Multi-Modal Deep Neural Networks With Applications to Voice Activity Detection and Guided Super Resolution

Ido Ariav

Multi-Modal Deep Neural Networks With Applications to Voice Activity Detection and Guided Super Resolution

Research Thesis

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

Ido Ariav

Submitted to the Senate

of the Technion — Israel Institute of Technology

Tamuz 5783 Haifa July 2023

This research was carried out under the supervision of Prof. Israel Cohen in the Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering.

The author of this thesis states that the research, including the collection, processing and presentation of data, addressing and comparing to previous research, etc., was done entirely in an honest way, as expected from scientific research that is conducted according to the ethical standards of the academic world. Also, reporting the research and its results in this thesis was done in an honest and complete manner, according to the same standards.

List of Publications

Journal Papers

- Ariav, I., Dov, D. and Cohen, I., "A Deep Architecture for Audio-Visual Voice Activity Detection in the Presence of Transients," *Signal Processing*, Vol. 142, pp.69–74, January 2018.
- Ariav, I. and Cohen, I., "An End-to-End Multimodal Voice Activity Detection Using WaveNet Encoder and Residual Networks," *IEEE Journal of Selected Topics in Signal Processing*, special issue on Data Science: Machine Learning for Audio Signal Processing, Vol. 13, Issue 2, pp.265–274, May 2019.
- Ariav, I. and Cohen, I., "Depth Map Super-Resolution Via Cascaded Transformers Guidance," *Frontiers in Signal Processing*, Vol. 3, Article 847890, pp. 1–-12, March 2022.
- Ariav, I. and Cohen, I., "Fully Cross Attention Transformer for Guided Depth Super Resolution," Sensors, Special Issue on Deep Learning Technology and Image Sensing, Vol. 23, No. 5, Article 2753, pp. 1-17, March 2023.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Israel Cohen, for his guidance. His great advice, consistent support, and helpful ideas had a huge impact on my work.

Finally, I wish to thank my family for their understanding and help during the journey of my Ph.D.

Contents

List of Figures

Abstract

Multimodal learning has advanced rapidly in several areas over the last decade, including computer vision and audio processing. A multimodal deep learning framework enables models to predict based on multiple modalities. This thesis introduces multimodal deep neural networks with specific applications in audio-visual voice activity detection and guided super-resolution.

The first task this thesis focuses on is audio-visual voice activity detection (VAD). We investigate VAD in challenging acoustic environments, such as high noise levels and transients found in real-life scenarios. An audio and video signal is captured by a camera pointed at the speaker's face in a multimodal setting. Accordingly, speech detection translates to the question of how to properly fuse the audio and video signals, which we address within the framework of deep learning. An architecture based on a variant of auto-encoders is presented, which combines the two modalities, providing a new representation of the signal that reduces interference. The new representation is fed into a recurrent neural network for speech detection, which is trained in a supervised manner to further encode differences between speech dynamics and interfering transients.

In follow-up work, we propose incorporating auditory and visual modalities into an end-to-end deep neural network for VAD. In noisy conditions, robust features must be extracted from both modalities to accurately distinguish speech from noise. We use a deep residual network to extract features from the video signal, while WaveNet encoders are used for feature extraction in the audio modality. To form a joint representation of speech, features from both modalities are fused using multimodal compact bilinear pooling. We then train the system in an end-to-end supervised fashion using a long short-term memory network to encode the temporal information.

The second topic this thesis addresses is guided super-resolution of depth information. Depth information captured by affordable depth sensors is characterized by low spatial resolution, which limits potential applications. Several methods have been proposed for guided super-resolution of depth maps using convolutional neural networks to overcome this limitation. In a guided super-resolution scheme, high-resolution depth maps are inferred from low-resolution ones with the additional guidance of a corresponding high-resolution intensity image. However, these methods are still prone to texture copying issues due to improper guidance by the intensity image. Specifically, in most existing methods, guidance from the color image is achieved by a naive concatenation of color and depth features. We propose a multi-scale residual deep network for depth map super-resolution. A cascaded transformer module incorporates high-resolution structural information from the intensity image into the depth upsampling process.

We additionally propose a fully transformer-based network for depth map super resolution. A cascaded transformer module extracts deep features from a low-resolution depth. It incorporates a novel cross-attention mechanism to seamlessly and continuously provide guidance from the color image into the depth upsampling process. Using a window partitioning scheme, linear complexity in image resolution can be achieved, so it can be applied to high-resolution images. The proposed methods of guided depth super resolution outperform other state-of-the-art methods through extensive experiments.

Abbreviations and Notations

Abbreviations

AUC	:	area under the curve
CAGM	:	cross-attention guidance module
CATB	:	cross-attention transformer block
CATL	:	cross-attention transformer layer
CFG	:	color feature guidance
CNN	:	convolutional neural network
CTGM	:	cascaded transformer guidance module
DBN	:	deep-belief network
DEM	:	dynamic elevation model
DL	:	deep learning
DSR	:	depth super resolution
FFT	:	fast Fourier transform
HR	:	high-resolution
LN	:	layer normalization
LR	:	low-resolution
LSTM	:	long short-term memory
MCB	:	multimodal compact bilinear
MFCC	:	mel frequency cepstral coefficients
MLP	:	multilayer perceptron
MSA	:	multiheaded self-attention
PLP	:	perceptual linear prediction
RDG	:	residual dilated groups

RELU	:	rectified linear unit
RMSE	:	root mean square error
RNN	:	recurrent neural network
ROC	:	receiver operating characteristic
SGD	:	stochastic gradient descent
SNR	:	signal-to-noise ratio
SR	:	super resolution
TN	:	true negative
TP	:	true positive
VAD	:	voice activity detection

Notations

$\mathbf{a}$	:	an audio signal
$\mathbf{a}_n$	:	feature representations of the nth frame of the *clean* audio signal
$\hat{\mathbf{a}}_n$	:	the nth frame of the *clean* raw audio signal
$\tilde{\mathbf{a}}_n$	:	the nth frame of the audio signal contaminated by background noises and transient i
$\{\mathbf{a}_n\}_1^N$	:	a dataset of length N frames of the *clean* audio signal
$\mathbf{A}$	:	an attention matrix
$\mathbf{b}$	:	a bias vector
$\mathbf{Conv}_3(\cdot)$	:	a convolution layer with a kernel size of 3×3
$\mathbf{Conv}_1(\cdot)$	:	a convolution layer with a kernel size of 1×1
$\mathbf{D}_{\mathrm{LR}}$	:	a low resolution depth map
$\mathbf{D}_{\mathrm{HR}}$	:	a reconstructed high resolution depth map
$\mathbf{F}$	:	the nonlinear mapping learned by a deep neural network
$\mathbf{h}$	:	a hidden layer in some neural network
$\mathcal{H}_0$	:	a hypothesis denoting speech absence
$\mathcal{H}_1$	:	a hypothesis denoting speech presence
$\mathbf{I}$	:	a 2D input image
$\mathbf{I}_{\mathrm{HR}}$	:	a high resolution guidance intensity image
$\mathbb{I}(n)$	:	a speech indicator of frame n
$\mathbf{K}$	:	keys vector in a transformer module
$\{\mathbf{K}_q\}$	:	a set of learnable filters in a convolution layer
$\mathcal{L}()$	:	a loss function
N	:	the input sequence length for a transformer encoder
(P, P)	:	an image patch of size $P \times P$
$\mathbf{Q}$	:	queries vector in a transformer module
s	:	a given scaling factor in a super resolution scenario
$\mathbf{S}_a^i$	:	the i^{th} audio sequences of length T
$\mathbf{S}_v^i$	:	the i^{th} video sequences of length T
$\tilde{\mathbf{S}}_a^i$	:	the i^{th} audio sequences of length T contaminated with noise
$\mathbf{v}$	:	a video signal

$\mathbf{v}_n$	:	feature representations of the nth frame of the *clean* video signal
$\hat{\mathbf{v}}_n$	:	the nth frame of the *clean* raw video signal
$\mathbf{V}$	:	values vector in a transformer module
$\mathbf{W}$	:	a weight matrix in a neural network
$\mathbf{x}$	:	an input vector to some neural network
$\mathbf{x}_n$	:	an input vector to some neural network representing the nth frame of a signal
$\mathbf{y}$	:	an output vector of some neural network
$\mathbf{y}_n$	:	an output vector of some neural network representing the nth frame of a signal
θ	:	the learned parameters of a deep neural network
σ	:	an element-wise activation function
Ψ	:	the count sketch projection function

Chapter 1

Introduction

1.1 Background and Motivation

Multimodality is inherent to the world around us. Our senses help us gather and fuse data we will be able to use later in our cognition process. The ability to leverage multiple modalities of perception data collectively is a fundamental mechanism in our sensory perception. It enables us to engage with the world under dynamic and unconstrained circumstances, with each modality serving as a unique source of information with different statistical properties. As an example, images convey a visual impression of a "walking dog" by means of thousands of pixels, whereas a corresponding text describes this moment using several words. See Fig. 1.1.

This has led to significant research efforts in multimodal deep learning (DL). In general, multimodal DL involves building models that can extract and relate information from multimodal data. A multimodal approach can significantly improve performance, as opposed to a unimodal approach that only provides partial insight into a research problem. This information from multiple sources is contextually related and often provides additional necessary information. This allows for more accurate predictions since it reveals features that would otherwise be hidden in individual data sources.

The use of multimodal deep learning involves a variety of technical challenges, including representation, translation, alignment, fusion, and co-learning. Learning how to fuse the extracted heterogeneous features into a common representation space by learning how to extract features from two or more modalities, often with different

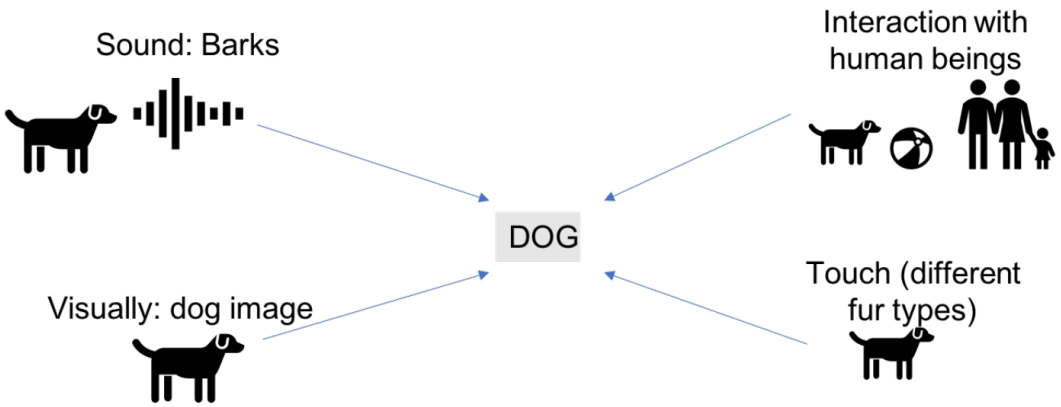

Figure 1.1: Conceptual representation of the animal dog from different modalities. [Dim22]

dimensions and properties (e.g., textual, visual, and auditory modalities), is of critical importance.

Notable recent advances and trends in multimodal DL include audio-visual speech recognition [YGS89], multimodal emotion recognition [Boa20], image and video captioning [BA18], Visual Question-Answering [AAL+15, ZPZ+20] and more.

The first application this thesis addresses is the one of voice activity detection (VAD). Voice activity detection is the problem of distinguishing between sections of a speech signal containing speech and non-speech sections. It fulfills an essential part in many modern speech-based systems such as those for speech and speaker recognition, speech enhancement, emotion recognition and dominant speaker identification.

Traditional methods of voice activity detection mostly rely on the assumption of quasi-stationary noise, i.e., the noise spectrum changes at a much lower rate than the speech signal. One group of such methods are those based on simple acoustic features such as zero-crossing rate and energy values in short time intervals [KN91, JMR94, VGX]. More advanced methods are model-based methods that focus on estimating a statistical model for the noisy signal [CK11, CKM06, SKS99, RSB+04]. The performance of such methods usually significantly deteriorates in the presence of even moderate levels of noise. Moreover, they cannot correctly model highly non-stationary noise and transient interferences, which are common in real-life scenarios and are within the main scope of this study, since the spectrum of transients, similarly to the spectrum

of speech, often rapidly varies over time [DTC16].

Apart from the statistical approaches noted above, more recent methods have been developed using machine learning techniques [SCK10, WZ11], and more specifically, using deep neural networks. Deep networks were successfully used to extract useful signal representations from raw data, and more specifically, several studies have also shown the favorable property of deep networks to model the inherent structure contained in the speech signal [HDY+12]. Deep neural networks were recently utilized in several modern voice activity detectors; Zhang and Wu [ZW13] proposed to extract a set of predefined acoustic features, e.g., Mel Frequency Cepstral Coefficients (MFCC), from a speech signal and then feed these features to a deep belief network (DBN) in order to obtain a more meaningful representation of the signal. They then used a linear classifier to perform speech detection. Thomas et al. [TGSS14] fed a convolutional neural network (CNN) with the log-Mel spectrogram together with its delta and delta-delta coefficients and trained the CNN to perform voice activity detection.

Despite showing improved performance compared to traditional methods, these networks classify each time frame independently, thus ignoring existing temporal relations between consecutive time frames. To alleviate this issue, several studies have suggested methods for modeling temporal relations between consecutive time frames [TR07]. More modern methods rely on recurrent neural networks (RNN) to incorporate previous inputs into the classification process, thus utilizing the signal's temporal information [LHB15, GMH13, HL13]. Hughes and Mierle [HM13] extracted Perceptual Linear Prediction (PLP) features from a speech signal and fed them to a multi-layered RNN with quadratic polynomial nodes to perform speech detection. Lim et al. [LJL16] proposed to transform the speech signal using a short-time Fourier transform and use a CNN to extract a high-level representation for the signal. This new representation was then fed to an LSTM to exploit the temporal structure of the data. These methods however still mostly rely on hand-crafted audio features, and often misclassify frames that contain both speech and transients as non-speech frames, since transients often appear more dominant than speech.

Although most of the current work on voice activity detection concentrates around the subject's audio signal, recent methods proposed to make use of other modalities,

such as visual information, to improve the voice detection [SRG+06, SRG+09, ARH+07, LWJ11, AHC10, TJY+12, AM08, YNO10, MLS+13]. The advantages of using a multimodal setting are most prominent when dealing with demanding acoustic environments, where high levels of acoustic noise and transient interferences are present since the video signal is entirely invariant to the acoustic environment. Therefore, proper incorporation of the video signal can significantly improve voice detection, as we show in this work. Ngiam et al. [NKK+11] proposed a Multimodal Deep Autoencoder for feature extraction from audio and video modalities. A bimodal DBN was used for the initialization of the deep autoencoder, and then the autoencoder was fine-tuned to minimize the reconstruction error of the two modalities. Zhang et al. [ZZHG16] used a CNN for classifying emotions in a multimodal setting. They applied two separate CNNs, the first operating on the mel-spectrogram of an audio signal and the second on a video recording of a subject's face. The features extracted using the two CNNs were concatenated and fed to a deep fully connected neural network to perform the classification. Dov et al. [DTC15] proposed to obtain separate low dimensional representations of the audio and video signals using diffusion maps [CL06]. The two modalities were then fused by a combination of speech presence measures, which are based on spatial and temporal relations between samples of the signal in the low dimensional domain.

The great majority of works described above still make use of commonly hand-crafted features in the audio or visual modality. To alleviate the need for hand-crafted features, a few studies proposed to adopt an end-to-end approach and use the raw, unprocessed data as input while utilizing as little human apriori knowledge as possible [GJ]. The motivation behind this is that a deep network can ultimately automatically learn an intermediate representation of the raw input signal that better suits the task at hand which in turn leads to improved overall performance. Trigeorgis et al. [TRB+16] proposed an end-to-end model for emotion recognition from raw audio signals. A CNN was used to extract features from the raw signal which were then fed to an LSTM network in order to capture the temporal information in the data. Tzirakis et al. [TTN+17] recently proposed another multimodal end-to-end method for emotion recognition. Features were extracted separately from audio and video signals using two CNNs and then concatenated to form a joint representation, which in turn was fed to

a multi-layered LSTM for classification. Hou et al. [HWL+18] proposed a multimodal end-to-end setting for speech enhancement. They used two CNNs to extract features from audio spectrograms and raw video, and these features were then concatenated and fed into several fully connected layers to produce an enhanced speech signal. This work, however, uses only simple concatenation to fuse the two modalities and does not utilize temporal information which we solve by incorporating LSTM. Petridis et al. [PSM+18] proposed an end-to-end approach for audio-visual speech recognition. They extracted features from each modality using a CNN and then fed these features to modality-specific RNN layers. The modalities were then fused by feeding the outputs of those RNNs to another RNN layer.

The second application this thesis addresses is the one of guided depth super resolution (DSR). High-resolution (HR) depth information of a scene plays a significant part in many applications such as 3D reconstruction [IKH+11], driving assistance [SSG+09], and mobile robots. Nowadays, depth sensors such as LIDAR or time-of-flight cameras are becoming more widely used. However, they often suffer from low spatial resolution, which does not always suffice for real-world applications. Thus, ongoing research has been done on reconstructing a high-resolution depth map from a corresponding low-resolution (LR) counterpart in a process termed depth super resolution (DSR).

The upsampling of depth information is not a trivial task since the fine details in the HR depth map are often missing or severely distorted in the LR depth map, because of the sensor's limited spatial resolution. Moreover, there often exists an inherent ambiguity in super-resolving the distorted fine details. A naive upsampling of the LR image, e.g., bicubic interpolation, usually produces unsatisfactory results with blurred and unsharp edges. Therefore, numerous advanced methods have been proposed recently for the upsampling, commonly termed super-resolution (SR), of depth information.

In recent years, many learning-based approaches based on elaborate convolutional neural networks (CNN) architectures for DSR were proposed [GLG+18, HLT16, RRB16, SDQ18, ZFY+19]. These methods surpassed the more classic approaches such as filter-based methods [HST12, YYDN07], and energy minimization-based methods [FRR+13, JHY+18, YYL+14] in terms of computation speed and the quality of the reconstructed HR information. Although CNN-based methods improved the performance significantly

compared with traditional methods, they still suffer from several drawbacks. To begin with, feature maps derived from a convolution layer have a limited receptive field, making long-range dependency modeling difficult. Second, a kernel in a convolution layer operates similarly on all parts of the input, making it content-independent and likely not the optimal choice. In contrast to CNN, Transformers, [VSP+17] have recently shown promising results in several vision-related tasks due to their use of attention. The attention mechanism enables the transformer to operate in a content-dependent manner, where each input part is treated differently according to the task.

In many cases, an LR depth map is accompanied by a corresponding HR intensity image. Many of the more recent methods propose to use this additional image to guide or enhance the SR of the depth map [GLG+18, HLT16, KHK13, KTL15, PKT+11, ZFY+19, ZZX+21, LDWS19, KJB+21, LDL20, YSW+20, CLYX21]. These methods assume that correspondence can be established between an edge in the intensity image and the matching edge in the depth map. Then, given that the intensity image has a higher resolution, its edges can determine depth discontinuities in the super-resolved HR depth map. However, there could be cases in which an edge in the intensity image does not correspond to a depth discontinuity in the depth map or vice versa, e.g., in the case of smooth, highly textured surfaces in the intensity image. These cases lead to texture copying, where color textures are over-transferred to the super-resolved depth map at the boundaries between textured and homogeneous regions. Hence, a more sophisticated guidance scheme needs to be considered.

1.2 Main Contributions

In this research thesis, the aforementioned topics are addressed. In essence, the thesis provides four main contributions. Two of them relate to the topic of audio-visual voice activity detection; the other two concern guided super resolution of depth information:

- In a multimodal setting, in which a speech signal is captured by a microphone and a video camera, we present a neural network architecture based on a variant of auto-encoders, which combines the audio-visual modalities, and provides a new representation of the signal, in which the effect of interferences is reduced.

To further encode differences between the dynamics of speech and interfering transients, the signal, in this new representation, is fed into a recurrent neural network, which is trained in a supervised manner for speech detection.

- In the same audio-visual multimodal setting, we utilize a deep residual network to extract features from the video signal, while for the audio modality, we employ a variant of WaveNet encoder for feature extraction. The features from both modalities are fused using multimodal compact bilinear pooling to form a joint representation of the speech signal. To further encode the temporal information, we feed the fused signal to a long short-term memory network and the system is then trained in an end-to-end supervised fashion.

- We propose a multi-scale residual deep network for guided depth map super-resolution. A cascaded transformer module incorporates high-resolution structural information from the intensity image into the depth upsampling process. The proposed cascaded transformer module achieves linear complexity in image resolution, making it applicable to high-resolution images.

- We propose a fully transformer-based network for depth map super resolution. A cascaded transformer module extracts deep features from a low-resolution depth. It incorporates a novel cross-attention mechanism to seamlessly and continuously provide guidance from the color image into the depth upsampling process. Using a window partitioning scheme, linear complexity in image resolution can be achieved, so it can be applied to high-resolution images.

1.3 Overview of the Thesis

In Chapter 3, a deep neural network architecture for audio-visual voice activity detection is presented. The architecture is based on specifically designed auto-encoders providing an underlying representation of the signal, in which, simultaneously, data from audio and video modalities are fused, and the effect of transients is reduced. The new representation is then incorporated into an RNN, which, in turn, is trained for speech presence/absence classification by incorporating temporal relations between

samples of the signal in the new representation. The classification is performed in a frame-by-frame manner without temporal delay, which makes the proposed deep architecture suitable for online applications. The proposed deep architecture is evaluated in the presence of highly non-stationary noises and transient interferences.

The proposed deep architecture is evaluated in the presence of highly non-stationary noises and transient interferences. Experimental results show improved performance of the proposed architecture compared to single-modal approaches that exploit only the audio or video signals, thus demonstrating the advantage of audio-video data fusion. In addition, we show that the proposed architecture outperforms competing multimodal detectors.

In Chapter 4, we present a deep end-to-end neural network architecture for audio-visual voice activity detection. First, we extract meaningful features from the raw audio and video signals; for the video signal, we employ a ResNet-18 network [HZRS] as a feature extractor, and for the audio signal, we employ a variant of a WaveNet [VDODZ$^+$] encoder. Then, instead of merely concatenating the feature vectors extracted from the two modalities, as is common in most multimodal networks, we propose to fuse the two vectors into a new joint representation via a Multimodal Compact Bilinear (MCB) pooling [GBZD] module, which was shown to efficiently and expressively combine multimodal features. The output of the MCB module is fed to several stacked LSTM layers in order to explore even further the temporal relations between samples of the speech signal in the new representation. Finally, a fully connected layer is used to perform the classification of each time frame to speech/non-speech, and the entire network is trained in a supervised end-to-end manner. The proposed deep end-to-end architecture is evaluated in the presence of highly non-stationary noises and transient interferences.

In Chapter 5, we propose a deep architecture for guided super resolution of depth information. To alleviate the texture copying problem, we propose a Cascaded Transformer Guidance Module (CTGM) for guided depth map SR. Our proposed CTGM is constructed by stacking several transformer blocks, each operating locally within non-overlapping windows that partition the entire input. Window shift is introduced between consecutive transformer blocks to enable inter-window connections to be learned. The CTGM is fed with HR features extracted from the intensity image and is trained

to pass only salient and consistent features that are then incorporated into the depth upsampling process. Our proposed CTGM is capable of learning structural and content information from a large receptive field, which was shown to be beneficial for SR tasks [ZZGZ17].

Our proposed CTGM exhibits linear memory constraints, making it applicable even for very large images. Furthermore, Unlike other transformer architectures, our architecture can handle different input resolutions, both during training and inference, making it highly applicable to real-world tasks.

Our overall architecture can be divided into three main parts: a depth branch, an intensity branch, and the CTGM. The proposed depth branch comprises several Residual Dilated Groups (RDG) [ZLL$^+$18] and performs the upsampling of the given LR depth map in a multi-scale manner, as in, e.g., [HLT16]. Meanwhile, the intensity image is fed into the intensity branch, which extracts HR features and complements the LR depth structures in the depth branch via the CTGM. This process is repeated according to the desired upsampling factor. This closely guided multi-scale scheme allows the network to learn rich hierarchical features at different levels, and better adapt to the upsampling of both fine-grained and coarse patterns. Moreover, this enables the network to seamlessly utilize the guidance from HR intensity features in multiple scales.

In Chapter 6, we propose a novel, fully transformer-based architecture for guided DSR. Specifically, the proposed architecture consists of three modules: shallow feature extraction, deep feature extraction and fusion, and an upsampling module. In this paper, we term the feature extraction and fusion module the cross-attention guidance module (CAGM). The shallow feature extraction module uses convolutional layers to extract shallow features from LR depth and HR color images, which are directly fed to the CAGM to preserve low-frequency information. Next, several transformer blocks are stacked to form the CAGM, each operating in non-overlapping windows from the previous block. Guidance from the color image is introduced via a cross-attention mechanism. In this manner, guidance from the HR color image is seamlessly integrated into the deep feature extraction process, enabling the network to focus on salient and meaningful features and to enhance the edge structures in the depth features while

suppressing textures in the color features. Moreover, using transformer blocks allows learning of structure and content from a wide receptive field, which is beneficial for SR tasks [ZZGZ17]. As a final step, shallow and deep features are fused in the upsampling module to reconstruct HR depth.

Our transformer-based architecture with a novel guidance mechanism that leverages cross-attention to seamlessly integrate guidance features from a color image to the DSR process. Also, linear memory constraints make the proposed architecture applicable even for large inputs.

1.4 Organization

This research thesis is organized as follows. Chapter 2 provides a high-level scientific background for the performed research. The contribution of this thesis is elaborated in Chapters 3 to 6. Chapters 3 and 4 are dedicated to the presentation of multimodal voice activity detection in an audio-visual setting. Chapters 5 and 6 introduce two transformer-based architectures for the task of guided depth super resolution. Finally, Chapter 7 concludes this thesis and proposes directions for future research.

Chapter 2

Background and Formulation

2.1 Audio-Visual voice activity detection

In this section, we provide a brief overview of multimodal voice activity detection. We consider a speech signal simultaneously recorded via a single microphone and a video camera pointed at a front-facing speaker. The video $\mathbf{v}$ signal usually comprises the mouth region of the speaker. It is aligned to the audio signal $\mathbf{a}$ by a proper selection of the frame length and the overlap of the audio signal.

The clean audio signal $\mathbf{a}$ can be used to label each time frame n according to the presence or absence of speech, and then assign each frame in $\mathbf{a}$ and $\mathbf{v}$, with the appropriate label. Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be two hypotheses denoting speech absence and presence, respectively, and let $\mathbb{I}(n)$ be a speech indicator of frame n, given by:

$$\mathbb{I}(n) = \begin{cases} 1, & n \in \mathcal{H}_1 \\ 0, & n \in \mathcal{H}_0 \end{cases} . \tag{2.1}$$

A voice activity detector aims to estimate $\mathbb{I}(n)$, i.e., to classify each frame n as a speech or non-speech frame.

In most cases, VADs operate on some feature representation of the frames in $\mathbf{a}$ and $\mathbf{v}$. Denote, $\mathbf{a}_n \in \mathbb{R}^A$ and $\mathbf{v}_n \in \mathbb{R}^V$ as feature representations of the nth frame of the *clean* audio and video signals, respectively, where A and V are the number of features. Alternatively, other VADs operate on raw input signals. Here $\hat{\mathbf{a}}_n \in \mathbb{R}$ and $\hat{\mathbf{v}}_n \in \mathbb{R}^{H \times W \times 3}$ represent the audio and video frames, respectively, where H and

Algorithm 2.1 Inject Random Background Noise and Transient Interference

▷ *Init noises and transients lists:*
Noises : {white Gaussian noise, musical instruments noise, babble noise, none}
Trans : {door-knocks, hammering, keyboard typing, metronome, scissors, none}

▷ *For each sequence* $\mathbf{S}_a^i$ *in the dataset:*
for $i = 1 \rightarrow$ *#sequences in dataset* **do**
  $L \leftarrow$ Length of sequence $\mathbf{S}_a^i$ in frames
  ▷ *randomly choose noises to inject:*
  noise $\leftarrow$ uniformly random noise from "Noises"
  trans $\leftarrow$ uniformly random transient from "Trans"
  $SNR \leftarrow$ uniformly random value from [0,20]
  ▷ *randomly inject sequences of length* L *from* N, T*:*
  if *noise* is not "none" **then**
    $R_{noise} \leftarrow$ random sequence of length L from *noise*
    $R_{noise} \leftarrow R_{noise}/std(R_{noise})$
    ▷ update R_{noise}'s SNR
    $R_{noise} \leftarrow R_{noise} \times (std(\mathbf{S}_a^i)/(10^{(SNR/20)})$
    $\mathbf{S}_a^i \leftarrow \mathbf{S}_a^i + R_{noise}$
  end if
  if *trans* is not "none" **then**
    $R_{trans} \leftarrow$ random sequence of length L from *trans*
    $\mathbf{S}_a^i \leftarrow \mathbf{S}_a^i + 2 \times R_{trans}$
  end if
end for

return $\mathbf{S}_a^i$ as $\tilde{\mathbf{S}}_a^i$

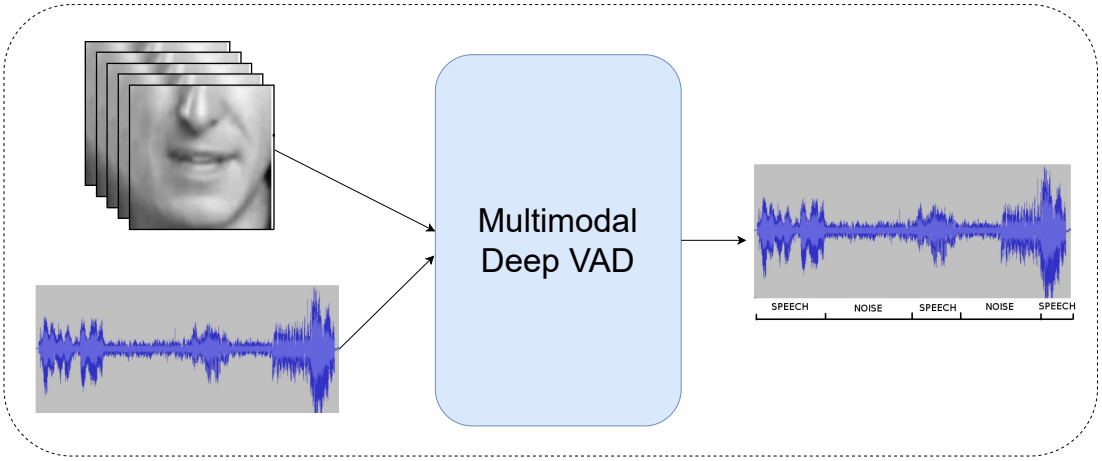

Figure 2.1: An illustration of an audio-visual voice activity detection system.

W are the height and width of each video frame in pixels, respectively. We define $\tilde{\mathbf{a}}_n$ to be the nth frame of the audio signal contaminated by background noises and transient interferences. Each $\tilde{\mathbf{a}}_n$ can be generated from $\mathbf{a}_n$(or $\hat{\mathbf{a}}_n$) by randomly adding background and transient noises, as presented in Alg. 2.1, for the case of $L = 1$.

Other VAds take into account L past frames in the classification process. For that we denote $\mathbf{S}_a^i$ and $\mathbf{S}_v^i$, respectively, the i^{th} audio and video sequences. Since only past frames are used in the classification process, each sequence is labeled according to the label of the last frame in the sequence. That is, the sequence $\mathbf{S}_v^i$ containing the video frames $\{\mathbf{v}_{i-L}, \mathbf{v}_{i-(L-1)}, \mathbf{v}_{i-(L-2)}, ..., \mathbf{v}_i\}$ is assigned the label given to $\mathbf{a}_i$, the i^{th} frame of the audio signal $\mathbf{a}$. Each sequence $\mathbf{S}_a^i$ is contaminated with noise according to the procedure outlined in Alg. 2.1 to produce the contaminated sequence $\tilde{\mathbf{S}}_a^i$.

It is worth noting that VAD is especially challenging in the presence of transients, which are typically more dominant than speech due to their short duration, high amplitudes and fast variations of the spectrum [DTC16]. Specifically, frames that contain both speech and transients, for which $\mathcal{H}_1$ holds, are often similar in the feature space to non-speech frames that contain only transients so that they are often wrongly classified as non-speech frames.

2.2 Autoencoders

In this section, we provide a short review of deep autoencoders [HS06]. An autoencoder is a feed-forward neural network with input and output layers of the same size, which we denote by $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{y} \in \mathbb{R}^D$, respectively. They are connected by one hidden layer $\mathbf{h} \in \mathbb{R}^M$, such that the input layer $\mathbf{x}$ is mapped into the hidden layer $\mathbf{h}$ through an affine mapping:

$$\mathbf{h} = \sigma\left(\mathbf{W}\mathbf{x} + \mathbf{b}\right), \tag{2.2}$$

where $\mathbf{W}$ is a $D \times M$ weight matrix, $\mathbf{b}$ is a bias vector and σ is an element-wise activation function. Then, $\mathbf{h}$ is mapped into the output layer $\mathbf{y}$:

$$\mathbf{y} = \tilde{\sigma}\left(\tilde{\mathbf{W}}\mathbf{h} + \tilde{\mathbf{b}}\right), \tag{2.3}$$

where $\tilde{\mathbf{W}}, \tilde{\mathbf{b}}, \tilde{\sigma}$ are defined similarly to $\mathbf{W}, \mathbf{b}$ and σ.

Optimal parameters (weights) $\tilde{\mathbf{W}}, \mathbf{W}, \tilde{\mathbf{b}}, \mathbf{b}$ are those that allow reconstructing the signal $\mathbf{x}$ at the output $\mathbf{y}$ of the autoencoder, and they are obtained via a training procedure, by optimizing a certain loss function $\mathcal{L}(\mathbf{x}, \mathbf{y})$, e.g., a square error.

It has been shown [VLL$^+$10, RHW88] that minimization of the autoencoder's loss function $\mathcal{L}(\mathbf{x}, \mathbf{y})$ is equivalent to maximization of a lower bound on the retained information between the input and output of the autoencoder. Thus, the hidden layer $\mathbf{h}$, obtained by (2.2) with optimized parameters $\mathbf{W}$ and $\mathbf{b}$, has the maximal mutual information with the input signal $\mathbf{x}$. The activation functions $\sigma, \tilde{\sigma}$ are usually chosen to be non-linear functions. e.g. a sigmoid function $\sigma(z) = \frac{1}{1+\exp -z}$, so that the hidden layer $\mathbf{h}$ incorporates non-linear relations between different parts of the input signal [HS06, BL$^+$07]. In addition, the dimension M of $\mathbf{h}$ is typically set smaller than that of the input signal D. Therefore, the hidden layer $\mathbf{h}$ is often considered as a non-linear low-dimensional representation of the input signal.

A deep architecture of autoencoders is constructed by stacking L autoencoders such that the lth hidden layer, denoted by $\mathbf{h}^l$, is used as an input for the $(l+1)$th layer. The training is performed one layer at a time in a bottom-up fashion. The first layer of the deep architecture is trained with $\mathbf{x}$ as input, and once trained, $\mathbf{h}^1$ is calculated by (2.2)

using the optimized parameters $\mathbf{W}^1$ and $\mathbf{b}^1$, where $\mathbf{W}^l$ and $\mathbf{b}^l$ denote the parameters of the lth layer. Then, parameters $\mathbf{W}^1$ and $\mathbf{b}^1$ are fixed, and the obtained $\mathbf{h}^1$ is used as input for the training procedure of the second layer and similarly for all layers up-to L.

2.3 Recurrent Neural Networks and LSTM

An RNN is a feed-forward multi-layered neural network in which loop connections, which are added to the hidden layers, allow to the incorporation of temporal information in the decision process.

Given an input vector $\mathbf{x}_n$, an RNN with one hidden layer $\mathbf{h}_n$ computes the output layer $\mathbf{y}_n$ using a hidden layer at time frame $n-1$, according to:

$$\mathbf{h}_n = \sigma\left(\hat{\mathbf{W}}\mathbf{x}_n + \grave{\mathbf{W}}\mathbf{h}_{n-1} + \mathbf{b}_n\right) \tag{2.4}$$

$$\mathbf{y}_n = \bar{\sigma}\left(\bar{\mathbf{W}}\mathbf{h}_n + \bar{\mathbf{b}_n}\right) \tag{2.5}$$

where $\hat{\mathbf{W}}, \grave{\mathbf{W}}$ and $\bar{\mathbf{W}}$ are weight matrices, $\mathbf{b}$ and $\bar{\mathbf{b}}$ are the bias parameters, and $\sigma, \bar{\sigma}$ are the corresponding activation functions. The case of an RNN with one hidden layer is extended to the case of an RNN with $L > 1$ layers by iteratively calculating the hidden layers for $l = 1$ to L:

$$\mathbf{h}_n^l = \sigma\left(\hat{\mathbf{W}}^l\mathbf{h}_n^{l-1} + \grave{\mathbf{W}}^l\mathbf{h}_{n-1}^l + \mathbf{b}_n^l\right) \tag{2.6}$$

where $\mathbf{h}_n^l$ is the lth hidden layer at time n, and $\hat{\mathbf{W}}, \grave{\mathbf{W}}$ and $\mathbf{b}$ are defined as in (2.4). The first layer is the input layer, i.e., $\mathbf{h}_n^0 \triangleq \mathbf{x}_n$, and the output layer $\mathbf{y}_n$ is calculated from (2.5) using the last hidden layer $\mathbf{h}_n^{\bar{L}}$.

We note, that RNN has two beneficial properties for voice activity detection. First, the length of the temporal window used for speech detection is implicitly incorporated in the weights $\{\grave{\mathbf{W}}^l\}_1^L$, and is automatically learned during the training process rather than being arbitrarily predefined. Second, the speech indicator is obtained via a supervised procedure, which exploits the true labels of the presence of speech and allows for accurate detection of speech.

When applied to long sequences, RNNs suffer from vanishing gradients, which hinder learning. The gradients carry information used in the RNN parameter update. The parameter updates become insignificant when the gradient becomes smaller and smaller, which means no real learning takes place. Additionally, RNNs struggle to learn long-term dependencies in input data.

In light of these limitations, Long Short Term Memory networks were proposed in [hoc]. LSTMs are a special kind of RNN, explicitly designed to avoid the long-term dependency problem and overcome vanishing gradients.

A common LSTM unit consists of a cell, an input gate, an output gate, and a forget gate. Three gates control the flow of information into and out of the cell, which allows it to remember values over arbitrary periods of time. The forget gate determines what information to discard by comparing a previous state to a current input and assigning a value between 0 and 1. A value of 1 indicates that the information is kept, and a value of 0 indicates that it is discarded. Input gates decide which pieces of newly received information to store in the current state, using the same system as forget gates. The output gates determine which pieces of information from the current state are output. This allows the LSTM network to maintain useful, long-term dependencies in the data by selectively outputting relevant information from the current state.

Formally, the LSTM cell model is characterized as follows:

$$
\begin{aligned}
\mathbf{f}_n &= \sigma\left(\hat{\mathbf{W}}_f \mathbf{x}_n + \grave{\mathbf{W}}_f \mathbf{h}_{n-1} + \mathbf{b}_f\right) \\
\mathbf{i}_n &= \sigma\left(\hat{\mathbf{W}}_i \mathbf{x}_n + \grave{\mathbf{W}}_i \mathbf{h}_{n-1} + \mathbf{b}_i\right) \\
\mathbf{o}_n &= \sigma\left(\hat{\mathbf{W}}_o \mathbf{x}_n + \grave{\mathbf{W}}_o \mathbf{h}_{n-1} + \mathbf{b}_o\right) \\
\tilde{\mathbf{c}}_n &= \tilde{\sigma}\left(\hat{\mathbf{W}}_c \mathbf{x}_n + \grave{\mathbf{W}}_c \mathbf{h}_{n-1} + \mathbf{b}_c\right) \\
\mathbf{c}_n &= \mathbf{f}_n \otimes \mathbf{c}_{n-1} + \mathbf{i}_n \otimes \tilde{\mathbf{c}}_n \\
\mathbf{h}_n &= \mathbf{o}_n \otimes \tilde{\sigma}\left(\mathbf{c}_n\right)
\end{aligned}
\tag{2.7}
$$

where $\mathbf{x}_n$ is the input vector, $\mathbf{f}_n$,$\mathbf{i}_n$, $\mathbf{o}_n$ and $\tilde{\mathbf{c}}_n$ are the activation vectors of the forget, input, output and cell gates, respectively. $\mathbf{c}_n$ is the cell state vector, and $\mathbf{h}_n$ is the hidden state vector which is also the output vector of the LSTM unit. $\hat{\mathbf{W}}$, $\grave{\mathbf{W}}$, and $\mathbf{b}$ are weight matrices and a bias vector that need to be learned during training. Here,

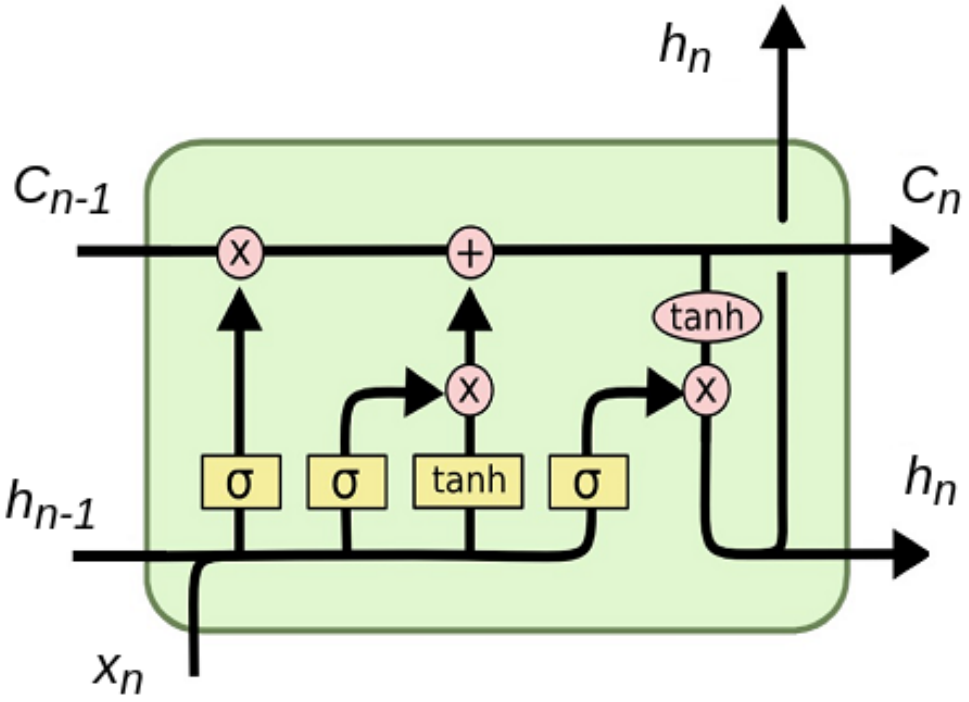

Figure 2.2: The general structure of an LSTM cell.

σ is a sigmoid function, and $\tilde{\sigma}$ is the hyperbolic tangent function.

2.4 Dilated Convolution

Convolutions are one of the main building blocks of modern neural networks. A convolution layer's parameters consist of a set of learnable filters $\{\mathbf{K}_q\}$, that have the same number of dimensions as the input they are convolved with. During the forward pass, these filters are convolved with the input, computing the dot product between the entries of the filter and the input to produce a new feature map. In most modern networks, relatively small filters are often used, thus each entry of the feature map has only a small receptive field of the input. In order to increase this receptive field, larger filters can be used, or many layers of small filters can be stacked, both at the expense of more burdensome computations. Alternatively, dilated convolutions can be utilized to increase the receptive field of each feature map entry without substantially increasing the computational cost.

In a dilated convolution, the filter $\mathbf{K}_q$ is applied over an area that is larger than

its size by skipping input values with a predetermined dilation factor. E.g, in the 2-D case, a convolution operation between a 3×3 filter $\mathbf{K}_q$ with a dilation factor of 2 and a 2-D input volume $\mathbf{I}$ at location (p, o) is given by:

$$(\mathbf{I} * \mathbf{K}_q)(p, o) = \sum_{l=-1}^{1} \sum_{m=-1}^{1} \mathbf{I}(p - 2\,l, o - 2\,m)\, \mathbf{K}_q(l + 1, m + 1) \tag{2.8}$$

where $*$ denotes the convolution operator. By skipping input values, a dilated convolution effectively operates on a larger receptive field than a standard convolution.

In order to increase the receptive field of all feature maps' entries even further, without increasing the computational load, several dilated convolutions can be stacked [YK16]. Notably, a block constructed of 10 dilated convolutions with exponentially increasing dilation factors of 1, 2, 4, . . . , 512, has a receptive field of size 1024 and can be considered a more efficient and discriminative non-linear counterpart of a 1×1024 regular convolution layer.

2.5 Multimodal Compact Bilinear Pooling

Bilinear pooling is a fusion method for two vectors, $\mathbf{x} \in \mathbb{R}^{B_1}$ and $\mathbf{q} \in \mathbb{R}^{B_2}$, in which the joint representation is simply the outer product between the two vectors. This allows, in contrast to an element-wise product or simple concatenation, a multiplicative interaction between all elements of both vectors. Bilinear pooling models [TF00] have recently been used for fine-grained classification tasks [LRM]. However, their main drawback is their high dimensionality of $B_1 \times B_2$, which can be very large in most real-life cases and leads to an infeasible number of learnable parameters. For example, if B_1 and B_2 are the lengths of two vector embeddings where $B_1 = B_2 = 512$, the joint representation is of length 512^2. This inevitably leads to very high memory consumption and high computation times and can also result in over-fitting.

To overcome the aforementioned limitation, multimodal compact bilinear pooling was presented in [FPY+16]. The multimodal compact bilinear pooling is approximated by projecting the joint outer product to a lower dimensional space while also avoiding computing the outer product directly. This is accomplished via the count sketch projection function suggested in [CCFC02], denoted as Ψ, which is a method of projecting

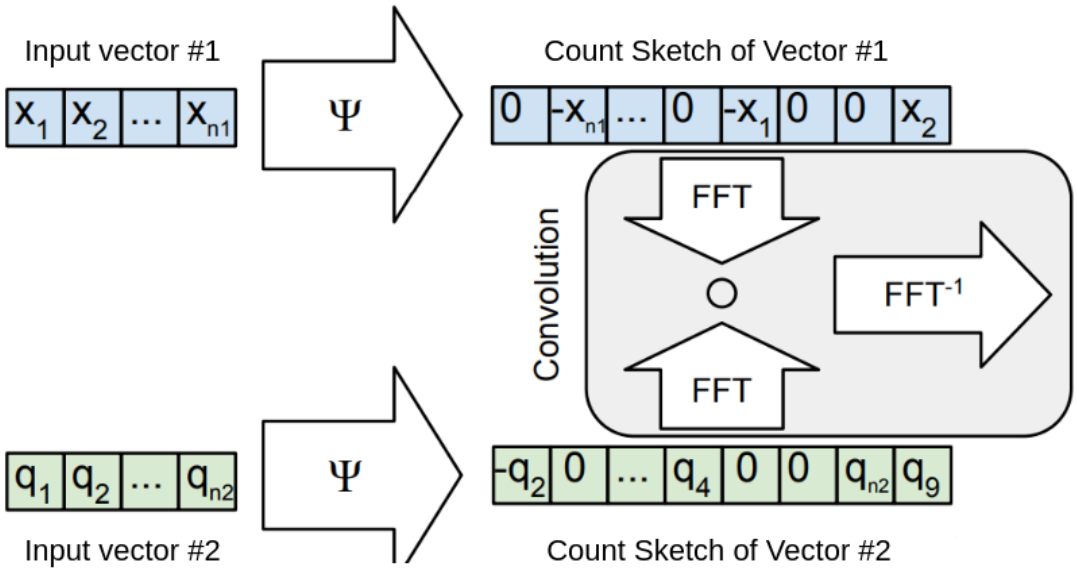

Figure 2.3: An illustration of multimodal compact bilinear pooling for two input vectors. [FPY+16]

a vector $\mathbf{c} \in \mathbb{R}^m$ to a lower dimensional representation $\hat{\mathbf{c}} \in \mathbb{R}^d$ for $d < m$. Instead of applying Ψ directly on the outer product of the two embeddings, [PP13] showed that explicitly computing the outer product of the two vectors can be avoided since the count sketch of the outer product can be expressed as a convolution of both count sketches. Additionally, the convolution theorem states that a circular convolution of two discrete signals can be performed with lower asymptotic complexity by performing multiplication in the frequency domain (note that linear convolution in the time domain may be replaced with a circular convolution by applying a proper zero-padding to the time-domain signals). Therefore, both vectors $\mathbf{x}$ and $\mathbf{q}$ are projected separately to a lower dimensional representation using Ψ, and then the element-wise products of fast Fourier transforms (FFT) of the lower dimensional representations is calculated to form the MCB output.

We note that MCB can easily be extended and remains effective for more than two modalities as the fusion of the modalities is achieved by the element-wise product of FFTs. Another advantage to using MCB for vector fusion is being able to choose the desired size for the joint vector. In contrast, when feature vectors are fused by simple concatenation, element-wise multiplication, or dot product, the joint representation is given by the sizes of the feature vectors. In MCB, the size of the joint representation

can be selected according to performance or computational requirements.

2.6 Guided super resolution of depth maps

A method for guided depth SR aims to find the nonlinear mapping between an LR depth map and the corresponding HR depth map. An HR intensity image guides the process of finding this nonlinear relation. For a given scaling factor $s = 2^m$ we denote the LR depth map as $\mathbf{D}_{\text{LR}} \in \mathbb{R}^{H/s \times W/s}$ and the respective HR guidance intensity image as $\mathbf{I}_{\text{HR}} \in \mathbb{R}^{H \times W}$. Then, the corresponding HR depth map $\mathbf{D}_{\text{HR}} \in \mathbb{R}^{H \times W}$ can be found from:

$$\mathbf{D}_{\text{HR}} = \mathbf{F}(\mathbf{D}_{\text{LR}}, \mathbf{I}_{\text{HR}}; \theta) \tag{2.9}$$

where $\mathbf{F}$ denotes the nonlinear mapping learned by a deep neural network, and θ represents the learned network's parameters.

2.7 Vision Transformers

In recent years, transformer-based architectures [VSP+17] achieved great success in natural language processing tasks, enabling long-range dependencies in the data to be learned via their sophisticated attention mechanism. Their tremendous success in the language domain has led researchers to investigate their adaptation to computer vision, where it has recently demonstrated promising results on certain tasks, specifically image classification [DBK+20, WXL+21, LLC+21] and object detection [ZSL+20, CMS+20].

A vanilla transformer encoder, as proposed in [VSP+17], usually consists of alternating layers of multiheaded self-attention (MSA) and MLP blocks, with Layer Normalization (LN) before every block and residual connections after every block.

An MSA block takes as input a sequence of length N of d-dimensional embeddings

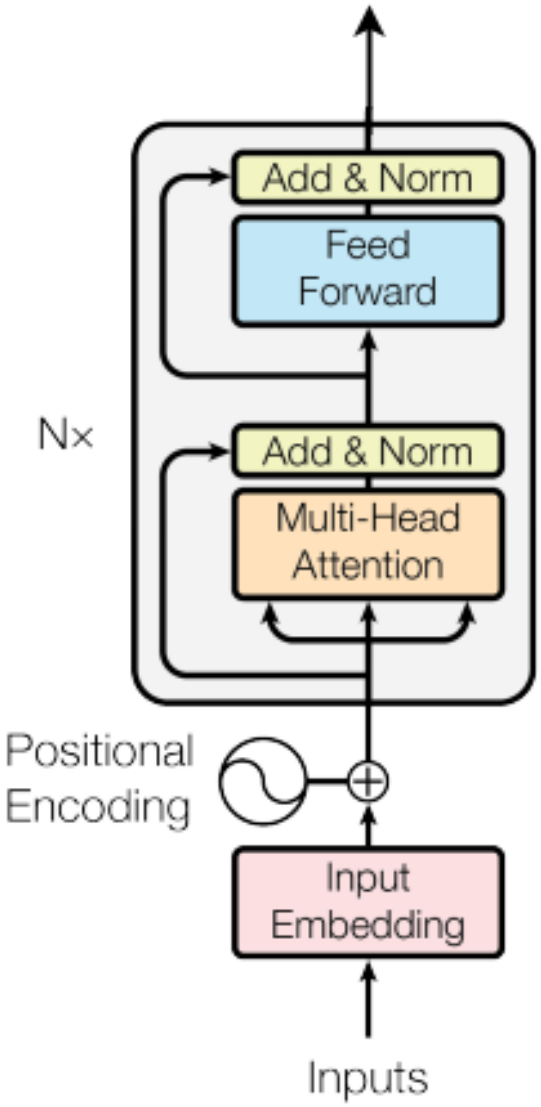

Figure 2.4: A vanilla transformer encoder. [VSP+17]

$\mathbf{x} \in \mathbb{R}^{N \times d}$ and produces an output sequence $\mathbf{y} \in \mathbb{R}^{N \times d}$ via:

$$\begin{aligned} \mathbf{Q} &= \mathbf{x}\mathbf{W}_Q, \mathbf{K} = \mathbf{x}\mathbf{W}_K, \mathbf{V} = \mathbf{x}\mathbf{W}_V \\ \mathbf{A} &= \text{Softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{\mathbf{d}}) \\ \mathbf{y} &= \mathbf{A}\mathbf{V} \end{aligned} \tag{2.10}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are $D \times D$ parameter matrices of 1×1 convolutions responsible for projecting the entries of the sequence $\mathbf{x}$ into the three standard transformer paradigms; keys, queries, and values, respectively. Each entry of the output sequence

$\mathbf{y}$ is a linear combination of values in $\mathbf{V}$ weighted by the attention matrix $\mathbf{A}$, which itself is computed from similarities between all pairs of query and key vectors.

To allow transformers to handle 2D images, an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ is first divided into non-overlapping patches of size (P, P). Each patch is flattened and projected to a d-dimensional vector via a trainable linear projection, forming the patch embeddings $\mathbf{x} \in \mathbb{R}^{N \times d}$ where H, W are the height and width of the image, respectively, C is the number of channels, and $N = H \times W / P^2$ is the total number of patches. Finally, N is the effective input sequence length for the transformer encoder. Patch embeddings are enhanced with position embeddings to retain 2D image positional information.

Transformers derive their modeling capabilities from computing self-attention $\mathbf{A}$ and $\bar{\mathbf{X}}$. Since self-attention has a quadratic cost in time and space, it cannot be applied directly to images as N quickly becomes unmanageable. As a result of this inherent limitation, modality-aware sequence length restrictions have been applied to preserve model performance while restricting sequence length. [DBK+20] showed that a transformer architecture could be directly applied to medium-sized image patches for different vision tasks. The aforementioned memory constraints are mitigated by this local self-attention.

Although the above self-attention module can effectively exploit intra-modality relationships in the input image, in a multi-modality setting, the inter-modality relationships, e,g., the relationships between different modalities, also need to be explored. Thus, a cross-attention mechanism was introduced in which attention masks from one modality highlight the extracted features in another. Contrary to self-similarity, wherein query, key, and value are based on similarities within the same feature array, in cross-attention, keys, and values are calculated from features extracted from one modality, while queries are calculated from the other. Formally, an MSA block using cross-attention is given by -

$$\mathbf{Q} = \hat{\mathbf{x}}\mathbf{W}_Q \, , \mathbf{K} = \mathbf{x}\mathbf{W}_K \, , \mathbf{V} = \mathbf{x}\mathbf{W}_V \tag{2.11}$$

where $\mathbf{x}$ is the input sequence of one modality and $\hat{\mathbf{x}}$ is the input sequence of the second modality. The calculation of attention matrix $\mathbf{A}$ and output sequence $\bar{\mathbf{y}}$ remains the

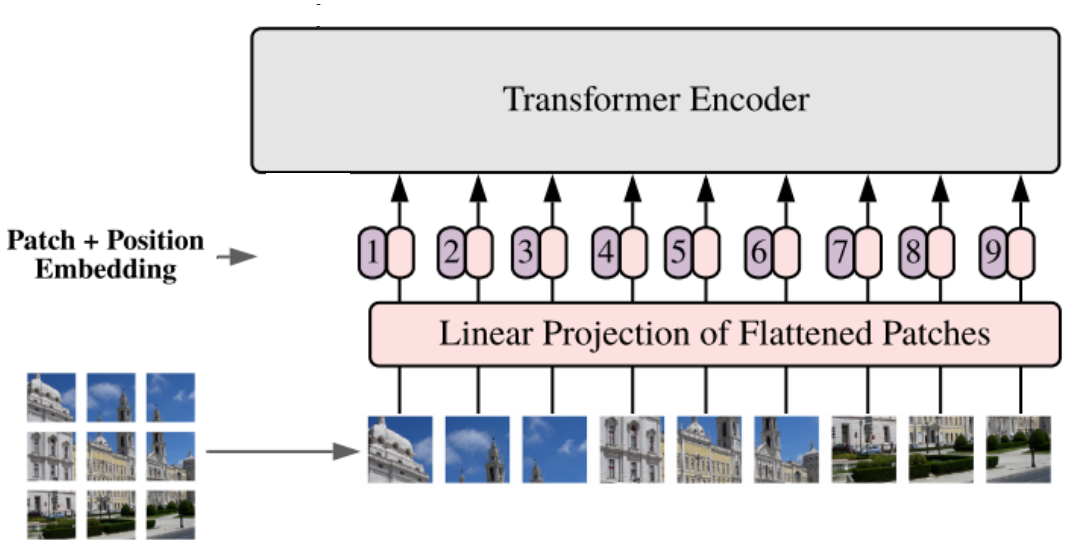

Figure 2.5: An illustration of a patch-based vision transformer. [DBK+20]

same.

Chapter 3

A Deep Architecture for Audio-Visual Voice Activity Detection in the Presence of Transients

Signal Processing 142 (2018) 69–74

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Short communication

A deep architecture for audio-visual voice activity detection in the presence of transients☆

Ido Ariav*, David Dov, Israel Cohen

The authors are with the Andrew and Erna Viterbi Faculty of Electrical Engineering, The Technion-Israel Institute of Technology, Haifa 32000, Israel

ARTICLE INFO

Article history:
Received 9 April 2017
Revised 5 July 2017
Accepted 11 July 2017
Available online 12 July 2017

Keywords:
Audio-visual speech processing
Voice activity detection
Auto-encoder
Recurrent neural networks

ABSTRACT

We address the problem of voice activity detection in difficult acoustic environments including high levels of noise and transients, which are common in real life scenarios. We consider a multimodal setting, in which the speech signal is captured by a microphone, and a video camera is pointed at the face of the desired speaker. Accordingly, speech detection translates to the question of how to properly fuse the audio and video signals, which we address within the framework of deep learning. Specifically, we present a neural network architecture based on a variant of auto-encoders, which combines the two modalities, and provides a new representation of the signal, in which the effect of interferences is reduced. To further encode differences between the dynamics of speech and interfering transients, the signal, in this new representation, is fed into a recurrent neural network, which is trained in a supervised manner for speech detection. Experimental results demonstrate improved performance of the proposed deep architecture compared to competing multimodal detectors.

1. Introduction

Voice activity detection is a segmentation problem of a given speech signal into sections that contain speech and sections that contain only noise and interferences. It constitutes an essential part in many modern speech-based systems such as those for speech and speaker recognition, speech enhancement, emotion recognition and dominant speaker identification. We consider a multimodal setting, in which speech is captured by a microphone, and a video camera is pointed at the face of the desired speaker. The multimodal setting is especially useful in difficult acoustic environments, where the audio signal is measured in the presence of high levels of acoustic noise and transient interferences, such as keyboard tapping and hammering [1,2]. The video signal is completely invariant to the acoustic environment, and nowadays, it is widely available in devices such as smart-phones and laptops. Therefore, proper incorporation of the video signal significantly improves voice detection, as we show in this paper.

In silent acoustic environments, speech segments in a given signal are successfully distinguished from the silence segments using methods based on simple acoustic features such as zero-crossing rate and energy values in short time intervals [3–5]. However, the performances of these methods significantly deteriorate in the presence of noise even with moderate levels of signal-to-noise ratios (SNR). Another group of methods assumes statistical models for the noisy signal, focusing on estimation of the model parameters. For example, the variances of speech and noise can be estimated by tracking the variations of the noisy signal over time [6–9]. The main drawback of such methods is that they cannot properly model highly non-stationary noise and transient interferences, which are in the main scope of this study. The spectrum of transients often rapidly varies over time, as does the spectrum of speech, and as a result, they are not properly distinguished [2].

More recent studies address the problem of voice activity detection from a machine learning point of view, in which the goal is to classify segments of the noisy signal into speech and non-speech classes [10,11]. Learning-based methods learn implicit models from training data instead of assuming explicit distributions for the noisy signal. A particular school of models, relevant to this paper, is deep neural networks, which have gained popularity in recent years in a variety of machine learning tasks. These models utilize multiple hidden layers for useful signal representations, and their potential for voice activity detection has been partially exploited in recent studies. Zhang and Wu [12] proposed using a deep-belief network to learn an underlying representation of a speech signal from predefined acoustic features. The new representation is then fed into a linear classifier for speech detection. Mendelev et al. [13] introduced a multi-layer perceptron network for speech detection, and proposed to improve its robustness to

☆ This research was supported by the Israel Science Foundation (grant no. 576/16).
* Corresponding author.
E-mail addresses: idoariav@tx.technion.ac.il, idoariav@gmail.com (I. Ariav), davidd@tx.technion.ac.il (D. Dov), icohen@ee.technion.ac.il (I. Cohen).

http://dx.doi.org/10.1016/j.sigpro.2017.07.006
0165-1684/

noise using the "Dropout" technique [14]. Despite the improved performance, the network in [13] classifies each time frame independently, thus ignoring temporal relations between segments of the signal. The studies presented in [15–18] propose using a recurrent neural network (RNN) to naturally exploit temporal information by incorporating previous inputs for voice detection. These methods however still struggle in frames that contain both speech and transients. Since transients are characterized by fast variations in time and high energy values, they often appear more dominant than speech. Therefore, frames containing only transients appear similar to frames containing both transients and speech, so that they are wrongly detected as speech frames.

A different school of studies suggests improving the robustness of speech detection to noise and transients by incorporating a video signal, which is invariant to the acoustic environment. Often, the video captures the mouth region of the speakers, and it is represented by specifically designed features, which model the shape and movement of the mouth in each frame. Examples of such features are the height and the width of the mouth [19,20], key-points and intensity levels extracted from the region of the mouth [21–24], and motion vectors [25,26].

Two common approaches exist in the literature concerning the fusion of audio and video signals, termed early and late fusion [27,28]. In early fusion, video and audio features are concatenated into a single feature vector and processed as single-modal data [29]. In late fusion, measures of speech presence and absence are constructed separately from each modality, and then combined using statistical models [30,31]. Dov et al. [32,33], for example, proposed to obtain separate low dimensional representations of the audio and video signals using diffusion maps. The two modalities are then fused by a combination of speech presence measures, which are based on spatial and temporal relations between samples of the signal in the low dimensional domain.

In this paper, we propose a deep neural network architecture for audio-visual voice activity detection. The architecture is based on specifically designed auto-encoders providing an underlying representation of the signal, in which simultaneous data from audio and video modalities are fused in order to reduce the effect of transients. The new representation is incorporated into an RNN, which, in turn, is trained for speech presence/absence classification by incorporating temporal relations between samples of the signal in the new representation. The classification is performed in a frame-by-frame manner without temporal delay, which makes the proposed deep architecture suitable for online applications.

The proposed deep architecture is evaluated in the presence of highly non-stationary noises and transient interferences. Experimental results show improved performance of the proposed architecture compared to single-modal approaches that exploit only the audio or video signals, thus demonstrating the advantage of audio-video data fusion. In addition, we show that the proposed architecture outperforms competing multimodal detectors.

The remainder of the paper is organized as follows. In Section 2, we formulate the problem. In Section 3, we introduce the proposed architecture. In Section 4, we demonstrate the performance of the proposed deep architecture for voice activity detection. Finally, in Section 5, we draw conclusions and offer some directions for future research.

2. Problem formulation

We consider a speech signal simultaneously recorded via a single microphone and a video camera pointed at a front-facing speaker. The video signal comprises the mouth region of the speaker. It is aligned to the audio signal by a proper selection of the frame length and the overlap of the audio signal as described in Section 4. Let $\mathbf{a}_n \in \mathbb{R}^A$ and $\mathbf{v}_n \in \mathbb{R}^V$ be feature representations of the nth frame of the *clean* audio and video signals, respectively, where A and V are the number of features. Similarly to $\mathbf{a}_n$, let $\tilde{\mathbf{a}}_n \in \mathbb{R}^A$ be a feature representation of the audio signal contaminated by background noises and transient interferences. The audio and the video features are based on the Mel Frequency Cepstral Coefficients (MFCC) and motion vectors, respectively, and their construction is described in Section 4.

We consider a dataset of N consecutive triplets of frames $(\mathbf{a}_1, \tilde{\mathbf{a}}_1, \mathbf{v}_1), (\mathbf{a}_2, \tilde{\mathbf{a}}_2, \mathbf{v}_2), \ldots, (\mathbf{a}_N, \tilde{\mathbf{a}}_N, \mathbf{v}_N)$ containing both speech and non-speech time intervals. We use the clean signal $\{\mathbf{a}_n\}_1^N$ to label each time frame n according to the presence or absence of speech. Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be two hypotheses denoting speech absence and presence, respectively, and let $\mathbb{I}(n)$ be a speech indicator of frame n, given by:

$$\mathbb{I}(n) = \begin{cases} 1, & n \in \mathcal{H}_1 \\ 0, & n \in \mathcal{H}_0 \end{cases}. \tag{1}$$

The goal in this study is to estimate $\mathbb{I}(n)$, i.e., to classify each frame n as a speech or non-speech frame.

Voice activity detection is especially challenging in the presence of transients, which are typically more dominant than speech due to their short duration, high amplitudes and fast variations of the spectrum [2]. Specifically, frames that contain both speech and transients, for which $\mathcal{H}_1$ holds, are often similar in the feature space to non-speech frames that contain only transients, so that they are often wrongly classified as non-speech frames. To address this challenge, we introduce a deep neural network architecture, which is designed to reduce the effect of transients by exploiting both the clean and the noisy audio signals, $\mathbf{a}_n$ and $\tilde{\mathbf{a}}_n$, respectively, and the video signal $\mathbf{v}_n$.

3. Deep architecture for audio-visual voice activity detection

3.1. Review of autoencoders

The proposed deep architecture is based on obtaining a transient reducing representation of the signal via the use of auto-encoders, which are shortly reviewed in this subsection for the sake of completeness [34]. An auto-encoder is a feed-forward neural network with an input and output layers of the same size, which we denote by $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{y} \in \mathbb{R}^D$, respectively. They are connected by one hidden layer $\mathbf{h} \in \mathbb{R}^M$, such that the input layer $\mathbf{x}$ is mapped into the hidden layer $\mathbf{h}$ through an affine mapping:

$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}), \tag{2}$$

where $\mathbf{W}$ is a $D \times M$ weight matrix, $\mathbf{b}$ is a bias vector and σ is an element-wise activation function. Then, $\mathbf{h}$ is mapped into the output layer $\mathbf{y}$:

$$\mathbf{y} = \tilde{\sigma}\left(\tilde{\mathbf{W}}\mathbf{h} + \tilde{\mathbf{b}}\right), \tag{3}$$

where $\tilde{\mathbf{W}}, \tilde{\mathbf{b}}, \tilde{\sigma}$ are defined similarly to $\mathbf{W}$, $\mathbf{b}$ and σ.

Optimal parameters (weights) $\tilde{\mathbf{W}}, \mathbf{W}, \tilde{\mathbf{b}}, \mathbf{b}$ are those that allow reconstructing the signal $\mathbf{x}$ at the output $\mathbf{y}$ of the auto-encoder, and they are obtained via a training procedure, by optimizing a certain loss function $L(\mathbf{x}, \mathbf{y})$, e.g., a square error, which we use here. It has been shown [35,36] that minimization of the auto-encoder's loss function $L(\mathbf{x}, \mathbf{y})$ is equivalent to maximization of a lower bound on the retained information between the input and output of the auto-encoder. Thus, the hidden layer $\mathbf{h}$, obtained by (2) with optimized parameters $\mathbf{W}$ and $\mathbf{b}$, has the maximal mutual information with the input signal $\mathbf{x}$. The activation functions $\sigma, \tilde{\sigma}$ are usually chosen to be non-linear functions; here, we use a sigmoid function $\sigma(z) = \frac{1}{1+\exp -z}$, so that the hidden layer $\mathbf{h}$ incorporates non-linear relations between different parts of the input signal [34,37]. In addition, the dimension M of $\mathbf{h}$ is typically set smaller than that of

the input signal D. Therefore, the hidden layer $\mathbf{h}$ is often considered as a non-linear low dimensional representation of the input signal.

A deep architecture of auto-encoders is constructed by stacking L auto-encoders such that the lth hidden layer, denoted by $\mathbf{h}^l$, is used as an input for the $(l+1)$th layer. The training is performed one layer at a time in a bottom-up fashion. The first layer of the deep architecture is trained with $\mathbf{x}$ as input, and once trained, $\mathbf{h}^1$ is calculated by (2) using the optimized parameters $\mathbf{W}^1$ and $\mathbf{b}^1$, where $\mathbf{W}^l$ and $\mathbf{b}^l$ denote the parameters of the lth layer. Then, we fix parameters $\mathbf{W}^1$ and $\mathbf{b}^1$ and use the obtained $\mathbf{h}^1$ as input for the training procedure of the second layer and similarly for all layers up to L.

3.2. Transient-reducing audio-visual autoencoder

We adopt ideas from [38,39] of using autoencoders to fuse multimodal signals. We propose a specifically designed deep architecture, based on feeding the auto-encoder with an audio-visual signal contaminated by acoustic noises and transients, while reconstructing the clean signal. Specifically, let $\mathbf{z}_n \in \mathbb{R}^{A+V}$ and $\tilde{\mathbf{z}}_n \in \mathbb{R}^{A+V}$ be feature vectors of frame n, obtained by concatenating the video features $\mathbf{v}_n$ along with the audio features $\mathbf{a}_n$ and $\tilde{\mathbf{a}}_n$, respectively, such that $\mathbf{z}_n = [\mathbf{a}_n^T, \mathbf{v}_n^T]^T$ and $\tilde{\mathbf{z}}_n = [\tilde{\mathbf{a}}_n^T, \mathbf{v}_n^T]^T$. The auto-encoder is fed by the noisy audio-visual feature vector $\tilde{\mathbf{z}}_n$, and is trained to reconstruct the clean signal $\mathbf{z}_n$, i.e., to minimize $L(\hat{\mathbf{z}}_n, \mathbf{z}_n)$ where $\hat{\mathbf{z}}_n \in \mathbb{R}^{A+V}$ is the output of the auto-encoder.

This approach simultaneously serves two purposes; it both allows fusing of the audio and the video modalities, and reduces the effect of transients. According to (2), the hidden layer $\mathbf{h}$ is obtained by a non-linear fusion between the entries of $\tilde{\mathbf{z}}$, and specifically, by the fusion of the audio and the video modalities. In addition, the effect of transients is reduced in the hidden layer $\mathbf{h}$ since the training process is designed to reconstruct the clean signal at the output. As a result, the hidden layer only captures factors that are related to the clean signal, as we demonstrate in Section 4.

We stack L such auto-encoders to form a deep neural network as described in Section 3.1. For layers $l > 1$ we can no longer use the clean and the noisy speech signals; instead, we follow the principle of a de-noising auto-encoder [35], i.e., corrupt each input $\mathbf{h}_n^l$ with random noise, and train the auto-encoder to reconstruct the uncorrupted input. Vincent et al. [35] have shown that stacking several auto-encoders yields an improved representation for the input data over an ordinary one layer auto-encoder, since the added layers allow the auto-encoder to learn more complex higher-order relations across the modalities. Assuming an architecture of L such auto-encoder layers, we consider the last layer of the network, denoted by $\mathbf{p}_n \triangleq \mathbf{h}_n^L$, as the new underlying representation of the audio-visual signal.

It is worth noting that the proposed representation significantly differs from the common early and late fusion approaches [27,28] since it is obtained via the exploration of complex relations between the audio and video signals.

3.3. Recurrent neural network for voice activity detection

Speech is an inherently dynamic process comprising rapidly alternating speech and non-speech segments, i.e., a speech segment followed by a non-speech segment (pause) and vice versa. Indeed, temporal information is widely used for improving voice activity detection by incorporating several consecutive frames in the decision process [8,9]. However, the number of previous frames that should be considered and their weight on the decision process is not straightforward, and can change over time. For example, a common assumption is that speech is present with a higher probability if it was present in previous frames rather than after a non-speech (silent) frame. Thus, predetermining the amount of past information considered in the classification process for all frames can result in suboptimal results. We address this issue by incorporating an RNN for the classification of each frame $\tilde{\mathbf{z}}_n$.

An RNN is a feed-forward multi-layered neural network in which loop connections, which are added to the hidden layers, allow to incorporate temporal information in the decision process.

Given the auto-encoder's output at time frame n, $\mathbf{p}_n$, an RNN with one hidden layer $\bar{\mathbf{h}}_n$ computes the output layer $\bar{\mathbf{y}}_n$ using a hidden layer at time frame $n-1$, according to:

$$\bar{\mathbf{h}}_n = \hat{\sigma}\left(\hat{\mathbf{W}}\mathbf{p}_n + \grave{\mathbf{W}}\bar{\mathbf{h}}_{n-1} + \hat{\mathbf{b}}_n\right) \tag{4}$$

$$\bar{\mathbf{y}}_n = \bar{\sigma}\left(\bar{\mathbf{W}}\bar{\mathbf{h}}_n + \bar{\mathbf{b}}_n\right) \tag{5}$$

where $\hat{\mathbf{W}}, \grave{\mathbf{W}}$ and $\bar{\mathbf{W}}$ are weight matrices, $\hat{\mathbf{b}}$ and $\bar{\mathbf{b}}$ are the bias parameters, and $\hat{\sigma}, \bar{\sigma}$ are the corresponding activation functions. The case of an RNN with one hidden layer is extended to the case of an RNN with $\bar{L} > 1$ layers by iteratively calculating the hidden layers for $l = 1$ to $\bar{L}$:

$$\bar{\mathbf{h}}_n^l = \hat{\sigma}\left(\hat{\mathbf{W}}^l\bar{\mathbf{h}}_n^{l-1} + \grave{\mathbf{W}}^l\bar{\mathbf{h}}_{n-1}^l + \hat{\mathbf{b}}_n^l\right) \tag{6}$$

where $\bar{\mathbf{h}}_n^l$ is the lth hidden layer at time n, and $\hat{\mathbf{W}}, \grave{\mathbf{W}}$ and $\hat{\mathbf{b}}$ are defined as in (4). The first layer is the input layer, i.e., $\bar{\mathbf{h}}_n^0 \triangleq \bar{\mathbf{x}}_n$, and the output layer $\bar{\mathbf{y}}_n$ is calculated from (5) using the last hidden layer $\bar{\mathbf{h}}_n^{\bar{L}}$.

We incorporate the proposed transient-reducing representation $\{\mathbf{p}_n\}_1^N$ into the deep RNN in order to exploit the temporal information inherent in speech for voice activity detection. Specifically, for each frame n, we feed the new representation $\mathbf{p}_n$ to the RNN and iteratively compute the hidden layers $\bar{\mathbf{h}}_n^l$ according to (6). Then, we use the output layer $\bar{\mathbf{y}}_n$, and apply a sigmoid function to constrain its values to the range of $0-1$. Thus, we consider the output as a probability measure for the presence of speech in frame n, and propose to estimate the speech indicator $\mathbb{I}(n)$ in (1) by comparing the output to a threshold t:

$$\mathbb{I}(n) = \begin{Bmatrix} 1, & \bar{\mathbf{y}}_n \geq t \\ 0, & \bar{\mathbf{y}}_n < t \end{Bmatrix}. \tag{7}$$

The RNN has two beneficial properties for voice activity detection. First, the length of the temporal window used for speech detection is implicitly incorporated in the weights $\{\grave{\mathbf{W}}^l\}_1^{\bar{L}}$, and is automatically learned during the training process rather than being arbitrarily predefined. Second, the speech indicator in (7) is obtained via a supervised procedure, which exploits the true labels of the presence of speech, and allows for an accurate detection of speech as we show in Section 4.

4. Experimental results

4.1. Experimental setting

4.1.1. Dataset

We evaluate the proposed deep architecture for voice activity detection using the dataset presented in [32]. The dataset includes audio-visual sequences of 11 speakers reading aloud an article chosen from the web, while making natural pauses every few sentences. Thus, the intervals of speech and non-speech range from several hundred ms to several seconds in length. The video signal uses a bounding box around the mouth region of the speaker, cropped from the original recording, and it is of 90×110 pixels. The audio signal is recorded at 8 kHz with an estimated SNR of ~ 25 dB. It is processed using short time frames of length 634 samples with 50% overlap such that it is aligned to the video frames which are processed at 25 frames/s. Each of the 11 sequences is

120 s long, and it is divided into two parts such that the first 60 s are used to train the algorithm and the rest of the sequence is used for evaluation.

The clean audio signal is contaminated with various background noises such as white Gaussian noise, musical instruments noise and babble noise, and with transients, such as a metronome, keyboard typing and hammering, taken from [40]. The training data extracted from each speaker contains all possible combinations of background noises and transients.

4.1.2. Feature selection

For the representation of the audio signal, we use MFCC[41], which represent the spectrum of speech in a compact form using the perceptually meaningful Mel-frequency scale. The MFCCs were found to perform well for voice activity detection under challenging conditions such as low SNR and non-stationary noise [32,42]. Each MFCC feature vector is composed of 12 cepstral coefficients, and their first and second derivatives, Δ and $\Delta\Delta$, respectively. Accordingly, the dimensions of the clean and contaminated audio feature vectors, $\mathbf{a}_n$ and $\tilde{\mathbf{a}}_n$, are $A = 36$. We note that by using Δ and $\Delta\Delta$ MFCCs, we incorporate temporal information into the process of learning the transient reducing representation. This allows for a better distinction between transients and speech, where the former typically vary faster over time. Even though the temporal information is also incorporated in the RNN, we found in our experiments that the use of Δ and $\Delta\Delta$ MFCCs further improves the detection results.

For the representation of the video signal, we use motion vectors, calculated using the Lucas–Kanade method [43,44]. Motion vectors are suitable for speech-related tasks since they capture both spatial and temporal information, i.e., the movement of the mouth, and they were previously exploited for voice activity detection in [25]. The feature representation, $\mathbf{v}_n$, is obtained by concatenating the absolute values of the velocities of each pixel from 3 consecutive frames $n-1, n, n+1$, so that its dimension is $V = 297$. We refer the reader to [32] for more details on the construction of the dataset and the audio-visual features.

4.1.3. Training process

The concatenated feature vector $\tilde{\mathbf{z}}_n$, of size $A+V = 333$, is fed as input to the transient-reducing audio-visual auto-encoder. The entries of $\tilde{\mathbf{z}}_n$ are normalized over the training set such that they have zero mean and unit variance in order to prevent saturation of the auto-encoder's neurons. We use an auto-encoder architecture with $L = 2$ hidden layers containing 200 neurons each, and with a logistic sigmoid activation function. During the training of the second hidden layer, in which we can no longer use the clean and contaminated signals for training, we contaminate the input for that layer with Gaussian noise with zero mean and variance 0.05 as described in Section 3.2.

The input layer of the RNN has 200 neurons, matching the output of the transient-reducing audio-visual auto-encoder, $\mathbf{p}_n$. The RNN comprises $\tilde{L} = 3$ hidden layers with 50,50, and 30 neurons, activated with a logistic sigmoid function, so that the full system architecture is of the form 333($\tilde{\mathbf{z}}_n$)-200($\mathbf{h}_n^1$)-200($\mathbf{p}_n$)-50($\tilde{\mathbf{h}}_n^1$)-50($\tilde{\mathbf{h}}_n^2$)-30($\tilde{\mathbf{h}}_n^3$)-1($\bar{\mathbf{y}}_n$).We used a sigmoid activation function in order to constrain the output of the entire network to be in the range [0, 1] so that it can be used as a probability measure for speech presence. For consistency, we also use the sigmoid for the activation of the hidden layers, and note that we found in our experiments that it performs similarly to the widely used ReLU [45]. We train the RNN layers in a supervised end-to-end manner using back propagation through time [36], and the whole system is optimized with gradient descent with a learning rate of 10^{-5} and momentum 0.9. All of the weights are initialized with values from a random normal distribution with zero mean and variance 0.01.

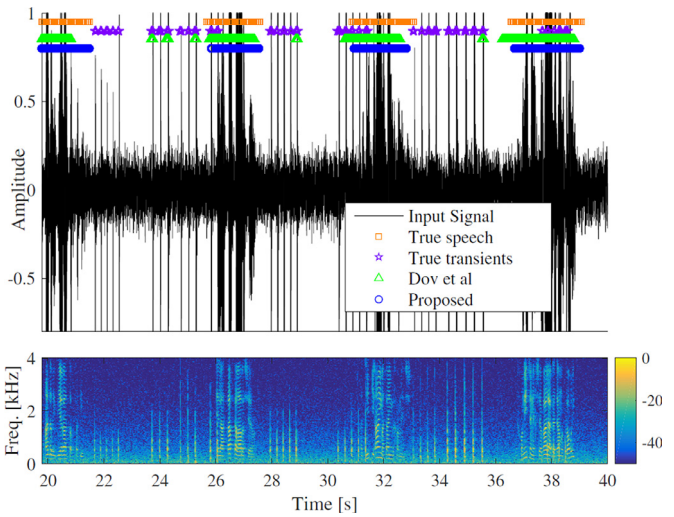

Fig. 1. Example of voice activity detection. Acoustic environment: colored Gaussian noise with 10 dB SNR and hammering transient interferences. (Top) Time domain, input signal – black solid line, true speech- orange squares, true transients – purple stars, competing method [32] with a threshold set for 90% correct detection rate – green triangles, proposed deep architecture with a threshold set for 90% correct detection rate – blue circles. (Bottom) Spectrogram of the input signal. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To prevent over-fitting, we use the early stopping procedure [46]; specifically, we use 30% of the training data as a validation set, on which we evaluate the network once every 5 epochs. The training procedure is stopped when the loss function of the network stops improving, and specifically, when 5 consecutive increases in validation error are obtained. To further increase robustness against convergence into suboptimal local minima, we train three realizations of the same network with different random weight initializations and average the predictions of the network over all realizations. The training time of all three realizations of the network took about 8 hours on an ordinary desktop computer.

4.2. Evaluation

To evaluate the performance of the proposed deep architecture, we compare it to the audio-visual voice activity detectors presented in [32] and [28], which are denoted in the plots by "Dov AV" and "Tamura", respectively. In Fig. 1 we present an example of speech detection in the presence of hammering transient. The performance of the proposed deep architecture is compared to the algorithm presented in [32] by setting the threshold value t in (7) to provide 90% correct detection rate, and comparing their false alarm rates. Fig. 1 shows that the proposed architecture yields significantly fewer false alarms compared to the competing detector, where the latter wrongly detects transients as speech, e.g., in seconds 33 – 36.

In Figs. 2–4 we compare the different algorithms in the form of receiver operating characteristic (ROC) curves, which present the probability of detection versus the probability of false alarm. The ROC curves are generated by spanning the threshold in (7) over all values between zero and one. Moreover, the maximal performance of each method for different acoustic environments is presented in Table 1. They are obtained using a threshold value that provides the best results in terms of true positive (TP) rate plus true negative (TN) rate.

In order to further demonstrate the benefit of the fusion of the audio and video signals for voice activity detection, we evaluated single modal versions of the proposed architecture based only on the audio or video modalities. The single modal versions are de-

Table 1
Comparison of different VADs in terms of TP + TN. The best result in each column is highlighted in bold fonts.

	Babble 10 dB SNR Keyboard	Musical 10 dB SNR Hammering	Colored 5 dB SNR Hammering	Musical 0 dB SNR Keyboard	Babble 15 dB SNR Scissors
Tamura	73.6	83.8	83.9	73.8	81.2
Dov - Audio	87.7	89.9	87.8	86.5	90.2
Dov - Video	89.6	89.6	89.6	89.6	89.6
Dov - AV	92.9	94.5	92.8	92.9	94.6
Proposed - AV	**95.8**	**95.4**	**95.9**	**95.1**	**97.2**

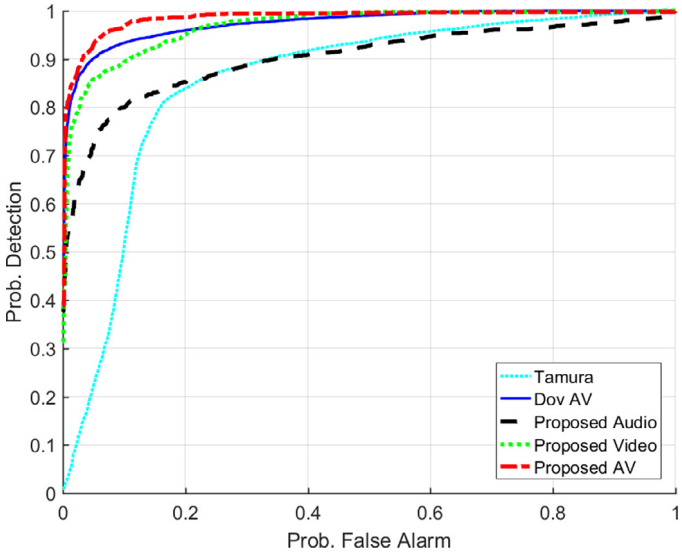

Fig. 2. Probability of detection versus probability of false alarm. Acoustic environment: Musical noise with 10 dB SNR and hammering transient interferences (best viewed in color).

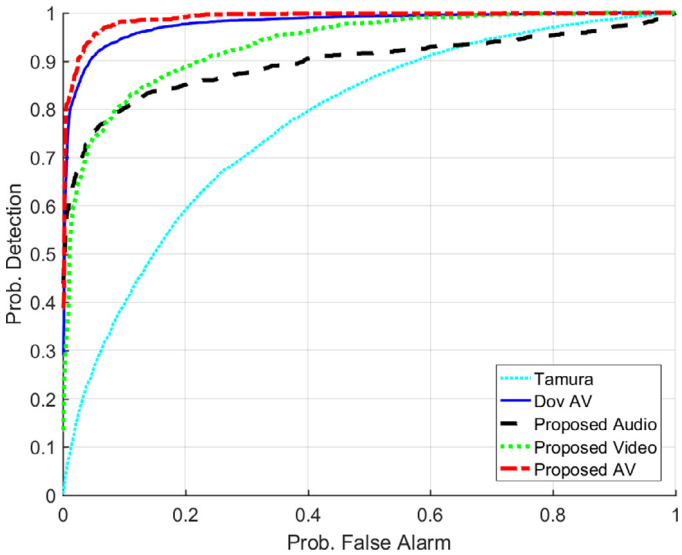

Fig. 3. Probability of detection versus probability of false alarm. Acoustic environment: Babble noise with 10 dB SNR and keyboard transient interferences (best viewed in color).

noted in the plots by "Proposed Audio" and "Proposed Video", respectively. When tested in a single modal version, the proposed deep architecture is fed only with features from one modality, after making proper changes to the input layer size. Then, the entire network is trained as described in Section 3. The benefit of fusing the audio and the video modalities is clearly shown in Figs. 2–4, where the proposed audio-visual architecture significantly outperforms the single modal versions. Also, the proposed deep architecture outperforms the audio-visual methods presented in [28] and [32].

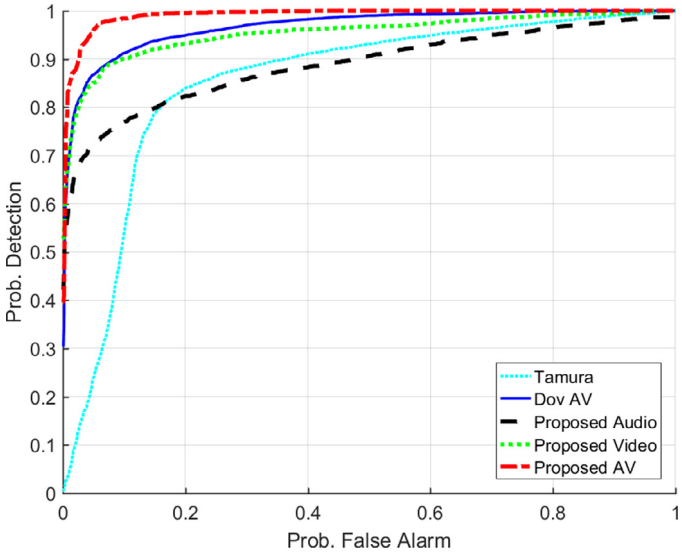

Fig. 4. Probability of detection versus probability of false alarm. Acoustic environment: Colored noise with 5 dB SNR and hammering transient interferences (best viewed in color).

In contrast to [32], where the modalities are merged only at the decision level, the proposed architecture exploits complex relations between the modalities learned by the transient-reducing auto-encoder. Moreover, in [32] the temporal context is only considered by concatenating features from a predefined number of consecutive frames, while in the proposed architecture the weights associated with previous frames are automatically learned by the supervised training process of the RNN, allowing for varying durations of temporal context to be exploited for voice activity detection.

5. Conclusions and future work

We have proposed a deep architecture for speech detection, based on specifically designed auto-encoders providing a new representation of the audio-visual signal, in which the effect of transients is reduced. The new representation is fed into a deep RNN, trained in a supervised manner to generate voice activity detection while exploiting the differences in the dynamics between speech and the transients. Experimental results have demonstrated that the proposed architecture outperforms competing state-of-the-art detectors providing accurate detections even under low SNR conditions and in the presence of challenging types of transients.

Future research directions include considering more complex variations of recurrent neural networks for the classification process. For example, bidirectional RNNs may be used to exploit the temporal context from future frames, and long short-term memory (LSTM) networks may facilitate learning even longer-term de-

pendencies between the inputs. Another next step is to perform a fine-tuning of the entire network from end to end in a supervised manner, while simultaneously updating the weights of the auto-encoder and the RNN via back propagation.

References

[1] D. Dov, I. Cohen, Voice activity detection in presence of transients using the scattering transform, in: Proc. 28th Convention of the Electrical & Electronics Engineers in Israel (IEEEI), IEEE, 2014, pp. 1–5.

[2] D. Dov, R. Talmon, I. Cohen, Kernel method for voice activity detection in the presence of transients, IEEE/ACM Trans. Audio, Speech Lang. Process. 24 (12) (2016) 2313–2326.

[3] D.A. Krubsack, R.J. Niederjohn, An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech, IEEE Trans. Signal Process. 39 (2) (1991) 319–329.

[4] J.-C. Junqua, B. Mak, B. Reaves, A robust algorithm for word boundary detection in the presence of noise, IEEE Trans. Speech Audio Process. 2 (3) (1994) 406–412.

[5] S. Van Gerven, F. Xie, A comparative study of speech detection methods, in: Proc. 5th European Conference on Speech Communication and Technology (EUROSPEECH), 1997, pp. 1095–1098.

[6] N. Cho, E.-K. Kim, Enhanced voice activity detection using acoustic event detection and classification, IEEE Trans. Consumer Electron. 57 (1) (2011) 196–202.

[7] J.-H. Chang, N.S. Kim, S.K. Mitra, Voice activity detection based on multiple statistical models, IEEE Trans. Signal Process. 54 (6) (2006) 1965–1976.

[8] J. Sohn, N.S. Kim, W. Sung, A statistical model-based voice activity detection, IEEE Signal Process. Lett. 6 (1) (1999) 1–3.

[9] J. Ramírez, J.C. Segura, C. Benítez, A. De La Torre, A. Rubio, Efficient voice activity detection algorithms using long-term speech information, Speech Commun. 42 (3) (2004) 271–287.

[10] J.W. Shin, J.-H. Chang, N.S. Kim, Voice activity detection based on statistical models and machine learning approaches, Comput. Speech Lang. 24 (3) (2010) 515–530.

[11] J. Wu, X.-L. Zhang, Maximum margin clustering based statistical VAD with multiple observation compound feature, IEEE Signal Process. Lett. 18 (5) (2011) 283–286.

[12] X.-L. Zhang, J. Wu, Deep belief networks based voice activity detection, IEEE Trans. Audio, Speech Lang. Process. 21 (4) (2013) 697–710.

[13] V.S. Mendelev, T.N. Prisyach, A.A. Prudnikov, Robust voice activity detection with deep maxout neural networks, Mod. Appl. Sci. 9 (8) (2015) 153.

[14] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[15] S. Leglaive, R. Hennequin, R. Badeau, Singing voice detection with deep recurrent neural networks, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 121–125.

[16] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6645–6649.

[17] T. Hughes, K. Mierle, Recurrent neural networks for voice activity detection, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 7378–7382.

[18] W.-T. Hong, C.-C. Lee, Voice activity detection based on noise-immunity recurrent neural networks, Int. J. Adv. Comput. Technol. (IJACT) 5 (5) (2013) 338–345.

[19] D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz, C. Jutten, An analysis of visual speech information applied to voice activity detection, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1, 2006, pp. 601–604.

[20] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, C. Jutten, A study of lip movements during spontaneous dialog and its application to voice activity detection, J. Acoustical Soc. Am. 125 (2) (2009) 1184–1196.

[21] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, C. Jutten, Two novel visual voice activity detectors based on appearance models and retinal filtering, in: Proc. 15th European Signal Processing Conference (EUSIPCO), 2007, pp. 2409–2413.

[22] E.-J. Ong, R. Bowden, Robust lip-tracking using rigid flocks of selected linear predictors, in: Proc. 8th IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2008.

[23] Q. Liu, W. Wang, P. Jackson, A visual voice activity detection method with adaboosting, in: in Proc. Sensor Signal Processing for Defence (SSPD), 2011, pp. 1–5.

[24] S. Siatras, N. Nikolaidis, M. Krinidis, I. Pitas, Visual lip activity detection and speaker detection using mouth region intensities, IEEE Trans. Circuits Syst. Vid. Technol. 19 (1) (2009) 133–137.

[25] A. Aubrey, Y. Hicks, J. Chambers, Visual voice activity detection with optical flow, IET Image Process. 4 (6) (2010) 463–472.

[26] P. Tiawongsombat, M.-H. Jeong, J.-S. Yun, B.-J. You, S.-R. Oh, Robust visual speakingness detection using bi-level HMM, Pattern Recognit. 45 (2) (2012) 783–793.

[27] P.K. Atrey, M.A. Hossain, A. El Saddik, M.S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimedia Syst. 16 (6) (2010) 345–379.

[28] S. Tamura, M. Ishikawa, T. Hashiba, S. Takeuchi, S. Hayamizu, A robust audio-visual speech recognition using audio-visual voice activity detection, in: Proc. 11th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2010, pp. 2694–2697.

[29] I. Almajai, B. Milner, Using audio-visual features for robust voice activity detection in clean and noisy speech, in: Proc. 16th European Signal Processing Conference (EUSIPCO), 2008.

[30] T. Yoshida, K. Nakadai, H.G. Okuno, An improvement in audio-visual voice activity detection for automatic speech recognition, in: Proc. 23rd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2010, pp. 51–61.

[31] V.P. Minotto, C.B. Lopes, J. Scharcanski, C.R. Jung, B. Lee, Audiovisual voice activity detection based on microphone arrays and color information, IEEE J. Selected Topics Signal Process. 7 (1) (2013) 147–156.

[32] D. Dov, R. Talmon, I. Cohen, Audio-visual voice activity detection using diffusion maps, IEEE/ACM Trans. Audio, Speech Lang. Process. 23 (4) (2015) 732–745.

[33] R.R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmonic Anal. 21 (1) (2006) 5–30.

[34] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[35] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.

[36] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back--propagating errors, Nature 323 (1986) 533–536.

[37] Y. Bengio, Y. LeCun, Scaling learning algorithms towards AI, Large-Scale Kernel Mach. 34 (5) (2007) 1–41.

[38] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proc. 28th International Conference on Machine Learning (ICML-11), 2011, pp. 689–696.

[39] N. Srivastava, R.R. Salakhutdinov, Multimodal learning with deep Boltzmann machines, in: Advances in Neural Information Processing Systems, 2012, pp. 2222–2230.

[40] [Online]. Available: http://www.freesound.org.

[41] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoustics, Speech Signal Process. 28 (4) (1980) 357–366.

[42] S. Mousazadeh, I. Cohen, Voice activity detection in presence of transient noise using spectral clustering, IEEE Trans. Audio, Speech Lang. Process. 21 (6) (2013) 1261–1271.

[43] J.L. Barron, D.J. Fleet, S.S. Beauchemin, Performance of optical flow techniques, Int. J. Comput. Vis. 12 (1) (1994) 43–77.

[44] A. Bruhn, J. Weickert, C. Schnörr, Lucas/kanade meets horn/schunck: combining local and global optic flow methods, Int. J. Comput. Vis. 61 (3) (2005) 211–231.

[45] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proc. 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814.

[46] L. Prechelt, Early stopping-but when? in: G. Montavon, G. B. Orr and K.-R. Müller (Eds.), Neural Networks: Tricks of the Trade, Springer, 2012, pp. 53–67.

Chapter 4

An End-to-End Multimodal Voice Activity Detection Using WaveNet Encoder and Residual Networks

An End-to-End Multimodal Voice Activity Detection Using WaveNet Encoder and Residual Networks

Ido Ariav and Israel Cohen, *Fellow, IEEE*

***Abstract*—Recently, there has been growing use of deep neural networks in many modern speech-based systems such as speaker recognition, speech enhancement, and emotion recognition. Inspired by this success, we propose to address the task of voice activity detection (VAD) by incorporating auditory and visual modalities into an end-to-end deep neural network. We evaluate our proposed system in challenging acoustic environments including high levels of noise and transients, which are common in real-life scenarios. Our multimodal setting includes a speech signal captured by a microphone and a corresponding video signal capturing the speaker's mouth region. Under such difficult conditions, robust features need to be extracted from both modalities in order for the system to accurately distinguish between speech and noise. For this purpose, we utilize a deep residual network, to extract features from the video signal, while for the audio modality, we employ a variant of WaveNet encoder for feature extraction. The features from both modalities are fused using multimodal compact bilinear pooling to form a joint representation of the speech signal. To further encode the temporal information, we feed the fused signal to a long short-term memory network and the system is then trained in an end-to-end supervised fashion. Experimental results demonstrate the improved performance of the proposed end-to-end multimodal architecture compared to unimodal variants for VAD. Upon the publication of this paper, we will make the implementation of our proposed models publicly available at https://github.com/iariav/End-to-End-VAD and https://israelcohen.com.**

***Index Terms*—Audio-visual speech processing, voice activity detection, deep neural networks, WaveNet.**

I. Introduction

VOICE activity detection constitutes an essential part of many modern speech-based systems, and its applications can be found in various domains. A partial list of such domains includes speech and speaker recognition, speech enhancement, dominant speaker identification, and hearing-improvement devices. In many cases, voice activity detection is used as a preliminary block to separate the segments of the signal that contain speech from those that contain only noise and interferences, thus enabling the overall system to, e.g., perform speech recognition only on speech segments, or change the noise reduction method between speech/noise segments.

Traditional methods of voice activity detection mostly rely on the assumption of quasi-stationary noise, i.e., the noise spectrum changes at a much lower rate than the speech signal. One group of such methods are those based on simple acoustic features such as zero-crossing rate and energy values in short time intervals [1]–[3]. More advanced methods are model-based methods that focus on estimating a statistical model for the noisy signal [4]–[7]. The performance of such methods usually significantly deteriorates in the presence of even moderate levels of noise. Moreover, they cannot correctly model highly non-stationary noise and transient interferences, which are common in real life scenarios and are within the main scope of this study, since the spectrum of transients, similarly to the spectrum of speech, often rapidly varies over time [8].

Apart from the statistical approaches noted above, more recent methods have been developed using machine learning techniques [9], [10], and more specifically, using deep neural networks. In recent years, deep neural networks achieved groundbreaking improvements on several pattern recognition benchmarks in areas such as image classification [11], speech and speaker recognition [12] and even multimodal tasks such as visual question answering [13]. Deep networks were successfully used to extract useful signal representations from raw data, and more specifically, several studies have also shown the favorable property of deep networks to model the inherent structure contained in the speech signal [14]. Deep neural networks were recently utilized in several modern voice activity detectors; Zhang and Wu [15] proposed to extract a set of predefined acoustic features, e.g., Mel Frequency Cepstral Coefficients (MFCC), from a speech signal and then feed these features to a deep-belief network (DBN) in order to obtain a more meaningful representation of the signal. They then used a linear classifier to perform speech detection. Thomas *et al.* [16] fed a convolutional neural network (CNN) with the log-Mel spectrogram together with its delta and delta-delta coefficients and trained the CNN to perform voice activity detection.

Despite showing improved performance compared to traditional methods, these networks classify each time frame independently, thus ignoring existing temporal relations between consecutive time frames. To alleviate this issue, several studies have suggested methods for modeling temporal relations between consecutive time frames [17]. More modern methods rely on recurrent neural networks (RNN) to incorporate previous inputs into the classification process, thus utilizing the

Manuscript received October 9, 2018; revised January 4, 2019 and February 12, 2019; accepted February 13, 2019. Date of publication February 22, 2019; date of current version May 16, 2019. This work was supported in part by the Israel Science Foundation under Grant 576/16, and in part by the ISF-NSFC joint research program under Grant 2514/17. The guest editor coordinating the review of this paper and approving it for publication was Prof. Juhan Nam. *(Corresponding author: Ido Ariav.)*

The authors are with the Andrew and Erna Viterbi Faculty of Electrical Engineering, The Technion-Israel Institute of Technology, Haifa 3200003, Israel (e-mail: idoariav@tx.technion.ac.il; icohen@ee.technion.ac.il).

Digital Object Identifier 10.1109/JSTSP.2019.2901195

signal's temporal information [18]–[20]. Hughes and Mierle [21] extracted Perceptual Linear Prediction (PLP) features from a speech signal and fed them to a multi-layered RNN with quadratic polynomial nodes to perform speech detection. Lim *et al.* [22] proposed to transform the speech signal using a short-time Fourier transform and use a CNN to extract high-level representation for the signal. This new representation was then fed to an LSTM to exploit the temporal structure of the data. These methods however still mostly rely on hand-crafted audio features, and often misclassify frames that contain both speech and transients as non-speech frames, since transients often appear more dominant than speech.

Although most of the current work on voice activity detection concentrates around the subject's audio signal, recent methods proposed to make use of other modalities, such as visual information, to improve the voice detection [23]–[31]. The advantages of using a multimodal setting are most prominent when dealing with demanding acoustic environments, where high levels of acoustic noise and transient interferences are present since the video signal is entirely invariant to the acoustic environment. Therefore, proper incorporation of the video signal can significantly improve voice detection, as we show in this paper. Ngiam *et al.* [32] proposed a Multimodal Deep Autoencoder for feature extraction from audio and video modalities. A bimodal DBN was used for the initialization of the deep autoencoder, and then the autoencoder was fine-tuned to minimize the reconstruction error of the two modalities. Zhang *et al.* [33] used a CNN for classifying emotions in a multimodal setting. They applied two separate CNNs, the first operating on the mel-spectrogram of an audio signal and the second on a video recording of a subject's face. The features extracted using the two CNNs were concatenated and fed to a deep fully connected neural network to perform the classification. Dov *et al.* [34] proposed to obtain separate low dimensional representations of the audio and video signals using diffusion maps [35]. The two modalities were then fused by a combination of speech presence measures, which are based on spatial and temporal relations between samples of the signal in the low dimensional domain. In our previous work [36], we proposed a deep architecture comprised of a transient reducing autoencoder and an RNN for voice activity detection. Features were extracted from both modalities and fed to the transient reducing autoencoder which was trained to both reduce the effect of transients and merge the modalities. The output of the autoencoder was fed to an RNN that incorporated temporal data to the speech presence/absence classification.

The great majority of works described above still make use of commonly hand-crafted features in audio or visual modality. To alleviate the need for hand-crafted features, a few studies proposed to adopt an end-to-end approach and use the raw, unprocessed data as input while utilizing as little human apriori knowledge as possible [37]. The motivation behind this being that a deep network can ultimately automatically learn an intermediate representation of the raw input signal that better suits the task at hand which in turn leads to improved overall performance. Trigeorgis *et al.* [38] proposed an end-to-end model for emotion recognition from raw audio signal. A CNN was used to extract features from the raw signal which were then fed to an LSTM network in order to capture the temporal information in the data. Tzirakis *et al.* [39] recently proposed another multimodal end-to-end method for emotion recognition. Features were extracted separately from audio and video signals using two CNNs and then concatenated to form a joint representation, which in turn was fed to a multi-layered LSTM for classification. Hou *et al.* [40] proposed a multimodal end-to-end setting for speech enhancement. They used two CNNs to extract features from audio spectrograms and raw video, and these features were then concatenated and fed into several fully connected layers to produce an enhanced speech signal. This work, however, uses only simple concatenation to fuse the two modalities and does not utilize temporal information which we solve by incorporating LSTM. Petridis *et al.* [41] proposed an end-to-end approach for audio-visual speech recognition. They extracted features from each modality using a CNN and then fed these features to modality-specific RNN layers. The modalities were then fused by feeding the outputs of those RNNs to another RNN layer.

In this paper, we propose a deep end-to-end neural network architecture for audio-visual voice activity detection. First, we extract meaningful features from the raw audio and video signals; for the video signal we employ a ResNet-18 network [42] as a feature extractor, and for the audio signal we employ a variant of a WaveNet [43] encoder. Then, instead of merely concatenating the feature vectors extracted from the two modalities, as is common in most multimodal networks, we propose to fuse the two vectors into a new joint representation via a Multimodal Compact Bilinear (MCB) pooling [44] module, which was shown to efficiently and expressively combine multimodal features. The output of the MCB module is fed to several stacked LSTM layers in order to explore even further the temporal relations between samples of the speech signal in the new representation. Finally, a fully connected layer is used to perform the classification of each time-frame to speech/non-speech, and the entire network is trained in a supervised end-to-end manner. To the best of our knowledge, this is the first time that such an end-to-end approach is applied for voice activity detection.

The proposed deep end-to-end architecture is evaluated in the presence of highly non-stationary noises and transient interferences. Experimental results show the benefits of our multimodal end-to-end architecture compared to unimodal approaches, and the advantage of audio-video data fusion is thus demonstrated. Also, we demonstrate the effectiveness of the MCB module for modality fusion compared to a simple concatenation/multiplication of feature vectors.

The remainder of the paper is organized as follows. In Section II, we formulate the problem of voice activity detection and present our dataset. In Section III, we introduce the proposed multimodal end-to-end architecture. In Section IV, we demonstrate the performance of the proposed deep end-to-end architecture for voice activity detection. Finally, in Section V, we conclude and offer some directions for future research.

II. Dataset and Problem Formulation

A. Problem Formulation

Voice activity detection is a segmentation problem in which segments of a given speech signal are classified as sections that contain speech and sections that contain only noise and interferences. We consider a speech signal recorded via a single microphone and a video camera simultaneously, pointed at a front-facing speaker reading an article aloud.

Let $\mathbf{a} \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^{H \times W \times 3}$ be the audio and video signals, respectively, where H and W are the height and width of each video frame in pixels, respectively. For alignment of the two signals, we artificially divide $\mathbf{a}$ into frames of length M and denote each video/audio frame with subscript n.

We use the clean audio signal $\mathbf{a}$ to label each time frame n according to the presence or absence of speech, and then assign each frame in $\mathbf{a}$ and $\mathbf{v}$, with the appropriate label. Our proposed architecture performs a frame-by-frame classification for voice activity detection, however, as part of the classification process of each frame, we also take into account T past frames. For this reason, we construct from $\mathbf{a}$ and $\mathbf{v}$ a dataset of N overlapping sequences of audio and video by concatenating consecutive frames into sequences of length T. We denote by $\mathbf{S}_a^i$ and $\mathbf{S}_v^i$, respectively, the i^{th} audio and video sequences. Since we only use past frames in the classification process, each sequence is labeled according to the label of the last frame in the sequence. That is, the sequence $\mathbf{S}_v^i$ containing the video frames $\{\mathbf{v}_{i-14}, \mathbf{v}_{i-13}, \mathbf{v}_{i-12}, \ldots, \mathbf{v}_i\}$ is assigned the label given to $\mathbf{a}_i$, the i^{th} frame of the audio signal $\mathbf{a}$. Each $\mathbf{S}_a^i$ is then contaminated with background noises and transient interferences, as described in Section II-B.

The goal in this study is to classify each frame n as a speech or non-speech frame.

B. Dataset

We evaluate the proposed deep end-to-end architecture for voice activity detection using the dataset presented in [34], [36]. The dataset includes audio and video recordings of 11 speakers reading aloud an article. The speakers are instructed to make natural pauses every few sentences so that the intervals of speech and non-speech range from several hundred ms to several seconds in length. The video signal $\mathbf{v}$ comprises the mouth region of the speaker, cropped from the original recording. It is processed at 25 frames/s and each frame is 90×110 pixels in size. The audio signal is recorded at 8 kHz with an estimated SNR of $\sim$25 dB, and we artificially divide the signal to frames, where each frame contains $M = 320$ audio samples without overlap. Each of the 11 recordings is 120 seconds long. Each clean audio recording $\mathbf{a}$ is normalized to the range $[-1, 1]$ and each video recording $\mathbf{v}$ is normalized to have zero mean and standard deviation 1 in each of the three R, G, and B channels. We refer the reader to [34] for more details on the creation of the dataset.

We divide the audio and video recordings, $\mathbf{a}$ and $\mathbf{v}$, to overlapping sequences, $\mathbf{S}_a^i$, $\mathbf{S}_v^i$, of length T frames. This procedure produces an overall of $\sim 33,000$ such sequences. Out of the 11 speakers, we randomly select 8 speakers for the training set, and the other 3 speakers were used as an evaluation set. Finally, our training set contains $\sim 24,000$ audio/video sequences and the evaluation set contains $\sim 9,000$ sequences.

Algorithm 1: Inject Random Background Noise and Transient Interference.

▷ *Init noises and transients lists:*
Noises : {white Gaussian noise, musical instruments noise, babble noise, none}
Trans : {door-knocks, hammering, keyboard typing, metronome, scissors, none}

▷ *For each sequence* $\mathbf{S}_a^i$ *in the dataset:*
for $i = 1 \rightarrow \#sequences\ in\ dataset$ **do**
  $L \leftarrow$ Length of sequence $\mathbf{S}_a^i$ in frames
  ▷ *randomly choose noises to inject:*
  $noise \leftarrow$ uniformly random noise from "Noises"
  $trans \leftarrow$ uniformly random transient from "Trans"
  $SNR \leftarrow$ uniformly random value from [0,20]
  if $noise$ is not "none" **then**
    $R_{noise} \leftarrow$ random sequence of length L from $noise$
    $R_{noise} \leftarrow R_{noise}/std(R_{noise})$ ▷ update R_{noise}'s SNR
    $R_{noise} \leftarrow R_{noise} \times (std(\mathbf{S}_a^i)/(10^{(SNR/20)})$
    $\mathbf{S}_a^i \leftarrow \mathbf{S}_a^i + R_{noise}$
  if $trans$ is not "none" **then**
    $R_{trans} \leftarrow$ random sequence of length L from $trans$
    $\mathbf{S}_a^i \leftarrow \mathbf{S}_a^i + 2 \times R_{trans}$
return $\mathbf{S}_a^i$ as $\tilde{\mathbf{S}}_a^i$

During the construction of the dataset, we randomly add background noise and transient interferences to each clean audio sequence $\mathbf{S}_a^i$ in the evaluation set according to the following procedure outlined in Algorithm 1. First, background noise is randomly selected from one of {white Gaussian noise, musical instruments noise, babble noise, none} and a transient interference is randomly selected from one of {door-knocks, hammering, keyboard typing, metronome, scissors, none}, all taken from [45]. Once a background noise and transient were selected randomly, we randomly choose an SNR in the range [0, 20] and then $\mathbf{S}_a^i$ is contaminated with the selected noise and transient at the selected SNR. We denote the contaminated sequence as $\tilde{\mathbf{S}}_a^i$. This way, our evaluation set contains all possible combinations of background noises and transients at different SNR levels. For the training set we use a similar procedure for injecting noise and transients. However, instead of adding the noise only once during the construction of the dataset, we inject the noise in each iteration of the training process. This way, during the entire training process, a specific sequence $\mathbf{S}_a^i$ can be injected with different combinations of background noise, transient, and SNR level. Note that in addition to noise injection, augmenting the video signal or altering the speakers' voice signals according to

the injected noise levels (Lombard effect), can be applied to the same dataset, but this will be explored in future work.

It is worth noting that even though we use the same speech recordings as in [34], [36], the dataset we construct from them is somewhat different and significantly more challenging. In our experiments, each sample of the evaluation set contains a different mixture of background noise, transient, and SNR. In contrast, [34], [36] contaminated their evaluation set with only one background noise and one transient at a predefined SNR in each experiment.

III. Deep Multimodal End-To-End Architecture for Voice Activity Detection

Voice activity detection becomes even more challenging in the presence of transients, which are typically more dominant than speech due to their short duration, high amplitudes and fast variations of the spectrum [8]. Specifically, audio features extracted from frames that contain both speech and transients are often similar to features extracted from frames that contain only transients, so that they are often wrongly classified as non-speech frames. To address this challenge, we introduce a deep end-to-end neural network architecture, in which meaningful features are learned from raw audio and video data. The overall system is trained in an end-to-end manner to predict speech presence/absence, and hence the learned features are those that maximize the classification accuracy. We propose to extract such features from both video and audio using two variants of known neural networks, ResNet and WaveNet, respectively, which we will now review. An overview of our deep multimodal end-to-end architecture for voice activity detection can be seen in Fig. 1.

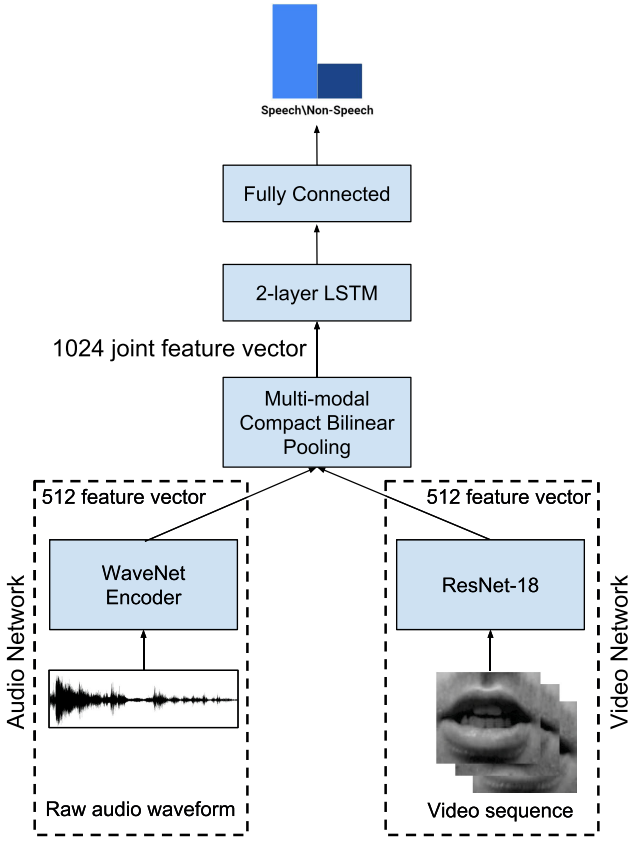

Fig. 1. Our proposed deep multimodal end-to-end architecture for voice activity detection.

A. Visual Network

In order to extract features from the raw video signal we use a deep residual network of 18 layers [42] denoted ResNet-18. Deep ResNets are formed by stacking several residual blocks of the form:

$$\mathbf{y}_k = F(\mathbf{x}_k, \mathbf{W}_k) + h(\mathbf{x}_k), \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the input and output of residual block k, F is the residual block's function to be learned, and $h(\mathbf{x}_k)$ is either an identity mapping or a linear projection so that the dimensions of function F and the input $\mathbf{x}$ will match. The first layer of a ResNet-18 model is a 7×7 convolutional layer with 64 feature maps, and it is followed by a 3×3 max pooling layer. These two layers are followed by four residual blocks, where after each residual block a shortcut connection is added. Each residual block contains two convolutional layers of sizes 3×3, and the outputs of these residual blocks contain 64, 128, 256 and 512 feature maps respectively. After the last residual block, an average pooling layer is inserted followed by a fully connected layer performing the classification.

In order to use the ResNet-18 model to generate an embedding for the video sequence $\mathbf{S}_v^i$, we drop the last fully connected layer of the model and use the output of the average pooling layer as the video embedding which we denote $\mathbf{Z}_v$. The average pooling layer has 512 neurons, and thus the size of the produced embedding $\mathbf{Z}_v$ in feature space is 512 in length, for each frame of $\mathbf{S}_v^i$. $\mathbf{Z}_v$ has a temporal size of T, to match the temporal length of $\mathbf{S}_v^i$ so that the overall size of $\mathbf{Z}_v$ is $T \times 512$. In order to avoid feeding each image in the sequence $\mathbf{S}_v^i$ to the network separately in a serial fashion, we concatenate all images into one batch, with T images, and feed this batch of images to the video network. The network then operates on all images in a parallel fashion, which provides substantial reduction of computation time.

It is worth noting that we also experimented with deeper residual models such as ResNet-50 and ResNet-101. However, they showed no improvement regarding the task of voice activity detection and some even experienced degradation in performance. This degradation can perhaps be explained by over-fitting of the model since these very deep models are usually used for tasks with hundreds or even thousands of classes whereas voice activity detection is a binary classification problem. These networks also produce longer embeddings, i.e., 2048 in size. We thus opted to use a shallower model with 18 layers, which also significantly reduces the memory consumption and computational load of the overall network.

B. Audio Network

In contrast to most previous machine learning works in audio processing, in which the first step is to extract hand-crafted

features from the data, we propose to learn the feature extraction and classification steps in one jointly trained model for voice activity detection. The input to our audio network is the sequence of noisy raw audio signal frames $\tilde{\mathbf{S}}_a^i$. The feature extraction from the raw audio signal is performed by a WaveNet encoder, comprised of stacked residual blocks of dilated convolutions, which exhibit very large receptive fields, and which we will now review. The WaveNet encoder is designed in such a manner that enables it to better deal with long-range temporal dependencies that exist in the audio signal, than ordinary CNN or feed-forward networks.

1) Dilated Convolution: Convolutions are one of the main building blocks of modern neural networks. A convolution layer's parameters consist of a set of learnable filters $\{\mathbf{K}_q\}$, that have the same number of dimensions as the input they are convolved with. During the forward pass, these filters are convolved with the input, computing the dot product between the entries of the filter and the input to produce a new feature map. In most modern networks, relatively small filters are often used, thus each entry of the feature map has only a small receptive field of the input. In order to increase this receptive field, larger filters can be used, or many layers of small filters can be stacked, both at the expense of more burdensome computations. Here, we opted to use dilated convolutions to increase the receptive field of each feature map entry, without substantially increasing the computational cost.

In a dilated convolution, the filter $\mathbf{K}_q$ is applied over an area that is larger than its size by skipping input values with a predetermined dilation factor. E.g, in the 2-D case, a convolution operation between a 3×3 filter $\mathbf{K}_q$ with dilation factor of 2 and a 2-D input volume $\mathbf{I}$ at location (p, o) is given by:

$$(\mathbf{I} * \mathbf{K}_q)(p, o) = \sum_{l=-1}^{1} \sum_{m=-1}^{1} \mathbf{I}(p - 2\,l, o - 2\,m) \mathbf{K}_q(l + 1, m + 1) \quad (2)$$

where $*$ denotes the convolution operator. By skipping input values, a dilated convolution effectively operates on a larger receptive field than a standard convolution.

In order to increase the receptive field of all feature maps' entries even further, without increasing the computational load, several dilated convolutions can be stacked [46]. Notably, a block constructed of 10 dilated convolutions with exponentially increasing dilation factors of 1, 2, 4, . . . , 512, has a receptive field of size 1024 and can be considered a more efficient and discriminative non-linear counterpart of a 1×1024 regular convolution layer. We utilize this property in our implementation as described in Section III-B2.

2) WaveNet Encoder: WaveNet [43] is a powerful generative approach to probabilistic modeling of raw audio. Recalling the original WaveNet architecture described in [43], a WaveNet network is a fully convolutional neural network constructed by stacked blocks of dilated convolutions, where the convolutional layers in each block have exponentially growing dilation factors that allow the receptive field to also grow exponentially with depth and cover thousands of time-steps. A WaveNet model

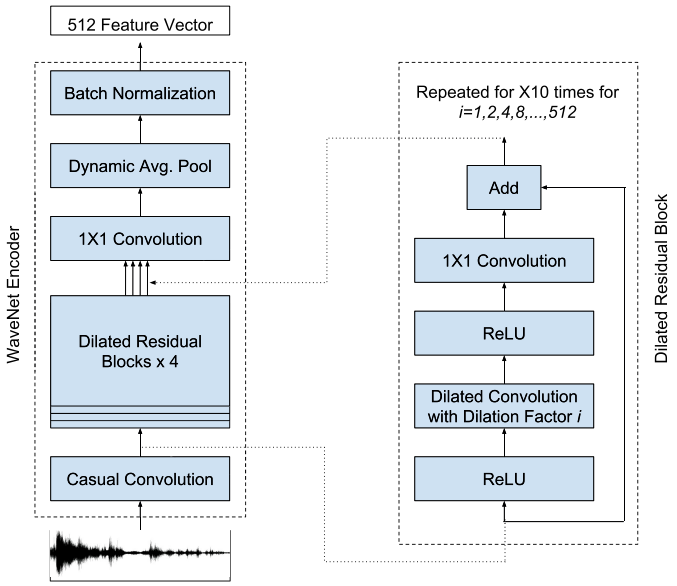

Fig. 2. Left - our WaveNet encoder architecture. Right - the structure of a dilated residual block.

makes use of both residual [42] and parameterized skip connections throughout the network, which facilitates faster convergence of the model and enables the training of much deeper models. For more details on the original WaveNet architecture, we refer the readers to [43].

A WaveNet network is trained to predict the next sample of audio from a fixed-size input of prior sample values. In this paper, we use similar building blocks as in WaveNet to construct a WaveNet encoder that operates on a raw audio signal and, instead of predicting the next sample, it produces an embedding $\mathbf{Z}_a$ for each $\tilde{\mathbf{S}}_a^i$. This embedding $\mathbf{Z}_a$ is then used as the feature vector for the given audio sequence.

Our implementation of a WaveNet encoder consists of a causal convolution layer followed by four stacked residual blocks, where each block is constructed by stacking 10 layers of dilated convolutions with exponentially growing dilation factors, ranging from 1 to 512 in each block. We found in our experiments that setting a fixed size of 32 channels for all convolution layers allows the model to be expressive enough while not making the network unnecessarily large. The output of all residual blocks is then aggregated and fed into a 1-D regular convolution with a kernel size of 1 and 512 channels so that the final dimension of $\mathbf{Z}_a$ in feature space is 512. We then finally apply an adaptive 1-D average pooling which operates on the temporal dimension, in order to further aggregate the activations of all residual blocks, and so that the temporal length of the embedding $\mathbf{Z}_a$ will be T, to match that of the video signal. Thus, the final size of $\mathbf{Z}_a$ is $T \times 512$. Throughout the network, we use 1-D filters of length 2 for all regular and dilated convolutions, and each convolution layer precedes a ReLU nonlinearity [47]. In Fig. 2 we show our WaveNet encoder overall architecture and the dilated residual block, which is stacked several times in the network.

Note that in contrary to the original WaveNet, we do not use μ-law compounding transformation to quantize the input,

and instead, we operate on the raw 1-D signal. Moreover, early experiments did not show any noticeable advantage to the non-linearity used in [43], comprised of the element-wise multiplication of a tanh and sigmoid functions, over ReLU, so we opted to use the latter.

C. Multimodal Compact Bilinear Pooling

Once the embeddings for the audio and video signals, $\mathbf{Z}_a$ and $\mathbf{Z}_v$, are obtained, via the audio and video networks respectively, both embeddings are fused to form a joint representation that is then fed to the classification layers. Here, we propose to rely on multimodal compact bilinear pooling (MCB) to obtain this joint representation, rather than on simple concatenation of the embeddings.

Bilinear pooling is a fusion method for two vectors, $\mathbf{f} \in \mathbb{R}^{B_1}$ and $\mathbf{g} \in \mathbb{R}^{B_2}$, in which the joint representation is simply the outer product between the two vectors. This allows, in contrast to an element-wise product or simple concatenation, a multiplicative interaction between all elements of both vectors. Bilinear pooling models [48] have recently been used for fine-grained classification tasks [49]. However, their main drawback is their high dimensionality of $B_1 \times B_2$, which can be very large in most real-life cases and leads to an infeasible number of learnable parameters. For example, in our paper B_1 and B_2 are the lengths of the embeddings $\mathbf{Z}_a$ and $\mathbf{Z}_v$ in feature space, namely $B_1 = B_2 = 512$, which results in a joint representation of length 512^2. This inevitably leads to very high memory consumption and high computation times and can also result in over-fitting.

Here, we adopt ideas from [50] to the multimodal case for audio and visual modalities, and use MCB to fuse the two embeddings $\mathbf{Z}_a$ and $\mathbf{Z}_v$. As discussed in detail in [50], the multimodal compact bilinear pooling is approximated by projecting the joint outer product to a lower dimensional space, while also avoiding computing the outer product directly. This is accomplished via the count sketch projection function suggested in [51], denoted as Ψ, which is a method of projecting a vector $\mathbf{c} \in \mathbb{R}^m$ to a lower dimensional representation $\hat{\mathbf{c}} \in \mathbb{R}^d$ for $d < m$. Instead of applying Ψ directly on the outer product of the two embeddings, we follow [52] which showed that explicitly computing the outer product of the two vectors can be avoided since the count sketch of the outer product can be expressed as a convolution of both count sketches. Additionally, the convolution theorem states that a circular convolution of two discrete signals can be performed with lower asymptotic complexity by performing multiplication in the frequency domain (note that linear convolution in the time-domain may be replaced with a circular convolution by applying a proper zero-padding to the time-domain signals). Therefore, we project both embeddings $\mathbf{Z}_a$ and $\mathbf{Z}_v$ separately to a lower dimensional representation using Ψ and then calculate element-wise products of fast Fourier transforms (FFT) of the lower dimensional representations.

We apply MCB to $\mathbf{Z}_a$ and $\mathbf{Z}_v$ to form a joint representation of the two modalities, denoted $\mathbf{Z}_{av}$. We choose the MCB output size to be 1024 in feature space, and its temporal size is T so that the final size of $\mathbf{Z}_{av}$ is $T \times 1024$. We then apply batch normalization before feeding $\mathbf{Z}_{av}$ as input to the classification layers detailed in Section III-D. We note that MCB can easily be extended and remains effective for more than two modalities as the fusion of the modalities is achieved by element-wise product of FFTs. Another advantage to using MCB for vector fusion is being able to choose the desired size for the joint vector. In contrast, when feature vectors are fused by simple concatenation, element-wise multiplication or dot product, the joint representation is given by the sizes of the feature vectors. In MCB, the size of the joint representation can be selected according to performance or computational requirements.

In Section IV we demonstrate the effectiveness of fusing the two modalities with MCB compared to a simple concatenation of the embeddings.

TABLE I
TENSOR NOTATIONS AND SIZES OF OUR FINAL MULTIMODAL NETWORK

Notation	Dimensions	Description
$\tilde{\mathbf{S}}_a^i$	$\mathbb{R}^{(T\times 320)\times 1}$	noisy sequence of T audio frames
$\mathbf{S}_v^i$	$\mathbb{R}^{T\times 90\times 110\times 3}$	sequence of T video frames
$\mathbf{Z}_a$	$\mathbb{R}^{T\times 512}$	embedding of audio sequence
$\mathbf{Z}_v$	$\mathbb{R}^{T\times 512}$	embedding of video sequence
$\mathbf{Z}_{av}$	$\mathbb{R}^{T\times 1024}$	joint embedding produced by MCB
$\mathbf{Y}_{av}$	$\mathbb{R}^{1\times 1024}$	the last temporal output of the last LSTM layer
$\mathbf{Out}_{av}$	$\mathbb{R}^{1\times 1}$	the final network's output representing speech presence/absence

The first dimension is always the temporal dimension, and the rest are dimensions in data/feature space.

D. Classification Layers

The joint embedding $\mathbf{Z}_{av}$ produced from the MCB module is fed to an LSTM block with two layers, each with 1024 cells, to explore even more temporal information embedded in the speech signal. We follow a "many-to-one" approach and feed only the last temporal response of the last LSTM layer, denoted $\mathbf{Y}_{av}$, to a fully connected layer with 1024 neurons, to match the size of the last LSTM layer. This fully conected layer is followed by another fully connected layer with just one neuron representing the output of the whole network. We apply a sigmoid activation function, so that the output of that final layer, denoted $\mathbf{Out}_{av}$, is constrained to the range of 0–1 and can be considered as the probability for speech absence/presence.

For regularization purposes, and to avoide over-fitting, due to the large number of parameters in the network compared to the number of training examples, we use dropout [53] at several points in our network. We use a dropout with probability $p = 0.5$ at the last fully connected layer and dropout with probability $p = 0.2$ before and after the MCB module. We also use batch normalization on the outputs of the audio and video networks and the MCB module's output.

We summarize all the tensors in the final implementation of our multimodal end-to-end network and their sizes in Table I.

IV. Experimental Results

A. Training Process

1) Unimodal Training: Prior to training our deep multimodal end-to-end architecture, we train two unimodal variants of our architecture, for audio and video. This allows us to initialize our multimodal network with weights from the learned unimodal networks, which enables the network to converge to a better minima while also improving convergence speed. To construct the audio unimodal network, we remove the MCB module from our multimodal architecture and merely feed the audio embedding $\mathbf{Z}_a$ directly to the classification block as described in Section III-D. Similarly, we construct the video network by removing the MCB and feeding the video embedding $\mathbf{Z}_v$ directly to the classification block.

For the training of the visual network, we initialize all the weights with values from a random normal distribution with zero mean and variance 0.01, including the weights of the ResNet-18 model. In our experiments, we found this leads to better results than initializing the ResNet-18 from a model that was pre-trained on, e.g., ImageNet [11]. We feed the network with sequences of $T = 15$ video frames, and they are all processed in parallel as discussed in Section III-A. The produced embedding, $\mathbf{Z}_v \in \mathbb{R}^{15\times512}$, is fed directly to the classification layers.

Similarly to the visual network, we initialize all the weights of the audio network with values from a random normal distribution with zero mean and variance 0.01. We feed the WaveNet encoder with sequences $\tilde{\mathbf{S}}_a^i$ of $T = 15$ raw audio frames, which corresponds to 4800 audio samples or 0.6 seconds of audio. The WaveNet encoder produces the embedding $\mathbf{Z}_a \in \mathbb{R}^{15\times512}$. As in the visual network, we feed $\mathbf{Z}_a$ directly to the classification layers.

We train each network separately for 150 epochs using stochastic gradient descent (SGD) with weight decay of 10^{-4} and momentum 0.9. We use an initial learning rate of 0.01 and divide this learning rate by a factor of 10 every 30 epochs. For the audio network we use a mini-batch of 96 samples and for the visual network a mini-batch of size 16. As described in Section III-D, we use dropout on the outputs of the WaveNet encoder and the ResNet-18 model and also on the last fully connected layer for regularization of the network.

2) Multimodal Training: Once both unimodal networks are trained, the classification block of each unimodal network is discarded, and we use the learned weights of the WaveNet encoder and ResNet-18 modules to initialize the corresponding parts of the multimodal network. We use MCB with an output size of 1024 to fuse the feature vectors extracted from both modalities. This joint representation is fed to a similar classification block used in the unimodal variants. The LSTM and fully connected layers in the classification block are initialized with random weights from a random normal distribution with zero mean and variance 0.01. The entire network is trained in an end-to-end manner, and the visual and speech networks are fine-tuned. The multimodal network is trained for an additional 50 epochs using SGD with weight decay of 10^{-4}, momentum 0.9 and a fixed learning rate of 0.001.

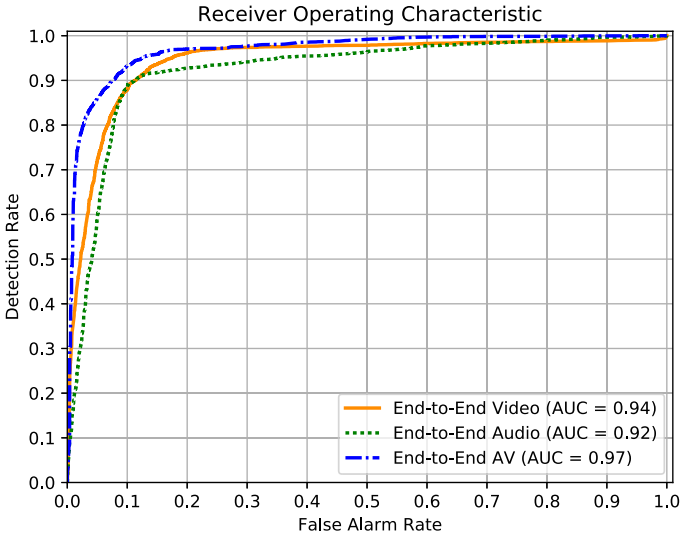

Fig. 3. Probability of detection versus probability of false alarm of our unimodal and multimodal architectures (best viewed in color).

For further regularization of the network, and in order to avoid the exploding gradients problem which can arise in LSTM cells, we enforce a hard constraint on the norm of the gradients by scaling it when it exceeds a threshold of 0.5 [54].

B. Evaluation

In order to demonstrate the benefit of the fusion of the audio and video signals for voice activity detection, we evaluated both the multimodal and the unimodal versions of the proposed architecture. The unimodal versions are denoted in the plots by "End-to-End Audio" and "End-to-End Video", respectively, and the multimodal version is denoted in the plots by "End-to-End AV". The unimodal and multimodal versions are constructed and trained as described in Section IV-A. The benefit of fusing the audio and the video modalities is clearly shown in Fig. 3, where the proposed audio-visual architecture significantly outperforms the unimodal versions. We compare the different networks in the form of receiver operating characteristic (ROC) curves, which present the probability of detection versus the probability of false alarm. We also give the area under the curve (AUC) measure for each of the architectures.

To evaluate the performance of the proposed deep multimodal end-to-end architecture, we compare it with the competing audio-visual voice activity detectors presented in [34] and [36]. We evaluated all voice activity detectors on the challenging evaluation set described in Section II-B. In [34], the authors used only the four largest components of their diffusion mapping to describe the audio/video signals. In our experiments, we found that using a larger number of components to describe the signals, can be beneficial. We experimented with a different number of diffusion mapping components between 4 and 20 and found that the performance improved up to the level of using ten components, and increasing the number of components even further did not provide any additional noticeable improvement in performance. This can be explained by the more complex

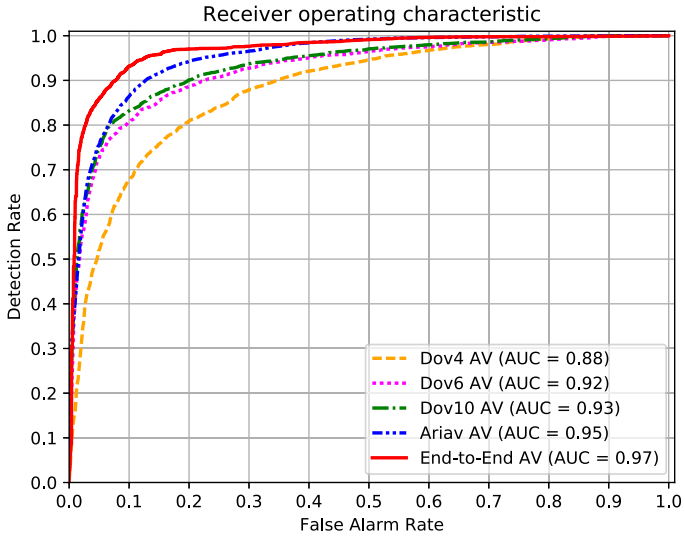

Fig. 4. Probability of detection versus probability of false alarm of our multimodal end-to-end architecture and the VADs presented in [34] and [36] (best viewed in color).

TABLE II
ACCURACY, PRECISION, RECALL, AND F1-SCORE ON EVALUATION SET FOR DIFFERENT END-TO-END MULTIMODAL ARCHITECTURES

Architecture	Acc.	Precision	Recall	F1 Score
separate-LSTM+concat	0.8962	0.8759	0.9205	0.8977
separate-LSTM+MCB	0.9084	**0.9125**	0.8836	0.8978
shared-LSTM+concat	0.8982	0.8705	**0.9356**	0.9018
shared-LSTM+MCB	**0.9152**	0.9033	0.9254	**0.9142**

We denote the proposed architecture as "Shared LSTM" and the architecture in which we use unimodal LSTMs as "Separate LSTM."

TABLE III
ACCURACY ON EVALUATION SET FOR DIFFERENT SEQUENCE LENGTHS T FED INTO OUR END-TO-END MULTIMODAL ARCHITECTURE

Sequence length	Acc.
15	0.9152
30	0.9152
45	0.9148

nature of our evaluation set, in which each frame is contaminated with a different combination of background noise, transient interference, and SNR level. We denote the VAD proposed in [34] using 4,6 and 10 components as "Dov4 AV", "Dov6 AV" and "Dov10 AV" respectively. The VAD proposed in our earlier work [36] was found to be superior to all versions of the VADs presented in [34] and is denoted in the plots by "Ariav AV". In Fig. 4 it is clearly shown that the performance of our proposed deep multimodal end-to-end architecture is superior to those presented in [34] and [36].

We perform several ablation experiments to demonstrate the effectiveness of our proposed end-to-end architecture. To demonstrate the advantages of fusing the embeddings $\mathbf{Z}_a$ and $\mathbf{Z}_v$ using the MCB module, we conducted an experiment in which we replaced the MCB module with a simple vector concatenation, which is standard practice today for multimodal problems. In another set of experiments, we construct a multimodal network from the two unimodal networks in which we do not remove the unimodal LSTM layers. Instead, we feed the MCB module with the outputs of the two unimodal LSTMs and the joint representation produced by the MCB is fed directly to a fully connected layer for classification. We denote this architecture as "separate-LSTM", as opposed to the originally proposed architecture which we denote in the table as "shared-LSTM". Throughout all of the above experiments, the rest of the architecture, including the number of neurons/cells in each layer and the training procedure remains unchanged. Moreover, since the MCB's output size was chosen to be 1024, it matches the size of the concatenated embedding, so it is a fair comparison between the two variants. Table II shows the results in terms of classification accuracy, precision, recall and f1-score on the evaluation set. It can be seen that fusing the two modalities using MCB gives better results than a simple concatenation of the feature vectors. Moreover, the architecture with the shared-LSTM shows better performance, and a possible explanation is that this way the LSTM can capture the dynamics of the speech signal which is shared across the modalities.

Furthermore, we experimented with different sequence lengths T to feed our end-to-end network. We conducted experiments using sequence lengths of $T = \{15, 30, 45\}$, which correspond to 0.6, 1.2 and 1.8 seconds of audio/video. In these experiments, we used our multimodal networks with MCB architecture. Note that the change of T merely affects the number of sequences in the training and evaluation sets in a negligible way, so it is a fair comparison between the different cases. Table III shows the results in terms of classification accuracy on the evaluation set. As seen in Table III, there is no real advantage to using longer sequences as input, but the computational cost and memory consumption are greater, mainly due to the vision network, and therefore we opted to use a sequence length of $T = 15$.

In contrast to our previous work [36], where the network operates on hand-crafted features extracted from the two modalities, the proposed architecture operates on raw signals to extract the most suitable features from each modality to the task of voice activity detection. Moreover, in [36] the features from the two modalities are merely concatenated before being fed to an autoencoder to exploit the relations between the modalities. However, in our proposed architecture the fusion is performed via an MCB module which allows for higher order relations between the two modalities to be exploited.

V. CONCLUSIONS AND FUTURE WORK

We have proposed a deep multimodal end-to-end architecture for speech detection, which utilizes residual connections and dilated convolutions and operates on raw audio and video signals to extract meaningful features in which the effect of transients is reduced. The features from each modality are fused into a joint representation using MCB which allows higher order relations between the two modalities to be explored. In order to further exploit the differences in the dynamics between speech

and the transients, the joint representation is fed into a deep LSTM. A fully connected layer is added, and the entire network is trained in a supervised manner to perform voice activity detection. Experimental results have demonstrated that our multimodal end-to-end architecture outperforms unimodal variants while providing accurate detections even under low SNR conditions and in the presence of challenging types of transients. Furthermore, the use of MCB for modality fusion has also been shown to outperform other methods for modality fusion.

Future research directions include applying the proposed end-to-end multimodal architecture to different voice-related tasks such as speech recognition or speech enhancement. Moreover, we will explore additional methods for noise injection, in which the video signal is augmented and the speakers' voices are modified according to the injected noise levels (Lombard effect).

Another direction worth exploring would be to use the proposed architecture on altogether different modalities, e.g., replacing the audio signal with an electrocardiogram (ECG) signal and training the network to perform ECG related tasks. This is possible since the architecture operates on raw signals and therefore does not depend on audio, or image, specific features.

REFERENCES

[1] D. A. Krubsack and R. J. Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech," *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 319–329, Feb. 1991.

[2] J.-C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 3, pp. 406–412, Jul. 1994.

[3] S. Van Gerven and F. Xie, "A comparative study of speech detection methods," in *Proc. 5th Eur. Conf. Speech Commun. Technol.*, 1997, pp. 1095–1098.

[4] N. Cho and E.-K. Kim, "Enhanced voice activity detection using acoustic event detection and classification," *IEEE Trans. Consum. Electron.*, vol. 57, no. 1, pp. 196–202, Feb. 2011.

[5] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

[6] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[7] J. Ramírez, J. C. Segura, C. Benítez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, no. 3, pp. 271–287, 2004.

[8] D. Dov, R. Talmon, and I. Cohen, "Kernel method for voice activity detection in the presence of transients," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2313–2326, Dec. 2016.

[9] J. W. Shin, J.-H. Chang, and N. S. Kim, "Voice activity detection based on statistical models and machine learning approaches," *Comput. Speech Lang.*, vol. 24, no. 3, pp. 515–530, 2010.

[10] J. Wu and X.-L. Zhang, "Maximum margin clustering based statistical VAD with multiple observation compound feature," *IEEE Signal Process. Lett.*, vol. 18, no. 5, pp. 283–286, May 2011.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[12] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," Linguistic Data Consortium, Philadelphia, PA, USA, 1993.

[13] S. Antol *et al.*, "VQA: Visual Question Answering," in *Proc. Int. Conf. Comput. Vision*, 2015, pp. 2425–2433.

[14] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[15] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 697–710, Apr. 2013.

[16] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 2519–2523.

[17] R. Tahmasbi and S. Rezaei, "A soft voice activity detection using GARCH filter and variance gamma distribution," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1129–1134, May 2007.

[18] S. Leglaive, R. Hennequin, and R. Badeau, "Singing voice detection with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 121–125.

[19] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 6645–6649.

[20] W.-T. Hong and C.-C. Lee, "Voice activity detection based on noise-immunity recurrent neural networks," *Int. J. Adv. Comput. Technol.*, vol. 5, no. 5, pp. 338–345, 2013.

[21] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7378–7382.

[22] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Proc. IEEE Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–4.

[23] D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 1, 2006, pp. 601–604.

[24] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, and C. Jutten, "A study of lip movements during spontaneous dialog and its application to voice activity detection," *J. Acoust. Soc. Amer.*, vol. 125, no. 2, pp. 1184–1196, 2009.

[25] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, and C. Jutten, "Two novel visual voice activity detectors based on appearance models and retinal filtering," in *Proc. 15th Eur. Signal Process. Conf.*, 2007, pp. 2409–2413.

[26] Q. Liu, W. Wang, and P. Jackson, "A visual voice activity detection method with adaboosting," in *Proc. Sens. Signal Process. Defence*, 2011, pp. 1–5.

[27] A. J. Aubrey, Y. A. Hicks, and J. A. Chambers, "Visual voice activity detection with optical flow," *IET Image Process.*, vol. 4, no. 6, pp. 463–472, 2010.

[28] P. Tiawongsombat, M.-H. Jeong, J.-S. Yun, B.-J. You, and S.-R. Oh, "Robust visual speakingness detection using bi-level HMM," *Pattern Recognit.*, vol. 45, no. 2, pp. 783–793, 2012.

[29] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," in *Proc. 16th Eur. Signal Process. Conf.*, 2008, pp. 1–5.

[30] T. Yoshida, K. Nakadai, and H. G. Okuno, "An improvement in audio-visual voice activity detection for automatic speech recognition," in *Proc. 23rd Int. Conf. Ind., Eng. Appl. Appl. Intell. Syst.*, 2010, pp. 51–61.

[31] V. P. Minotto, C. B. O. Lopes, J. Scharcanski, C. R. Jung, and B. Lee, "Audiovisual voice activity detection based on microphone arrays and color information," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 1, pp. 147–156, Feb. 2013.

[32] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[33] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Multimodal deep convolutional neural network for audio-visual emotion recognition," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 281–284.

[34] D. Dov, R. Talmon, and I. Cohen, "Audio-visual voice activity detection using diffusion maps," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 732–745, Apr. 2015.

[35] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, 2006.

[36] I. Ariav, D. Dov, and I. Cohen, "A deep architecture for audio-visual voice activity detection in the presence of transients," *Signal Process.*, vol. 142, pp. 69–74, 2018.

[37] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1764–1772.

[38] G. Trigeorgis *et al.*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5200–5204.

[39] P. Tzirakis, G. Trigeorgis, M. A Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[40] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, Apr. 2018.
[41] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2018, pp. 6548–6552.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[43] A. Van Den Oord *et al.*, "Wavenet: A generative model for raw audio," to be published.
[44] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 317–326.
[45] [Online]. Available: http://www.freesound.org
[46] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," to be published.
[47] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.
[48] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Comput.*, vol. 12, no. 6, pp. 1247–1283, 2000.
[49] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.
[50] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," 2016, arXiv:1606.01847.
[51] M. Charikar, K. Chen, and M. Farach-Colton, "Finding frequent items in data streams," in *Proc. Int. Colloq. Autom., Lang. Program.*, 2002, pp. 693–703.
[52] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proc. 19th ACM Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 239–247.
[53] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[54] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

Ido Ariav received the B.Sc. (*Cum Laude*) degree in electrical engineering and the B.A. degree in economics (*Cum Laude*) from the Technion–Israel Institute of Technology, Haifa, Israel, in 2012. He is currently working toward the Ph.D. degree in electrical engineering with the Technion–Israel Institute of Technology.

From 2013 to 2019, he worked in the field of computer vision as a Researcher with Elbit Systems, Ltd., Haifa, Israel. His research interests include theory and applications of deep learning methods for multimodal signal processing, speech enhancement, and super resolution.

Mr. Ariav is the recipient of the 2012 Wilk Award for excellent undergraduate project at the Signal and Image Processing Lab, Electrical Engineering Department, Technion.

Israel Cohen (M'01–SM'03–F'15) received the B.Sc. (*Summa Cum Laude*), M.Sc., and Ph.D. degrees in electrical engineering from the Technion – Israel Institute of Technology, Haifa, Israel, in 1990, 1993, and 1998, respectively.

He is a Professor of electrical engineering with the Technion – Israel Institute of Technology, Haifa, Israel. He is also a Visiting Professor with the Northwestern Polytechnical University, Xian, China. From 1990 to 1998, he was a Research Scientist with the RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate with the Computer Science Department, Yale University, New Haven, CT, USA. In 2001, he joined the Electrical Engineering Department of the Technion. He is a Co-editor of the Multichannel Speech Processing Section of the *Springer Handbook of Speech Processing* (Springer, 2008), and a coauthor of *Fundamentals of Signal Enhancement and Array Signal Processing* (Wiley-IEEE Press, 2018). His research interests include array processing, statistical signal processing, analysis and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, system identification, and adaptive filtering.

Dr. Cohen was the recipient of the Norman Seiden Prize for Academic Excellence (2017), the SPS Signal Processing Letters Best Paper Award (2014), the Alexander Goldberg Prize for Excellence in Research (2010), and the Muriel and David Jacknow Award for Excellence in Teaching (2009). He serves as Associate Member for the IEEE Audio and Acoustic Signal Processing Technical Committee, and as Distinguished Lecturer for the IEEE Signal Processing Society. He served as Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and the IEEE SIGNAL PROCESSING LETTERS, and as member for the IEEE Audio and Acoustic Signal Processing Technical Committee and the IEEE Speech and Language Processing Technical Committee.

Chapter 5

Depth Map Super-Resolution via Cascaded Transformers Guidance

frontiers in Signal Processing

ORIGINAL RESEARCH
published: 24 March 2022
doi: 10.3389/frsip.2022.847890

Depth Map Super-Resolution *via* Cascaded Transformers Guidance

Ido Ariav and Israel Cohen**

Andrew and Erna Viterby Faculty of Electrical and Computer Engineering, Technion–Israel Institute of Technology, Haifa, Israel

Depth information captured by affordable depth sensors is characterized by low spatial resolution, which limits potential applications. Several methods have recently been proposed for guided super-resolution of depth maps using convolutional neural networks to overcome this limitation. In a guided super-resolution scheme, high-resolution depth maps are inferred from low-resolution ones with the additional guidance of a corresponding high-resolution intensity image. However, these methods are still prone to texture copying issues due to improper guidance by the intensity image. We propose a multi-scale residual deep network for depth map super-resolution. A cascaded transformer module incorporates high-resolution structural information from the intensity image into the depth upsampling process. The proposed cascaded transformer module achieves linear complexity in image resolution, making it applicable to high-resolution images. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art techniques for guided depth super-resolution.

Keywords: depth maps, super-resolution, deep learning, attention, transformers

OPEN ACCESS

Edited by:
Dong Liu,
University of Science and Technology of China, China

Reviewed by:
Wenhan Yang,
Nanyang Technological University, Singapore
Jie Shao,
University of Electronic Science and Technology of China, China

****Correspondence:***
Ido Ariav
idoariav@campus.technion.ac.il
Israel Cohen
icohen@ee.technion.ac.il

Specialty section:
This article was submitted to Image Processing, a section of the journal Frontiers in Signal Processing

Received: *03 January 2022*
Accepted: *25 February 2022*
Published: *24 March 2022*

Citation:
Ariav I and Cohen I (2022) Depth Map Super-Resolution via Cascaded Transformers Guidance. Front. Sig. Proc. 2:847890. doi: 10.3389/frsip.2022.847890

1 INTRODUCTION

Depth information of a scene is crucial in many computer vision applications such as 3D reconstruction (Izadi et al., 2011), driving assistance (Schamm et al., 2009), and augmented reality. Recently, many low-cost consumer depth cameras were introduced, facilitating the use of depth information in various day-to-day scenarios. However, these low-cost sensors often suffer from low spatial resolution, limiting their potential applications. To facilitate such sensors, the depth information usually needs to undergo an upsampling process in which the corresponding high-resolution (HR) depth map is recovered from its low-resolution (LR) counterpart.

The upsampling of depth information is not a trivial task since the fine details in the HR depth map are often missing or severely distorted in the LR depth map, because of the sensor's limited spatial resolution. Moreover, there often exists an inherent ambiguity in super-resolving the distorted fine details. A naive upsampling of the LR image, e.g., bicubic interpolation, usually produces unsatisfactory results with blurred and unsharp edges. Therefore, numerous advanced methods have been proposed recently for the upsampling, commonly termed super-resolution (SR), of depth information.

Current methods for SR of depth maps can be generally categorized as filter-based methods (Yang et al., 2007; He et al., 2012), energy minimization based methods (Ferstl et al., 2013; Yang et al., 2014; Jiang et al., 2018) and learning-based methods, which rely heavily on the use of convolutional neural networks (CNN) (Riegler et al., 2016b; Hui et al., 2016; Guo et al., 2018; Song et al., 2018; Zuo et al., 2019a). Filter-based methods have a relatively low computational complexity but tend to introduce apparent artifacts in the resulting HR depth map. On the other hand, energy minimization methods often require cumbersome and time-consuming computations. They are heavily reliant on

regularization from a statistic or prior that is unavailable for some scenarios. Finally, learning-based methods have blossomed in recent years, and they now provide the best performances in terms of the quality of the upsampled depth map.

In many cases, an LR depth map is accompanied by a corresponding HR intensity image. Many of the more recent methods propose to use this additional image to guide or enhance the SR of the depth map (Park et al., 2011; Kiechle et al., 2013; Kwon et al., 2015; Hui et al., 2016; Guo et al., 2018; Zuo et al., 2019a; Lutio et al., 2019; Li et al., 2020; Ye et al., 2020; Cui et al., 2021; Kim et al., 2021; Zhao et al., 2021). These methods assume that correspondence can be established between an edge in the intensity image and the matching edge in the depth map. Then, given that the intensity image has a higher resolution, its edges can determine depth discontinuities in the super-resolved HR depth map. However, there could be cases in which an edge in the intensity image does not correspond to a depth discontinuity in the depth map or vice versa, e.g., in the case of smooth, highly textured surfaces in the intensity image. These cases lead to texture copying, where color textures are over-transferred to the super-resolved depth map at the boundaries between textured and homogeneous regions. Hence, a more sophisticated guidance scheme needs to be considered.

In this paper, to alleviate the texture copying problem, we propose a Cascaded Transformer Guidance Module (CTGM) for guided depth map SR. Transformers, designed initially for sequence modeling tasks (Vaswani et al., 2017), are notable for their use of attention to model long-range dependencies in the data. Recently, transformers were successfully adapted to computer vision tasks with promising results. Our proposed CTGM is constructed by stacking several transformer blocks, each operating locally within non-overlapping windows that partition the entire input. Window shift is introduced between consecutive transformer blocks to enable inter-window connections to be learned. The CTGM is fed with HR features extracted from the intensity image and is trained to pass only salient and consistent features that are then incorporated into the depth upsampling process. Our proposed CTGM is capable of learning structural and content information from a large receptive field, which was shown to be beneficial for SR tasks (Zhang et al., 2017).

Our overall architecture can be divided into three main parts: a depth branch, an intensity branch, and the CTGM. The proposed depth branch comprises several Residual Dilated Groups (RDG) (Zhang et al., 2018a), and performs the upsampling of the given LR depth map in a multi-scale manner, as in, e.g., Hui et al. (2016). Meanwhile, the intensity image is fed into the intensity branch, which extracts HR features and complements the LR depth structures in the depth branch via the CTGM. This process is repeated according to the desired upsampling factor. This closely guided multi-scale scheme allows the network to learn rich hierarchical features at different levels, and better adapt to the upsampling of both fine-grained and coarse patterns. Moreover, this enables the network to seamlessly utilize the guidance from HR intensity features in multiple scales. In **Section 4** we show that our proposed method achieves results with sharper boundaries, more complete details, and better quantitative values in terms of Root Mean Square Error (RMSE) compared to competing guided SR approaches. Our proposed architecture is shown in **Figure 1** for the case of an upsampling factor of 2.

Our contributions are as follows: (1) We introduce a novel cascaded transformer-based mechanism to produce salient guidance features from the intensity branch. (2) Our proposed CTGM exhibits linear memory constraints, making it applicable even for very large images. (3) Unlike other transformer architectures, our architecture can handle different input resolutions, both during training and inference, making it highly applicable to real-world tasks. (4) We achieve the state of the art performance on several depth upsampling benchmarks.

The remainder of this paper is organized as follows. In **Section 2**, we present a brief overview of the related work. In **Section 3**, we present our proposed architecture for depth map SR in detail. In **Section 4**, we conduct extensive experiments and report our results. Also, an ablation study is performed. Finally, in **Section 5**, we conclude and discuss future research directions.

2 RELATED WORK

Classic methods for depth map SR were inspired mainly by works from the related field of single color image SR. However, due to the limitation of single depth map SR, such methods usually work well only for small upsampling factors, e.g., 2 or 4. Guided depth map SR, on the other hand, is more robust even for more prominent upsampling factors, e.g., 8 or 16. This improved robustness is achieved by introducing guidance from cross domains, e.g., HR intensity image. A more detailed review of guided depth SR methods is given in the following subsections, emphasizing methods based on deep neural networks.

2.1 Single Depth Map Super Resolution

Earlier works for SR of depth maps, inspired by single image SR methods, mainly focused on filtering-based strategies. Mac Aodha et al. (2012) proposed that the matching HR depth candidate will be searched from a collected database for a given LR depth patch. Selecting the most probable candidate was then formulated as a Markov random field labeling problem. Hornacek et al. (2013) proposed to perform single depth map SR by exploring patch-wise scene self-similarity. Lei et al. (2017) proposed a view synthesis quality-based filtering, which jointly considers depth smoothness, texture similarity, and view synthesis quality.

Other works formulated depth map SR as a global optimization problem. Xie et al. (2015) offered an edge-guided depth map SR method, which applies Markov random field optimization to construct the HR edge map from the LR depth map. Later works considered dictionary learning strategies. Ferstl et al. (2015) used an external database to learn a dictionary of edge priors and then used the learned edge priors to guide the upsampling of an LR depth map in a variational sparse coding framework. Mandal et al. (2016) proposed an edge-preserving constraint to preserve the discontinuity in the depth map and a pyramidal

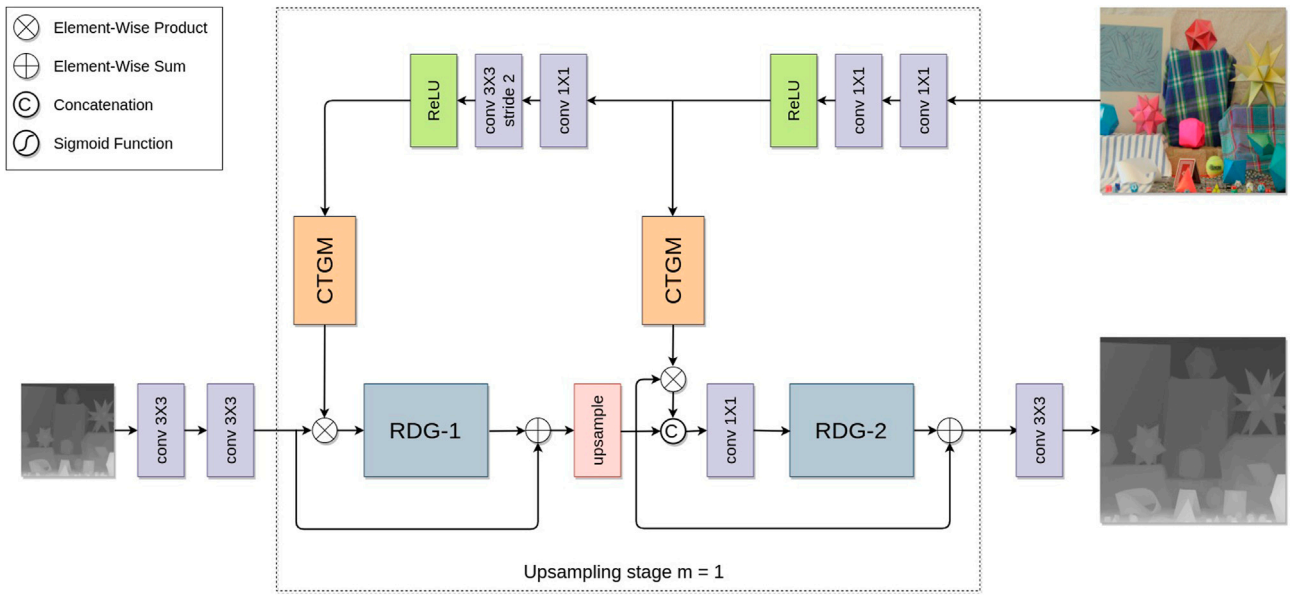

FIGURE 1 | Our proposed architecture for guided depth SR for an upsampling factor of 2. The dashed square marks a single upsampling stage *i* that can be duplicated according to the desired upsampling factor $s = 2^m$.

reconstruction framework to better deal with higher scaling factors.

Later, Riegler et al. (2016b) proposed ATGV-Net, which combined a deep CNN with a variational method to recover an accurate HR depth map. Recently, Huang et al. (2019) proposed a pyramid-structured network composed of dense residual blocks that use densely connected layers and residual learning to model the mapping between high-frequency residuals and LR depth maps. A deep supervision scheme in which auxiliary losses were added at various scales within the network was utilized to reduce the difficulty of model training.

2.2 Intensity Guided Depth Map Super Resolution

Unlike an HR depth map, an HR intensity image can usually be easily acquired by color cameras. Thus, in many real-life scenarios, a corresponding intensity image can be used to guide the upsampling process or enhance the low-quality depth maps. Various methods have been proposed to improve the quality of depth maps by the guidance of the HR intensity image. These methods can be categorized as filtering-based methods (He et al., 2010; Liu et al., 2013; Lu and Forsyth, 2015), global optimization-based methods Dong et al. (2016); Ferstl et al. (2013); Ham et al. (2015a,b); Jiang et al. (2018); Liu et al. (2016); Park et al. (2014, 2011); Yang et al. (2012, 2014), sparse representation-based methods (Kiechle et al., 2013; Kwon et al., 2015) and deep learning-based methods (Riegler et al., 2016a; Hui et al., 2016; Zhou et al., 2017; Guo et al., 2018; Zuo et al., 2019a,b; Zhao et al., 2021; Lutio et al., 2019; Kim et al., 2021; Li et al., 2020; Ye et al., 2020; Cui et al., 2021), which are in the focus of this paper.

Liu et al. (2013) utilized geodesic distances to upsample an LR depth map with the guidance of a corresponding HR color image. Lu and Forsyth (2015) used the correlation between edges in a segmentation image and boundaries in depth images to produce detailed HR depth structures. Yang et al. (2012) formulated the depth recovery problem to minimize auto-regressive prediction errors. Ferstl et al. (2013) developed a convex optimization problem for depth image upsampling, which guides the depth upsampling by an anisotropic diffusion tensor calculated from an HR intensity image. Park et al. (2014) extended the non-local structure regularization by combining the additional HR color input when upsampling an LR depth map. Kiechle et al. (2013) introduced a bimodal co-sparse analysis to capture the inter-dependency of registered intensity and depth information. Kwon et al. (2015) proposed a data-driven method for depth map refinement through multi-scale dictionary learning, based on the assumption that a linear combination of basis functions can efficiently represent local patches in both depth and RGB images. Jiang et al. (2018) proposed a depth map SR method in which non-local correlations are exploited by an auto-regressive model in the frequency domain. A multi-directional total variation prior is used in the spatial domain to characterize the geometrical structures.

Inspired by the great success in the SR of color images, deep learning methods for depth map SR have recently attracted more and more attention. Riegler et al. (2016a) used a fully convolutional network to produce an initial estimation for the HR depth. This estimation, in turn, was fed into a non-local variational model to optimize the final result. Hui et al. (2016) proposed an "MSG-Net," which infers the HR depth map from its LR counterpart in multiple stages, and uses a multi-scale fusion strategy to complement LR depth features with HR intensity

features in the high-frequency domain. Zhou et al. (2017) concluded that color images are more helpful for the depth map SR problem when noise is present, and the scaling factor is large. Guo et al. (2018) proposed exploiting multiple level receptive fields by constructing an input pyramid and extracting hierarchical features from the depth map and intensity image. These features are then concatenated, and the residual map between the interpolated depth map and the corresponding HR one is learned via a residual U-Net architecture. Zuo et al. (2019a) proposed a multi-scale upsampling network in which intensity features guide the upsampling process in multiple stages, and both low and high-level features are taken into account in the reconstruction of HR depth maps thanks to local and global connections. Zuo et al. (2019b) proposed another multi-scale network with global and local residual learning. The LR depth map is progressively upsampled, guided by the HR intensity image in multiple scales. Moreover, batch normalization layers were used to improve the performance of depth map denoising. Zhao et al. (2021) proposed to use a discrete cosine transform network in which pairs of color/depth images are fed into the semi-coupled feature extraction module to extract common and unique features from both modalities. The color features are then passed through an edge attention mechanism to highlight the edges useful for upsampling. Finally, a discrete cosine transform was employed to solve the SR optimization problem. Lutio et al. (2019) proposed to find a transformation from the guide image to the target HR depth map, which can be regarded as a pixel-wise translation. Kim et al. (2021) proposed to use deformable convolutions (Dai et al., 2017) for the upsampling of depth maps, where the spatial offsets for convolution are learned from the features of the given HR guidance image. Li et al. (2020) proposed a recumbent Y network for depth map SR. They built the network based on residual channel attention blocks and utilized spatial attention-based feature fusion blocks to suppress the artifacts of texture copying and depth bleeding. Ye et al. (2020) proposed a progressive multi-branch aggregation network by using the multi-branch fusion method to gradually restore the degraded depth map. Cui et al. (2021) proposed an architecture built on two U-Net branches for HR color images and LR depth maps. This architecture uses a dual skip connection structure to leverage the feature interaction of the two branches and a multi-scale fusion to fuse the deeper and multi-scale features of two branch decoders for more effective feature reconstruction.

However, the methods above still use simple schemes such as concatenation to fuse the guidance features extracted from the intensity image with the depth features. At the same time, we propose to use a CTGM, which directs the allocation of available processing resources towards the most informative components of the input, thus achieving superior results, as demonstrated in **Section 4**.

2.3 Vision Transformers

In recent years, transformer-based architectures (Vaswani et al., 2017) achieved great success in natural language processing tasks, enabling long-range dependencies in the data to be learned via their sophisticated attention mechanism. Their tremendous success in the language domain has led researchers to investigate their adaptation to computer vision, where it has recently demonstrated promising results on certain tasks, specifically image classification (Dosovitskiy et al., 2020; Liu et al., 2021; Wang et al., 2021) and object detection (Carion et al., 2020; Zhu et al., 2020).

A primary transformer encoder, as proposed in Vaswani et al. (2017), usually consists of alternating layers of multiheaded self-attention (MSA) and MLP blocks, with Layer Normalization (LN) before every block and residual connections after every block.

An MSA block takes as input a sequence of length N of d-dimensional embeddings $\mathbf{X} \in \mathbb{R}^{N\times d}$ and produces an output sequence $\mathbf{Y} \in \mathbb{R}^{N\times d}$ *via*:

$$\begin{aligned} &\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \mathbf{K} = \mathbf{X}\mathbf{W}_K, \mathbf{V} = \mathbf{X}\mathbf{W}_V \\ &\mathbf{A} = \mathrm{Softmax}\left(\mathbf{Q}\mathbf{K}^T \Big/ \sqrt{\mathbf{d}}\right) \\ &\mathbf{Y} = \mathbf{A}\mathbf{V} \end{aligned} \tag{1}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are $D \times D$ parameter matrices of 1×1 convolutions responsible for projecting the entries of the sequence $\mathbf{X}$ into the three standard transformer paradigms; keys, queries, and values, respectively. Each entry of the output sequence $\mathbf{Y}$ is a linear combination of values in $\mathbf{V}$ weighted by the attention matrix $\mathbf{A}$, which itself is computed from similarities between all pairs of query and key vectors.

The transformer's expressive power comes from computing the self-attention $\mathbf{A}$ and $\mathbf{Y}$. This computation, however, comes with a quadratic cost; it takes $O(N^2)$ time and space to compute the pairwise similarities between $\mathbf{Q}$ and $\mathbf{K}$ and to compute the linear combination of $\mathbf{V}$ vectors. This quadratic complexity makes it impractical to apply self-attention to images directly, as even for small images $\mathbf{X}$ quickly becomes too long a sequence for self-attention.

In light of this inherent limitation, efforts have been made to restrict these sequence lengths in a modality-aware manner while preserving modeling performance. The pioneering work of Dosovitskiy et al. (2020) proposed to directly apply a transformer architecture on non-overlapping medium-sized image patches for image classification. This local self-attention helps mitigate these memory constraints, as opposed to global self-attention.

3 PROPOSED METHOD

3.1 Formulation

A method for guided depth SR aims to find the nonlinear mapping between an LR depth map and the corresponding HR depth map. An HR intensity image guides the process of finding this nonlinear relation. For a given scaling factor $s = 2^m$ we denote the LR depth map as $\mathbf{D}_{\mathrm{LR}} \in \mathbb{R}^{H/s\times W/s}$ and the respective HR guidance intensity image as $\mathbf{I}_{\mathrm{HR}} \in \mathbb{R}^{H\times W}$. Then, the corresponding HR depth map $\mathbf{D}_{\mathrm{HR}} \in \mathbb{R}^{H\times W}$ can be found from:

$$\mathbf{D}_{\mathrm{HR}} = \mathbf{F}(\mathbf{D}_{\mathrm{LR}}, \mathbf{I}_{\mathrm{HR}}; \theta) \tag{2}$$

where $\mathbf{F}$ denotes the nonlinear mapping learned by our proposed architecture, and θ represents the learned network's parameters. We note that in our proposed architecture, contrary to some

other works, $\mathbf{D}_{\mathrm{LR}}$ is upsampled in a multi-stage scheme of m stages. In each stage, $\mathbf{D}_{\mathrm{LR}}$ is upsampled by a factor of two until it reaches the desired scaling factor s. We note that any upsampling stage m can also perform an upsampling by a factor of three. Thus the overall architecture is flexible enough for real applications and can be configured to achieve upsampling with scaling factors that are not the exponent of 2.

Throughout the section, we use $\mathbf{Conv}_3(\cdot)$ to denote a convolution layer with a kernel size of 3×3 and $\mathbf{Conv}_1(\cdot)$ to denote a convolution layer with a kernel size of 1×1.

3.2 Overall Network Architecture

As shown in **Figure 1**, our architecture mainly consists of three parts: intensity branch, depth branch, and CTGM, which provides guidance from the intensity branch to the depth branch. We now review the general structure of our depth and intensity branches, and more details of the proposed CTGM modules will be given in **Section 3.3**.

3.2.1 Intensity Branch

Our intensity branch consists of two primary modules; (1) a feature extraction module and (2) a downsampling module. These two basic modules are repeated in each upsampling stage $i \in \{1 \dots m\}$. The feature extraction module consists of two consecutive convolution layers with a kernel size of 1×1, followed by an element-wise rectified linear unit (ReLU) activation function. This module extracts essential features from the intensity image as guidance for the depth branch. The downsampling module performs a similar operation while also downsampling the feature maps by a factor of two. It consists of a convolution layer with a kernel size of 1×1 followed by another convolution layer with a kernel size of 3×3 and stride 2, which performs the downsampling. A ReLU activation then follows these two convolution layers. In this manner, the intensity frequency components are progressively downsampled to provide multiple-scale guidance for the depth branch via the CTGM, as elaborated in **Section 3.3**.

Specifically, a single upsampling stage $i \in \{1 \dots m\}$ of the intensity branch can be formulated as:

$$\mathbf{Y}^{i}_{\mathrm{FE}} = \sigma\left(\mathbf{Conv}_1\left(\mathbf{Conv}_1\left(\mathbf{Y}^{i-1}_{\mathrm{DS}}\right)\right)\right) \tag{3}$$

$$\mathbf{Y}^{i}_{\mathrm{DS}} = \sigma\left(\mathbf{Conv}_{3,2}\left(\mathbf{Conv}_1\left(\mathbf{Y}^{i}_{\mathrm{FE}}\right)\right)\right) \tag{4}$$

where $i \in \{1 \dots m\}$ denotes the current upsampling stage, σ is a ReLU activation function, and $\mathbf{Conv}_{3,2}(\cdot)$ is a convolution layer with a kernel size of 3×3 and stride 2. The input $\mathbf{Y}^{i-1}_{\mathrm{DS}}$ for upsampling stage i is either the output of upsampling stage i–1 or the input HR intensity image $\mathbf{I}_{\mathrm{HR}}$ in the case of $i = 1$, meaning $\mathbf{Y}^{0}_{\mathrm{DS}} = \mathbf{I}_{\mathrm{HR}}$.

3.2.2 Depth Branch

The depth branch plays the primary role of finding the nonlinear mapping between the LR and the super-resolved HR depth maps. Motivated by Zhang et al. (2018a), we use global and local residual learning, allowing for high-frequency details needed for super-resolving $\mathbf{D}_{HR}$ to be learned in the network and its sub-networks.

In our depth branch, we first extract shallow features from the LR input $\mathbf{D}_{LR}$ by feeding it through two consecutive convolution layers as suggested by, among others, Haris et al. (2018); Zhang et al. (2018b):

$$\mathbf{D}^{0} = \mathbf{Conv}_3\left(\mathbf{Conv}_3\left(\mathbf{D}_{\mathrm{LR}}\right)\right) \tag{5}$$

$\mathbf{D}_0$ is then progressively upsampled in m stages by a factor of two in each stage. Each such upsampling stage is composed of an RDG as proposed by Zhang et al. (2018a), followed by an upsampling module, and finally, a second RDG. This upsampling stage is duplicated according to the desired upscaling factor s.

An RDG is composed of stacking G Residual Dilated Blocks (Zhang et al., 2018a), where each such block is composed of stacking L layers of dilated convolution. A long skip connection connects the RDG's input to its output, stabilizing the training process Haris et al. (2018) and allowing the network to suppress low-frequency information from earlier layers and recover more useful information.

The first RDG in each upsampling stage performs a deep feature extraction. For the RDG to successfully recover high-frequency details from its LR input, we first scale its input via the CTGM, as elaborated in **Section 3.3**. Features extracted by the first RDG are upsampled by a factor of two via a pixel shuffle module (Shi et al., 2016). Thus the upsampling operators are learned in a data-driven way to improve the representation ability of our model. Finally, the upsampled feature maps are scaled once more by the output of another CTGM and fused with the unscaled feature maps via a convolution layer. The fused feature maps are then passed through a second RDG to explore deeper relations between the depth and intensity features.

A single upsampling stage i can be formulated as:

$$\mathbf{F}^{i}_{\mathrm{RDG1}} = \mathbf{H}^{1}_{\mathrm{RDG}}\left(\mathbf{D}^{i-1} \otimes \mathbf{H}^{1}_{\mathrm{CTGM}}\right) \oplus \mathbf{D}^{i-1} \tag{6}$$

$$\mathbf{F}^{i}_{\mathrm{UP}} = \mathbf{H}_{\mathrm{PS}}\left(\mathbf{F}^{i}_{\mathrm{RDG1}}\right) \tag{7}$$

$$\mathbf{D}^{i} = \mathbf{H}^{2}_{\mathrm{RDG}}\left(\mathbf{Conv}_1\left(\mathbf{F}^{i}_{\mathrm{UP}} \copyright \left(\mathbf{F}^{\mathrm{i}}_{\mathrm{UP}} \otimes \mathbf{H}^{2}_{\mathrm{CTGM}}\right)\right)\right) \oplus \mathbf{F}^{i}_{\mathrm{UP}} \tag{8}$$

where $\mathbf{H}_{\mathrm{RDG}}$ denotes the function learned by our RDG, $\mathbf{H}_{\mathrm{PS}}$ is a pixel shuffle upsampling module, $\otimes$ denotes element-wise product, $\oplus$ denotes element-wise sum, $\mathbf{H}_{\mathrm{CTGM}}$ denotes the scaling features produced from our CTGM, © denotes a concatenation operation and $\mathbf{D}^{i-1}$ is the output of the previous upsampling stage. More implementation details are given in **Section 4.1**.

3.3 Cascaded Transformer Guidance Module

Guidance from the intensity image in most previous CNN-based guided SR methods is usually achieved by extracting feature maps from the intensity image and concatenating them to features extracted in the depth branch. This guidance scheme effectively treats all features equally, in both spatial and channel domains, which is not optimal. Moreover, feature maps extracted from the intensity image via CNN usually have a limited receptive field, which affects the guidance quality. In comparison, we propose to exploit long-range dependencies in the guidance image via a

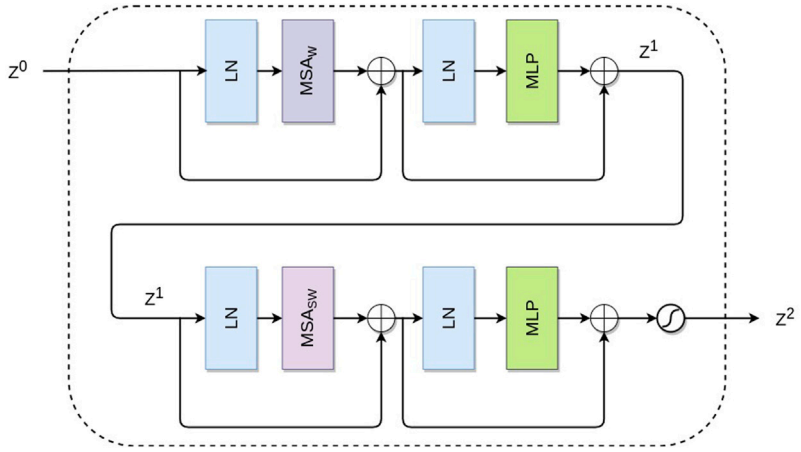

FIGURE 2 | Our proposed cascaded transformer guidance module.

novel cascaded transformer guidance module. The motivation is that in image SR, high-frequency features are more informative for HR reconstruction, and a large receptive field is also beneficial (Zhang et al., 2017). Our proposed CTGM is shown in **Figure 2**.

Before elaborating on the structure of our proposed CTGM, we observe significant challenges in transferring the transformer's high performance in the language domain to the visual domain, and specifically to low-level vision tasks. First, unlike word tokens that serve as the essential processing elements in language transformers, images can vary substantially in scale in real-life scenarios. However, in most existing transformer-based models, tokens must all be of a fixed scale. Another difference is the higher number of pixels in images than words text passages. Specifically, the task of SR requires dense prediction at the pixel level while avoiding down-scaling the input as much as possible to prevent loss of HR information. Working with such HR inputs would be intractable for transformers, as the computational complexity of its self-attention is quadratic to image size.

To overcome these issues, we build upon the work of Liu et al. (2021) and propose a cascaded transformer module, which operates on non-overlapping windows that partition the entire input image. The number of pixels in each such window is fixed, and by computing self-attention locally within each window, the complexity becomes linear to image size. Moreover, our proposed CTGM constructs hierarchical representations by applying several such transformer layers consecutively. We shift the partitioning windows with each layer, gradually merging neighboring patches in deeper transformer layers. The shifted windows overlap with the preceding layer windows, providing connections that significantly enhance modeling power. Our transformer model can conveniently extract meaningful information suitable for dense tasks such as SR and encode distant dependencies or heterogeneous interactions with these hierarchical feature maps. Furthermore, the window-based local self-attention is scalable; the linear complexity makes working with large inputs feasible while also enabling working with variable size inputs (given input size is divisible by window size).

Formally, given an intermediate feature map $F_{\mathrm{I}} \in \mathbb{R}^{C\times H\times W}$ as input, our CTGM first splits F_{I} into non-overlapping patches of size (P, P) to form $F_{I,p} \in \mathbb{R}^{\hat{C}\times\hat{H}\times\hat{W}}$ where $\hat{H} = H/P$ and $\hat{W} = W/P$. A trainable convolution layer with a kernel size of $P \times P$ and stride P is applied to construct an initial patch embedding $F_{I,p,emb} \in \mathbb{R}^{\hat{C}\times\hat{H}\times\hat{W}}$. Next, we apply window partitioning such that the windows partition $F_{I,p,emb}$ in a non-overlapping manner, where each window is of size $M \times M$. Every such window is flattened to form the window embeddings $F_{win} \in \mathbb{R}^{M^2\times\hat{C}}$, which forms the input sequence for the cascaded transformer module. During both the patches and windows partitioning, zero padding of the input is applied if necessary.

Similar to Hu et al. (2019), relative position embeddings are added to the window embeddings F_{win} to retain positional information. We use standard learnable 1D position embeddings since we have not observed significant performance gains from using more advanced 2D-aware or global position embeddings. We refer to this joint embedding $\mathbf{Z}^0$ which is the input to the following transformer module. Note that for computations efficiency, we batch all $\hat{H} \times \hat{W}/(M \times M)$ window embeddings before feeding them to the transformer module.

Our proposed transformer module receives $\mathbf{Z}^0$ as input and computes a hierarchical local self-attention within each window. We construct our transformer module by concatenating two modified transformer blocks, as shown in **Figure 2**. Each such transformer block is modified from the original transformer block by replacing the standard MSA module with a windows-based MSA ($\mathrm{MSA_w}$) while keeping the other layers unchanged. In an $\mathrm{MSA_w}$ module, we apply **Eq. 1** locally within each $M \times M$, thus avoiding computing self-attention on the entire input. Specifically, as illustrated in **Figure 2**, our modified transformer block consists of an $\mathrm{MSA_w}$ module, followed by a 2-layer MLP with GELU non-linearity in between. An LN layer is applied before each MLP and $\mathrm{MSA_w}$ module, and a residual connection is applied after each module.

We propose a shifted window partitioning approach that alternates between two partitioning configurations in the two

consecutive transformer blocks to increase modeling power and introduce cross-window connections. The first block uses a regular window partitioning strategy which starts from the top-left pixel. Then, the next block applies a windowing configuration shifted from the preceding block by displacing the windows by $M/2$, $M/2$ pixels from the regularly partitioned windows. We denote the MSA module that operates with the shifted window partitioning approach as MSA_{sw}. Finally, the two consecutive transformer blocks are computed as -

$$\hat{\mathbf{Z}}^1 = \mathbf{MSA}_{\mathbf{w}}\left(\mathbf{LN}\left(\mathbf{Z}^0\right)\right) + \mathbf{Z}^0 \tag{9}$$

$$\mathbf{Z}^1 = \mathbf{MLP}\left(\mathbf{LN}\left(\hat{\mathbf{Z}}^1\right)\right) + \hat{\mathbf{Z}}^1 \tag{10}$$

$$\hat{\mathbf{Z}}^2 = \mathbf{MSA}_{\mathbf{sw}}\left(\mathbf{LN}\left(\mathbf{Z}^1\right)\right) + \mathbf{Z}^1 \tag{11}$$

$$\mathbf{Z}^2 = \mathbf{MLP}\left(\mathbf{LN}\left(\hat{\mathbf{Z}}^2\right)\right) + \hat{\mathbf{Z}}^2 \tag{12}$$

where $\hat{\mathbf{Z}}^1$ and $\mathbf{Z}^1$ denote the output features of the MSA_{w} and MLP modules of the first block, respectively. Similarly, $\hat{\mathbf{Z}}^2$ and $\mathbf{Z}^2$ denote the output features of the MSA_{sw} and MLP modules of the second block, respectively. This shifted window partitioning approach introduces connections between neighboring non-overlapping windows in the previous layer, which improves modeling performance as shown in **Section 4**.

The overall CTGM module can be summarized as:

$$\mathbf{F}_{\text{CTGM}} = \hat{\sigma}\left(\mathbf{Z}^2\right) \tag{13}$$

where $\hat{\sigma}$ denotes the sigmoid function.

Finally, $\mathbf{F}_{\text{CTGM}}$, the features generated by the CTGM are used to scale the corresponding depth features in the depth branch by element-wise multiplication.

4 EXPERIMENTS

4.1 Training Details

To make a fair comparison to other competing depth map SR methods, we constructed our train and test data similarly to Guo et al. (2018); Huang et al. (2019); Hui et al. (2016), and more. We collected 92 pairs of depth and color images from the MPI Sintel depth dataset (Butler et al., 2012) and from the Middlebury dataset (Scharstein and Szeliski, 2002; Scharstein and Pal, 2007; Scharstein et al., 2014). We followed the same training and validation split as in Hui et al. (2016) and used 82 pairs for training and ten pairs for validation.

Instead of using the full-scale HR depth and intensity images as input in the training process, we randomly extracted patches and used these as input to our network. We opted to use an LR patch size of 96 × 96 pixels, having observed that using a larger patch size did not significantly improve the training accuracy. However, the memory requirements and computation time increase substantially with patch size. Therefore, for upsampling factors of 2 and 4, we extracted HR patches of sizes 192 × 192 and 384 × 384, respectively. For the upsampling factors of 8 and 16, we used smaller LR patch sizes of 48 × 48 and 24 × 24, respectively, due to memory limitations and the fact that some full-scale images had a resolution of <400. The LR patches were generated by downsampling each HR patch with bicubic interpolation according to the desired scaling factor. We augmented the extracted patches by randomly performing either a horizontal flip, a vertical flip, or a 90° rotation during training.

4.2 Implementation Details

In our proposed architecture, we set the number of RDBs in each RDG to $G = 20$ RDBs, and each such RDB has $L = 4$ dilated convolution layers as described in **Section 3.2.2**. These values for G and L provided the best performance to network size trade-off in our experiments. All convolution layers throughout our network have a stride of 1, and $C = 64$ filters, unless otherwise noted. A zero-padding strategy is used to keep size fixed for convolution layers with a kernel size of 3 × 3. The final convolution layer has only one filter, as we output depth values. In our CTGM implementation, we use a patch size of $p = 4$, an embedding dimension of $\hat{C} = 64$, and set the number of patches in each window to be $M = 12$ throughout the CTGM. Each MSA_{w} and MSA_{sw} module has four attention heads.

We trained a specific network for each upsampling factor $s \in 2$, 4, 8, 16, and used the Pytorch framework Paszke et al. (2019) to train our models. We used a batch size of 4 in all our experiments, with random initialization of the filter weights of each layer. We trained each network for 3×10^5 iterations of back-propagation. Our model was optimized using the $\mathcal{L}_1$ loss and the ADAM optimizer Kingma and Ba (2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The initial learning rate was set to 10^{-4} and then divided by 2 every 1×10^5 iterations. All our models were trained on a PC with an i9-9960x CPU, 64GB RAM, and two Titan RTX GPUs.

Our code and trained models will be made public upon publication.

4.3 Results

This section provides both quantitative and qualitative evaluations of our guided depth map SR method. We report the results of bicubic interpolation as a baseline, and compare our results to state-of-the-art global optimization-based methods Ferstl et al. (2013); Liu et al. (2016); Park et al. (2014), a sparse representation-based method (Kiechle et al., 2013) and mainly deep learning-based methods (Guo et al., 2018; Huang et al., 2019; Hui et al., 2016; Zuo et al., 2019a,b; Zhao et al., 2021; Kim et al., 2021; Li et al., 2020; Ye et al., 2020; Cui et al., 2021). We evaluated our proposed method on the noise-free Middlebury dataset and the more challenging noisy Middlebury dataset. Moreover, we demonstrate the generalization capability of our proposed method by evaluating on the NYU Depth v2 dataset.

4.3.1 Results on the Noise-Free Middlebury Dataset

Similar to recent works, we first evaluate the performances of the different methods on the noise-free hole-filled Middlebury RGB-D datasets for various scaling factors $s \in 2$, 4, 8, 16. The Middlebury datasets provide high-quality depth maps and RGB pairs in complex real-world scenes. In **Tables 1**, **2** we report the RMSE values for different scale factors, where the best RMSE for each evaluation is in boldface, whereas the second best one is underlined. All results in **Tables 1**, **2** are

TABLE 1 | Quantitative comparisons on "art." "books" and "laundry" from the noise-free middlebury dataset in terms of RMSE values for different scaling factors.

Method	Art				Books				Laundry			
	x2	x4	x8	x16	x2	x4	x8	x16	x2	x4	x8	x16
Bicubic	2.64	3.88	5.60	8.58	1.02	1.56	2.24	3.36	1.30	2.11	3.10	4.47
TGV Ferstl et al. (2013)	3.19	4.06	5.08	7.61	1.52	2.21	2.47	3.54	1.84	2.20	3.92	6.75
RDGE Liu et al. (2016)	2.31	3.26	4.31	6.78	1.14	1.53	2.18	2.92	1.47	2.06	2.87	4.22
NLMR Park et al. (2014)	3.01	4.24	6.32	10.04	1.25	1.96	2.92	4.34	1.88	2.64	3.78	6.13
JID Kiechle et al. (2013)	1.18	1.92	2.76	5.74	0.45	0.71	1.01	1.93	0.68	1.10	1.83	3.62
PSR Huang et al. (2019)	0.66	1.59	2.57	4.83	0.54	0.83	1.19	1.70	0.52	0.92	1.52	2.97
MSG Hui et al. (2016)	0.67	1.49	2.79	5.95	0.37	0.66	1.09	1.87	0.67	1.02	1.35	2.03
MFR Zuo et al. (2019b)	0.71	1.54	2.71	4.35	0.42	0.63	1.05	1.78	0.61	1.11	1.75	3.01
PMBA Ye et al. (2020)	0.61	1.19	2.47	4.37	0.41	0.53	1.10	1.51	0.38	0.80	1.54	2.72
RDN Zuo et al. (2019a)	0.56	1.47	2.60	4.16	0.36	0.62	1.00	1.68	0.48	0.96	1.63	2.86
DSR Guo et al. (2018)	0.53	1.21	2.23	3.95	0.42	0.60	0.89	1.51	0.44	0.75	1.21	1.89
RYN Li et al. (2020)	**0.26**	0.98	2.04	3.37	0.18	0.36	0.73	1.37	0.22	0.64	1.21	2.01
CUN Cui et al. (2021)	0.27	1.05	2.27	3.67	**0.16**	**0.35**	0.73	1.45	0.19	0.59	1.15	2.25
GDC Kim et al. (2021)	0.33	1.09	2.04	3.58	0.19	0.38	0.68	1.41	0.24	0.64	1.13	2.13
Ours	0.31	**0.73**	**1.89**	**2.76**	0.21	**0.35**	**0.66**	**1.22**	**0.18**	**0.43**	**0.87**	**1.62**

We report the RMSE values for different scale factors, where the best RMSE for each evaluation is in boldface, whereas the second best one is underlined.

TABLE 2 | Quantitative comparisons on "dolls," "moebius" and "reindeer" from the noise-free middlebury dataset in terms of RMSE values for different scaling factors.

Method	Dolls				Moebius				Reindeer			
	x2	x4	x8	x16	x2	x4	x8	x16	x2	x4	x8	x16
Bicubic	0.78	1.21	1.78	2.57	0.93	1.40	2.05	2.95	1.52	2.51	3.92	5.72
TGV Ferstl et al. (2013)	1.17	1.42	2.05	4.44	1.47	2.03	2.58	3.50	2.41	2.67	4.29	8.80
RDGE Liu et al. (2016)	1.14	1.49	1.94	2.45	0.97	1.44	2.21	2.79	1.82	2.58	3.24	4.90
NLMR Park et al. (2014)	1.16	1.64	2.39	3.71	1.12	1.76	2.62	4.07	2.25	3.20	4.63	6.94
JID Kiechle et al. (2013)	0.70	0.92	1.26	1.74	0.64	0.89	1.27	2.13	0.90	1.41	2.12	4.64
PSR Huang et al. (2019)	0.58	0.91	1.31	1.88	0.52	0.86	1.21	1.87	0.59	1.11	1.80	3.11
MSG Hui et al. (2016)	0.46	0.72	0.99	1.59	0.36	0.68	1.14	2.07	0.94	1.33	1.72	2.99
MFR Zuo et al. (2019b)	0.60	0.89	1.22	1.74	0.42	0.72	1.10	1.73	0.65	1.23	2.06	3.74
PMBA Ye et al. (2020)	0.36	0.66	1.08	1.75	0.39	0.55	1.13	1.62	0.40	0.92	1.76	2.86
RDN Zuo et al. (2019a)	0.56	0.88	1.21	1.71	0.38	0.69	1.06	1.65	0.51	1.17	1.60	3.58
DSR Guo et al. (2018)	0.49	0.81	1.10	1.60	0.43	0.67	0.96	1.57	0.51	0.96	1.57	2.54
RYN Li et al. (2020)	0.27	0.59	0.97	1.37	0.24	0.50	0.81	1.37	0.24	0.74	1.41	2.22
CUN Cui et al. (2021)	**0.22**	0.61	0.97	1.43	**0.20**	0.48	0.77	**1.31**	0.24	0.82	1.51	2.38
GDC Kim et al. (2021)	0.28	0.63	0.97	1.44	0.23	0.49	0.79	1.37	0.28	0.84	1.51	2.43
Ours	0.25	**0.50**	**0.90**	**1.49**	0.27	**0.46**	**0.76**	**1.31**	**0.21**	**0.43**	**1.19**	**1.84**

We report the RMSE values for different scale factors, where the best RMSE for each evaluation is in boldface, whereas the second best one is underlined.

generated by the authors' code or calculated directly from the upsampled depth maps provided by the authors.

From **Tables 1**, **2** it can be seen that deep learning based architectures for SR, such as Guo et al. (2018); Huang et al. (2019); Hui et al. (2016); Zuo et al. (2019a,b); Zhao et al. (2021); Kim et al. (2021); Li et al. (2020); Ye et al. (2020); Cui et al. (2021), have obvious advantages compared with other methods. Moreover, our proposed architecture, benefiting from our CTGM, achieves the best performance on almost all scaling factors in terms of RMSE values. This holds especially for large scaling factors, which are difficult for most methods. The average RMSE values obtained by our method on the whole test set are 0.48/1.04/1.70 for scaling factors x4/x8/x16, respectively. Compared to the second-best results, our results outperform them in terms of average RMSE values with a gain of 0.15/0.14/0.25, respectively. For a scale factor of x2, our method achieved similar average RMSE as the second best method.

To further demonstrate the performance of our method, **Figure 3** shows upsampled depth maps for different approaches on the "Art" and "Reindeer" datasets and a scale factor of 8. Our results are compared with bicubic interpolation as a baseline and 3 state-of-the-art methods which are RDGR (Liu et al., 2016), MSG (Hui et al., 2016), and DSR (Guo et al., 2018). It is observed that our proposed architecture can alleviate the blurring artifacts better and recover more detailed HR depth boundaries than the competing methods. Moreover, our approach can overcome the texture copying issue apparent with other methods, as marked by a red arrow. The evaluations suggest that our CTGM plays an important role in the success of depth map SR.

4.3.2 Results on the Noisy Middlebury Dataset

To further test the robustness of our proposed method, we carry additional experiments on the noisy Middlebury dataset. Depth maps are often corrupted by random noise during the acquisition

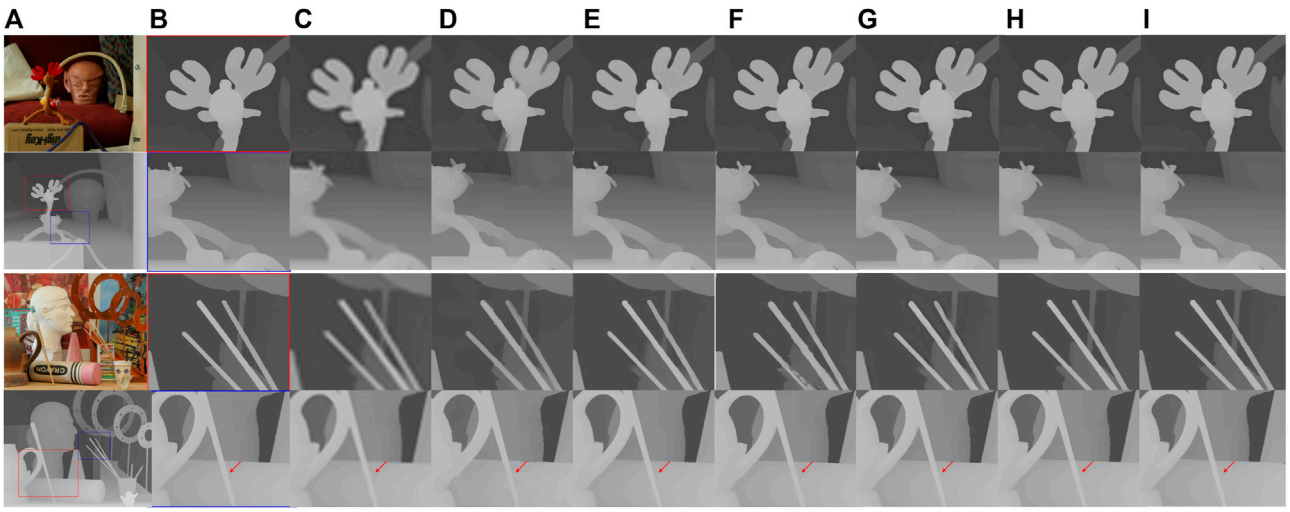

FIGURE 3 | A visual quality comparison for depth map SR at a scale factor of 8 on the noise-free "art" and "Reindeer" datasets. **(A)** HR color and depth images, **(B)** extracted ground truth patches, and upsampled patches by **(C)** Bicubic, **(D)** RDGE (Liu et al., 2016), **(E)** MSG (Hui et al., 2016), **(F)** DSR (Guo et al., 2018) **(G)** RDN (Zuo et al., 2019a) **(H)** PMBA (Ye et al., 2020) **(I)** our proposed method (best viewed on the enlarged electronic version).

TABLE 3 | Quantitative comparisons on the noisy middlebury dataset in terms of RMSE values for scaling factors 4 and 8.

Method	Art		Books		Laundry		Dolls		Moebius		Reindeer	
	x8	x16	x8	x16	x8	x16	x8	x16	x8	x16	x8	x16
Bicubic	6.74	9.04	4.68	5.30	5.35	6.53	4.51	4.90	4.54	5.02	5.71	7.12
TGV Ferstl et al. (2013)	7.26	12.05	2.88	4.73	4.45	8.06	2.82	5.14	3.01	6.11	4.65	9.03
NLMR Park et al. (2014)	8.01	11.01	3.29	4.91	4.51	6.35	3.33	4.45	3.27	4.61	5.33	7.56
MSG Hui et al. (2016)	4.24	7.42	2.48	4.19	3.31	4.88	2.53	3.41	2.47	3.76	3.36	4.95
MFR Zuo et al. (2019b)	3.97	6.14	2.13	3.17	2.82	4.57	2.25	3.30	2.13	3.33	3.01	4.86
RDN Zuo et al. (2019a)	4.09	6.62	2.11	3.36	2.88	5.11	2.33	3.59	2.18	3.69	3.09	4.93
DSR Guo et al. (2018)		6.96		5.66		7.54		4.28		3.39		5.25
RYN Li et al. (2020)	3.47		1.88		2.47		1.97		1.87		2.68	
GDC Kim et al. (2021)	3.31	4.77	1.69	**2.46**	2.20	**3.36**	1.89	2.59	1.72	2.68	2.57	**3.44**
Ours	**3.26**	**4.72**	**1.61**	2.96	**1.63**	3.47	**1.64**	**2.16**	**1.63**	**2.24**	**1.79**	3.59

we report the RMSE values for different scale factors, where the best RMSE for each evaluation is in boldface, whereas the second best one is underlined.

process. Thus we added random Gaussian noise with mean 0 and a SD of 5 to our LR training data, similarly to Guo et al. (2018); Zuo et al. (2019a,b). We retrained all our models and evaluated them on a test set corrupted with the same Gaussian noise. We report the RMSE values for the noisy case in **Table 3**.

It is observed that noise added to the LR depth maps significantly affects the reconstructed HR depth maps across all methods. However, our proposed architecture still manages to outperform competing methods and generate clean and sharp reconstructions.

To further test our method's robustness to noise, we added Gaussian noise with a mean 0 and SD of 5 to the guidance HR color images of our training and test data. This simulates a realistic scenario in which data acquired by both the depth and intensity sensors is corrupted with noise. We retrained our models and report the obtained average RMSE values in **Table 4**.

Table 4 demonstrate our method's insensitivity to noise added to the color image. We observe that models evaluated with noise in both LR depth and HR guidance image achieve very similar results to models evaluated with LR depth noise alone, thus demonstrating the effectiveness of our proposed CTGM.

TABLE 4 | Average RMSE Values of Our Proposed architecture for Different Scaling Factors on Various Datasets.

Middlebury dataset version	x2	x4	x8	x16
Noise-Free	0.23	0.48	1.04	1.70
Depth Noise	1.05	1.37	1.92	3.19
Depth and Color Noise	1.17	1.69	2.08	3.41

4.3.3 Results on NYU Depth v2 Dataset

In this subsection, to demonstrate the generalization ability of our proposed architecture, we carry out experiments on the challenging NYU Depth v2 public benchmark dataset (Silberman et al., 2012). The NYU Depth v2 dataset comprises

TABLE 5 | Quantitative comparisons on the NYU depth v2 dataset in terms of average RMSE values for a scaling factor of 4.

Method	Average RMSE on NYU depth v2 dataset
Bicubic	2.36
ATGV-Net Riegler et al. (2016b)	1.28
MSG Hui et al. (2016)	1.31
RDN Zuo et al. (2019a)	1.21
DSR Guo et al. (2018)	1.34
RYN Li et al. (2020)	1.06
PMBA Ye et al. (2020)	1.06
Ours	**0.95**

we report the RMSE values for different scale factors, where the best RMSE for each evaluation is in boldface, whereas the second best one is underlined.

TABLE 6 | Average inference time (seconds) for different scaling factors.

Method	x2	x4	x8	x16
Bicubic	0.01	0.01	0.01	0.01
TGV Ferstl et al. (2013)	45.73	49.78	46.34	46.20
AR Yang et al. (2014)	158.01	157.73	157.95	158.77
RDGE Liu et al. (2016)	68.07	67.69	68.45	68.17
MSG Hui et al. (2016)	0.26	0.30	0.38	0.42
DSR Guo et al. (2018)	0.22	0.23	0.23	0.23
RYN Li et al. (2020)	0.46	0.63	0.72	0.88
Ours	0.15	0.38	0.48	0.53

1449 high-quality RGB-D image pairs of natural indoors scenes, taken with a Microsoft Kinect camera. In this dataset, there are unavoidable local structural misalignments between depth maps and color images, which may affect the performance of guided SR methods. We note that no RDB-D pair from the NYU Depth v2 dataset was included in the training data of our models.

In **Table 5** we report the RMSE value averaged across all RGB-D pairs in the NYU Depth v2 dataset for a scaling factor of x4, where the best achieved RMSE is boldface. We compare our results with Bicubic interpolation as a baseline and competing guided SR methods; ATGV-Net (Riegler et al., 2016b), MSG (Hui et al., 2016), DSR (Guo et al., 2018), RDN (Zuo et al., 2019a), RYN (Li et al., 2020) and PMBA (Ye et al., 2020). Our proposed architecture achieves the lowest average RMSE, improving over the second-best method by 0.11, demonstrating our proposed method's generalization ability and robustness.

4.3.4 Inference Time

To show the real-world applicability of our proposed method, we compare the inference time of our proposed architecture to competing approaches. Inference times were measured using the same setup described in 4.2 running on a 1320 × 1080 pixels image. We report the average inference times in seconds in **Table 6**.

From **Table 6** we conclude that deep learning based methods, such as our proposed architecture as well as Hui et al. (2016); Li et al. (2020); Guo et al. (2018), achieve significantly faster inference times than traditional methods. Moreover, our proposed method performs similarly to Hui et al. (2016); Guo et al. (2018) and faster then Li et al. (2020) while achieving lower RMSE values. In contrast, the speed of Yang et al. (2014); Liu et al. (2016); Ferstl et al. (2013), which require multiple optimization iterations to achieve good reconstructions, is slower, limiting their practical applications. We also note that methods such as Liu et al. (2016) and Guo et al. (2018) which upsample the LR depth as an initial preprocess step exhibit very similar inference times across different scaling factors.

4.4 Ablation Study

We carry out an ablation study to demonstrate the effectiveness of each component in the proposed architecture. We conduct the following experiments: (1) Our architecture without intensity guidance and the CTGM denoted as "Depth-Only." (2) Our architecture with fewer RDBs in each RDG, i.e., $G = 4$, denoted as "proposed (S)." (3) Our architecture without shifted windows in the transformer block denoted as "proposed w/o ws." (4) Our architecture with absolute position embedding, instead of relative position embedding in the CTGM module denoted as "proposed - ape". In these experiments, we use the same network parameters with the settings as mentioned earlier. We evaluate the performance using average RMSE on our evaluation dataset at scaling factors 4, 8 and 16. As shown in **Table 7**, we observe that: (1) As expected, our guided architecture with CTGM performs better than a non-guided version that operates on depth data alone. (2) Our architecture with fewer RDBs still achieves competitive results. However, it is inferior to our full architecture. This implies that the network's depth also plays a significant role in the success of SR architectures. Our proposed architecture with long and short skip connections and close guidance from the CTGM module enables the effective training of such deep networks. (3) our cascaded transformer with the shifted window partitioning outperforms the counterpart built on a single-window partitioning. The results indicate the effectiveness of using shifted windows to build connections among windows in the preceding layers. (4) Relative position bias improves over absolute position bias, indicating the effectiveness of the relative position bias. Although recent image classification models Dosovitskiy et al. (2020) opted to abandon translation invariance, using an inductive bias that

TABLE 7 | Quantitative comparisons of our ablation experiments. Reported results are average RMSE on the noise-free middlebury dataset for scaling factors 4, 8 and 16.

Depth-only			Proposed (S)			Proposed w/o ws			Proposed - ape			Proposed		
x4	x8	x16	x4	x8	x16	x4	x8	x16	x4	x8	x16	x4	x8	x16
0.55	1.30	2.67	0.57	1.57	2.90	0.51	1.26	2.29	0.51	1.38	2.51	**0.48**	**1.04**	**1.70**

we report the RMSE values for different scale factors, where the best RMSE for each evaluation is in boldface, whereas the second best one is underlined.

encourages certain translation invariance is still preferable for dense prediction tasks such as SR. Moreover, we observe from **Table 7** that the advantages of our complete proposed architecture are more prominent in larger scaling factors, e.g., 8 and 16.

5 CONCLUSION

We have introduced a new method to address the problem of depth map upsampling by using a cascaded transformer module for guided depth SR. An LR depth map is progressively upsampled using residual dilated blocks and a novel guidance module, based on the cascaded transformer that operates on shifted window partitioning of the image, scales the intermediate feature maps of the network. Our proposed architecture achieves state-of-the-art performance for super-resolving depth maps using such a design.

In future work, we intend to explore even more realistic noise and artifacts in our test sets (e.g., missing depth values, misregistration between RGB image and depth map, etc.). Moreover, we will examine the proposed architecture on upsampling Dynamic Elevation Model (DEM) data using LR DEM and HR raster data, which acts as guidance.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://sintel.is.tue.mpg.de/ https://vision.middlebury.edu/stereo/data /https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported by the PMRI—Peter Munk Research Institute - Technion.

REFERENCES

Butler, D. J., Wulff, J., Stanley, G. B., and Black, M. J. (2012). "A Naturalistic Open Source Movie for Optical Flow Evaluation," in European conference on computer vision (Berlin, Germany: Springer), 611–625. doi:10.1007/978-3-642-33783-3_44

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end Object Detection with Transformers," in European Conference on Computer Vision (Berlin, Germany: Springer), 213–229. doi:10.1007/978-3-030-58452-8_13

Cui, Y., Liao, Q., Yang, W., and Xue, J.-H. (2021). "Rgb Guided Depth Map Super-resolution with Coupled U-Net," in 2021 IEEE International Conference on Multimedia and Expo (ICME) (Shenzhen, China: IEEE), 1–6. doi:10.1109/icme51207.2021.9428096

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., et al. (2017). "Deformable Convolutional Networks," in Proceedings of the IEEE international conference on computer vision, 764–773. doi:10.1109/iccv.2017.89

Dong, W., Shi, G., Li, X., Peng, K., Wu, J., and Guo, Z. (2016). Color-guided Depth Recovery via Joint Local Structural and Nonlocal Low-Rank Regularization. *IEEE Trans. Multimedia* 19, 293–301.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An Image Is worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*.

Ferstl, D., Reinbacher, C., Ranftl, R., Rüther, M., and Bischof, H. (2013). "Image Guided Depth Upsampling Using Anisotropic Total Generalized Variation," in Proceedings of the IEEE International Conference on Computer Vision, 993–1000. doi:10.1109/iccv.2013.127

Ferstl, D., Ruther, M., and Bischof, H. (2015). "Variational Depth Superresolution Using Example-Based Edge Representations," in Proceedings of the IEEE International Conference on Computer Vision, 513–521. doi:10.1109/iccv.2015.66

Guo, C., Li, C., Guo, J., Cong, R., Fu, H., and Han, P. (2018). Hierarchical Features Driven Residual Learning for Depth Map Super-resolution. *IEEE Trans. Image Process.* 28, 2545–2557. doi:10.1109/TIP.2018.2887029

Ham, B., Cho, M., and Ponce, J. (2015a). "Robust Image Filtering Using Joint Static and Dynamic Guidance," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4823–4831. doi:10.1109/cvpr.2015.7299115

Ham, B., Dongbo Min, D., and Kwanghoon Sohn, K. (2015b). Depth Superresolution by Transduction. *IEEE Trans. Image Process.* 24, 1524–1535. doi:10.1109/tip.2015.2405342

Haris, M., Shakhnarovich, G., and Ukita, N. (2018). "Deep Back-Projection Networks for Super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition, 1664–1673. doi:10.1109/cvpr.2018.00179

He, K., Sun, J., and Tang, X. (2012). Guided Image Filtering. *IEEE Trans. Pattern Anal. Mach Intell.* 35, 1397–1409. doi:10.1109/TPAMI.2012.213

He, K., Sun, J., and Tang, X. (2010). "Guided Image Filtering," in European conference on computer vision (Berlin, Germany: Springer), 1–14. doi:10.1007/978-3-642-15549-9_1

Hornacek, M., Rhemann, C., Gelautz, M., and Rother, C. (2013). "Depth Super Resolution by Rigid Body Self-Similarity in 3d," in Proceedings of the IEEE conference on computer vision and pattern recognition, 1123–1130. doi:10.1109/cvpr.2013.149

Hu, H., Zhang, Z., Xie, Z., and Lin, S. (2019). "Local Relation Networks for Image Recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 3464–3473. doi:10.1109/iccv.2019.00356

Huang, L., Zhang, J., Zuo, Y., and Wu, Q. (2019). Pyramid-structured Depth Map Super-resolution Based on Deep Dense-Residual Network. *IEEE Signal. Process. Lett.* 26, 1723–1727. doi:10.1109/lsp.2019.2944646

Hui, T.-W., Loy, C. C., and Tang, X. (2016). "Depth Map Super-resolution by Deep Multi-Scale Guidance," in European conference on computer vision (Berlin, Germany: Springer), 353–369. doi:10.1007/978-3-319-46487-9_22

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., et al. (2011). "Kinectfusion: Real-Time 3d Reconstruction and Interaction Using a Moving Depth Camera," in Proceedings of the 24th annual ACM symposium on User interface software and technology, 559–568.

Jiang, Z., Hou, Y., Yue, H., Yang, J., and Hou, C. (2018). Depth Super-resolution from Rgb-D Pairs with Transform and Spatial Domain Regularization. *IEEE Trans. Image Process.* 27, 2587–2602. doi:10.1109/tip.2018.2806089

Jun Xie, J., Feris, R. S., and Ming-Ting Sun, M.-T. (2015). Edge-guided Single Depth Image Super Resolution. *IEEE Trans. Image Process.* 25, 428–438. doi:10.1109/TIP.2015.2501749

Kiechle, M., Hawe, S., and Kleinsteuber, M. (2013). "A Joint Intensity and Depth Co-sparse Analysis Model for Depth Map Super-resolution," in Proceedings of the IEEE international conference on computer vision, 15451552. doi:10.1109/iccv.2013.195

Kim, J.-Y., Ji, S., Baek, S.-J., Jung, S.-W., and Ko, S.-J. (2021). Depth Map Super-resolution Using Guided Deformable Convolution. *IEEE Access* 9, 66626–66635. doi:10.1109/access.2021.3076853

Kingma, D. P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Kwon, H., Tai, Y.-W., and Lin, S. (2015). "Data-driven Depth Map Refinement via Multi-Scale Sparse Representation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 159–167. doi:10.1109/cvpr.2015.7298611

Lei, J., Li, L., Yue, H., Wu, F., Ling, N., and Hou, C. (2017). Depth Map Super-resolution Considering View Synthesis Quality. *IEEE Trans. Image Process.* 26, 1732–1745. doi:10.1109/tip.2017.2656463

Li, T., Dong, X., and Lin, H. (2020). Guided Depth Map Super-resolution Using Recumbent Y Network. *IEEE Access* 8, 122695–122708. doi:10.1109/access.2020.3007667

Liu, M.-Y., Tuzel, O., and Taguchi, Y. (2013). "Joint Geodesic Upsampling of Depth Images," in Proceedings of the IEEE conference on computer vision and pattern recognition, 169–176. doi:10.1109/cvpr.2013.29

Liu, W., Chen, X., Yang, J., and Wu, Q. (2016). Robust Color Guided Depth Map Restoration. *IEEE Trans. Image Process.* 26, 315–327. doi:10.1109/TIP.2016.2612826

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in International Conference on Computer Vision (ICCV).

Lu, J., and Forsyth, D. (2015). "Sparse Depth Super Resolution," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2245–2253. doi:10.1109/cvpr.2015.7298837

Lutio, R. d., D'aronco, S., Wegner, J. D., and Schindler, K. (2019). "Guided Super-resolution as Pixel-To-Pixel Transformation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 8829–8837. doi:10.1109/iccv.2019.00892

Mac Aodha, O., Campbell, N. D. F., Nair, A., and Brostow, G. J. (2012). "Patch Based Synthesis for Single Depth Image Super-resolution," in European conference on computer vision (Berlin, Germany: Springer), 71–84. doi:10.1007/978-3-642-33712-3_6

Mandal, S., Bhavsar, A., and Sao, A. K. (2016). Depth Map Restoration from Undersampled Data. *IEEE Trans. Image Process.* 26, 119–134. doi:10.1109/TIP.2016.2621410

Park, J., Kim, H., Tai, Y.-W., Brown, M. S., and Kweon, I. (2011). "High Quality Depth Map Upsampling for 3d-Tof Cameras," in 2011 International Conference on Computer Vision (Barcelona, Spain: IEEE), 1623–1630. doi:10.1109/iccv.2011.6126423

Park, J., Kim, H., Tai, Y.-W., Brown, M. S., and Kweon, I. S. (2014). High-quality Depth Map Upsampling and Completion for Rgb-D Cameras. *IEEE Trans. Image Process.* 23, 5559–5572. doi:10.1109/tip.2014.2361034

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: An Imperative Style, High-Performance Deep Learning Library," in Advances in Neural Information Processing Systems, 8024–8035.

Riegler, G., Ferstl, D., Rüther, M., and Bischof, H. (2016a). A Deep Primal-Dual Network for Guided Depth Super-resolution. *arXiv preprint arXiv:1607.08569*. doi:10.5244/c.30.7

Riegler, G., Rüther, M., and Bischof, H. (2016b). "Atgv-net: Accurate Depth Super-resolution," in European conference on computer vision (Berlin, Germany: Springer), 268–284. doi:10.1007/978-3-319-46487-9_17

Schamm, T., Strand, M., Gumpp, T., Kohlhaas, R., Zollner, J. M., and Dillmann, R. (2009). "Vision and Tof-Based Driving Assistance for a Personal Transporter," in 2009 International Conference on Advanced Robotics (Munich, Germany: IEEE), 1–6.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., et al. (2014). "High-resolution Stereo Datasets with Subpixel-Accurate Ground Truth," in German conference on pattern recognition (Berlin, Germany: Springer), 31–42. doi:10.1007/978-3-319-11752-2_3

Scharstein, D., and Pal, C. (2007). "Learning Conditional Random fields for Stereo," in 2007 IEEE Conference on Computer Vision and Pattern Recognition (Minneapolis, MN, USA: IEEE), 1–8. doi:10.1109/cvpr.2007.383191

Scharstein, D., and Szeliski, R. (2002). A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *Int. J. Comput. Vis.* 47, 7–42. doi:10.1023/a:1014573219977

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., et al. (2016). "Real-time Single Image and Video Super-resolution Using an Efficient Sub-pixel Convolutional Neural Network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 1874–1883. doi:10.1109/cvpr.2016.207

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). "Indoor Segmentation and Support Inference from Rgbd Images," in European conference on computer vision (Berlin, Germany: Springer), 746–760. doi:10.1007/978-3-642-33715-4_54

Song, X., Dai, Y., and Qin, X. (2018). Deeply Supervised Depth Map Super-resolution as Novel View Synthesis. *IEEE Trans. circuits Syst. video Technol.* 29, 2323–2336.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention Is All You Need," in Advances in neural information processing systems, 5998–6008.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., et al. (2021). Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. *arXiv preprint arXiv:2102.12122*.

Yang, J., Ye, X., Li, K., and Hou, C. (2012). "Depth Recovery Using an Adaptive Color-Guided Auto-Regressive Model," in European conference on computer vision (Berlin, Germany: Springer), 158–171. doi:10.1007/978-3-642-33715-4_12

Yang, J., Ye, X., Li, K., Hou, C., and Wang, Y. (2014). Color-guided Depth Recovery from Rgb-D Data Using an Adaptive Autoregressive Model. *IEEE Trans. Image Process.* 23, 3443–3458. doi:10.1109/tip.2014.2329776

Yang, Q., Yang, R., Davis, J., and Nistér, D. (2007). "Spatial-depth Super Resolution for Range Images," in 2007 IEEE Conference on Computer Vision and Pattern Recognition (Minneapolis, MN, USA: IEEE), 1–8. doi:10.1109/cvpr.2007.383211

Ye, X., Sun, B., Wang, Z., Yang, J., Xu, R., Li, H., et al. (2020). Pmbanet: Progressive Multi-branch Aggregation Network for Scene Depth Super-resolution. *IEEE Trans. Image Process.* 29, 7427–7442. doi:10.1109/tip.2020.3002664

Zhang, K., Zuo, W., Gu, S., and Zhang, L. (2017). "Learning Deep Cnn Denoiser Prior for Image Restoration," in Proceedings of the IEEE conference on computer vision and pattern recognition, 3929–3938. doi:10.1109/cvpr.2017.300

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. (2018a). "Image Super-resolution Using Very Deep Residual Channel Attention Networks," in Proceedings of the European Conference on Computer Vision (ECCV), 286–301. doi:10.1007/978-3-030-01234-2_18

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., and Fu, Y. (2018b). "Residual Dense Network for Image Super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2472–2481. doi:10.1109/cvpr.2018.00262

Zhao, Z., Zhang, J., Xu, S., Zhang, C., and Liu, J. (2021). Discrete Cosine Transform Network for Guided Depth Map Super-resolution. *arXiv preprint arXiv:2104.06977*.

Zhou, W., Li, X., and Reynolds, D. (2017). "Guided Deep Network for Depth Map Super-resolution: How Much Can Color Help?," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (New Orleans, LA, USA: IEEE), 1457–1461.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2020). Deformable Detr: Deformable Transformers for End-To-End Object Detection. *arXiv preprint arXiv:2010.04159*.

Zuo, Y., Fang, Y., Yang, Y., Shang, X., and Wang, B. (2019a). Residual Dense Network for Intensity-Guided Depth Map Enhancement. *Inf. Sci.* 495, 52–64. doi:10.1016/j.ins.2019.05.003

Zuo, Y., Wu, Q., Fang, Y., An, P., Huang, L., and Chen, Z. (2019b). "Multi-scale Frequency Reconstruction for Guided Depth Map Super-resolution via Deep Residual Network," in IEEE Transactions on Circuits and Systems for Video Technology.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Chapter 6

Fully Cross-Attention Transformer for Guided Depth Super Resolution

sensors

MDPI

Article

Fully Cross-Attention Transformer for Guided Depth Super-Resolution

Ido Ariav * and Israel Cohen *

Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering, Technion—Israel Institute of Technology, Haifa 3200003, Israel

* Correspondence: idoariav@campus.technion.ac.il (I.A.); icohen@ee.technion.ac.il (I.C.)

Abstract: Modern depth sensors are often characterized by low spatial resolution, which hinders their use in real-world applications. However, the depth map in many scenarios is accompanied by a corresponding high-resolution color image. In light of this, learning-based methods have been extensively used for guided super-resolution of depth maps. A guided super-resolution scheme uses a corresponding high-resolution color image to infer high-resolution depth maps from low-resolution ones. Unfortunately, these methods still have texture copying problems due to improper guidance from color images. Specifically, in most existing methods, guidance from the color image is achieved by a naive concatenation of color and depth features. In this paper, we propose a fully transformer-based network for depth map super-resolution. A cascaded transformer module extracts deep features from a low-resolution depth. It incorporates a novel cross-attention mechanism to seamlessly and continuously guide the color image into the depth upsampling process. Using a window partitioning scheme, linear complexity in image resolution can be achieved, so it can be applied to high-resolution images. The proposed method of guided depth super-resolution outperforms other state-of-the-art methods through extensive experiments.

Keywords: super-resolution; deep learning; depth maps; attention; multimodal; transformers

Citation: Ariav, V.; Cohen, I. Fully Cross-Attention Transformer for Guided Depth Super-Resolution. *Sensors* **2023**, *23*, 2723. https://doi.org/10.3390/s23052723

Academic Editors: Sukho Lee and Dae-Ki Kang

Received: 16 February 2023
Revised: 26 February 2023
Accepted: 27 February 2023
Published: 2 March 2023

1. Introduction

High-resolution (HR) depth information of a scene plays a significant part in many applications, such as 3D reconstruction [1], driving assistance [2], and mobile robots. Nowadays, depth sensors such as LIDAR or time-of-flight cameras are becoming more widely used. However, they often suffer from low spatial resolution, which does not always suffice for real-world applications. Thus, ongoing research has been done on reconstructing a high-resolution depth map from a corresponding low-resolution (LR) counterpart in a process termed depth super-resolution (DSR).

The LR depth map does not contain the fine details of the HR depth map, so reconstructing the HR depth map can be challenging. Bicubic interpolation, for example, often produces blurry depth maps when upsampling the LR depth, which limits the ability to, e.g., separate between different objects in the scene.

In recent years, many learning-based approaches based on elaborate convolutional neural network (CNN) architectures for DSR were proposed [3–7]. These methods surpassed the more classic approaches such as filter-based methods [8,9], and energy minimization-based methods [10–12] in terms of computation speed and the quality of the reconstructed HR information. Although CNN-based methods improved the performance significantly compared with traditional methods, they still suffer from several drawbacks. To begin with, feature maps derived from a convolution layer have a limited receptive field, making long-range dependency modeling difficult. Second, a kernel in a convolution layer operates similarly on all parts of the input, making it content-independent and likely not the optimal choice. In contrast to CNN, transformers [13] have recently shown promising results in

several vision-related tasks due to their use of attention. The attention mechanism enables the transformer to operate in a content-dependent manner, where each input part is treated differently according to the task.

LR depth information is often accompanied by HR color or intensity images in real-life situations. Thus, numerous methods proposed to use this HR image to guide the DSR process [3,4,7,14–23] since the HR image might provide some additional information that does not exist in the LR depth image, e.g., the edges of a color image can be used to identify discontinuities in a reconstructed HR depth image. However, one major limitation, termed texture-copying, still exists in these guided DSR methods. Texture copying may occur when intensity edges do not correspond to depth discontinuities in-depth maps, for example, a flat and highly textured surface. Consequently, the reconstruction of HR depth is then degraded due to the over-transfer of texture information.

This paper proposes a novel, fully transformer-based architecture for guided DSR. Specifically, the proposed architecture consists of three modules: shallow feature extraction, deep feature extraction and fusion, and an upsampling module. In this paper, we term the feature extraction and fusion module the cross-attention guidance module (CAGM). The shallow feature extraction module uses convolutional layers to extract shallow features from LR depth and HR color images, which are directly fed to the CAGM to preserve low-frequency information. Next, several transformer blocks are stacked to form the CAGM, each operating in non-overlapping windows from the previous block. Guidance from the color image is introduced via a cross-attention mechanism. In this manner, guidance from the HR color image is seamlessly integrated into the deep feature extraction process. This enables the network to focus on salient and meaningful features and enhance the edge structures in the depth features while suppressing textures in the color features. Moreover, contrary to CNN-based methods, which can only use local information, transformer blocks can exploit the input image's local and global information. This allows learning of structure and content from a wide receptive field, which is beneficial for SR tasks [24]. As a final step, shallow and deep features are fused in the upsampling module to reconstruct HR depth. Section 4 shows that the proposed architecture provides better visual results with sharper boundaries and better root mean square error (RMSE) values than competing guided DSR approaches. We also show how the proposed architecture helps to mitigate the texture-copying problem of guided DSR. The proposed architecture is shown in Figure 1.

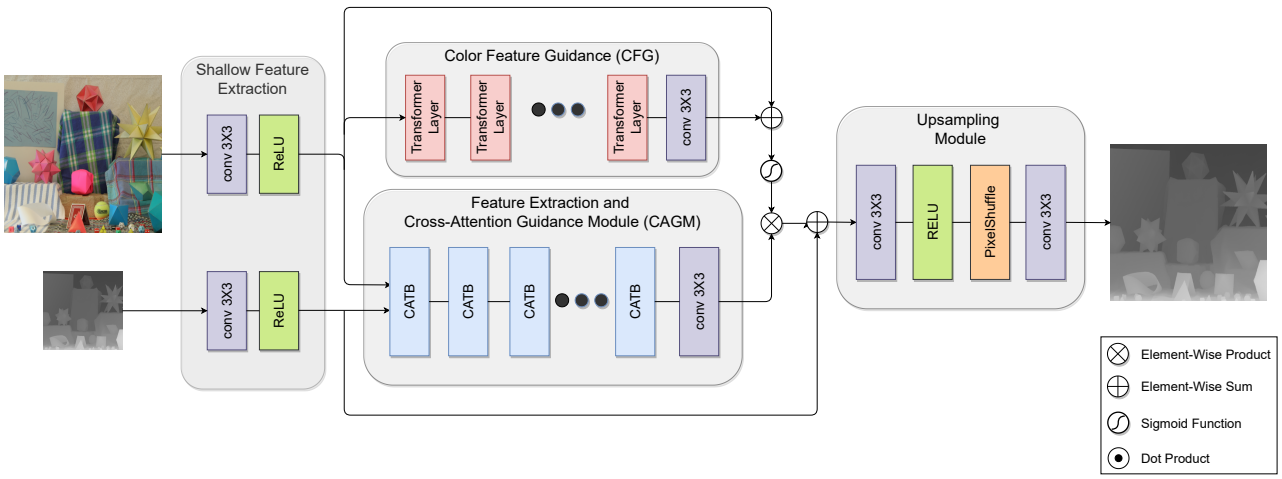

Figure 1. The proposed FCTN architecture for guided depth SR with a 2× upsampling factor.

Our main contributions are as follows: (1) We introduce a transformer-based architecture with a novel guidance mechanism that leverages cross-attention to seamlessly integrate guidance features from a color image to the DSR process. (2) Linear memory constraints make the proposed architecture applicable even for large inputs. (3) This architecture is not

limited to a fixed input size, so it can be applied to a variety of real-world problems. (4) Our system achieves state-of-the-art results on several depth-upsampling benchmarks.

The remainder of this paper is organized as follows. A summary of related work is presented in Section 2. We describe our architecture for guided DSR in Section 3. Section 4 reports the results of extensive experiments conducted on several popular DSR datasets. Additionally, an ablation study is conducted. We conclude and discuss future research directions in Section 5.

2. Related Work

2.1. Guided Depth Map Super-Resolution

A number of methods for reconstructing the HR depth map only from LR depth have been proposed in earlier works for single depth map SR. ATGV-Net was proposed by [5] combining a deep CNN in tandem with a variational method designed to facilitate the recovery of the HR depth map. Reference [25] modeled the mapping between HR and LR depth maps by utilizing densely connected layers coupled with a residual learning model. Auxiliary losses were tabulated at various scales to improve training.

However, it is pertinent to note that in most real-life scenarios, the LR depth image is coupled with a HR intensity image. Recently, several methods have been proposed to improve depth image quality, relying on the HR intensity image to guide the upsampling process. We group these methods under four sub-categories: filtering-based methods [26–28], global optimization-based methods [10–12,16,29–34], sparse representation-based methods [14,15], and deep learning-based methods [3,4,7,17–23,35–40], which are the focus of this paper.

Notable amongst the more classical works are [10], which formulated the upsampling of depth as a convex optimization problem. The upsampling process was guided by a HR intensity image. A bimodal co-sparse analysis was presented in [14] to describe the interdependency of the registered intensity and depth information. Reference [15] proposed a multi-scale dictionary as a method for depth map refinement, where local patches were represented in both depth and color via a combination of select basis functions.

Deep learning methods for SR of depth images have gained increasing attention due to recent success in SR of color images. A fully convolutional network was proposed in [35] to estimate the HR depth. To optimize the final result, this HR estimation was fed into a non-local variational model. Reference [4] proposed an "MSG-Net", in which both LR (depth) and HR (color) features are combined within the high-frequency domain using a multi-scale fusion strategy. Reference [3] proposed extracting hierarchical features from depth and color images by building a multi-scale input pyramid. The hierarchical features are further concatenated to facilitate feature fusion, whilst the residual map between the reconstructed and ground truth HR depth is learned with a U-Net architecture. Reference [37] proposed another multi-scale network in which the LR depth map upsampling, guided by the HR color image, was performed in stages. Global and local residual learning is applied within each scale. Reference [17] proposed a cosine transform network in which features from both depth and color images were extracted using a semi-coupled feature extraction module. To improve depth upsampling, edges were highlighted by an edge attention mechanism operating on color features. Reference [19] proposed to use deformable convolutions [41] for the upsampling of depth maps, using the features of the HR guidance image to determine the spatial offsets. Reference [42] also applied deformable convolutions to enhance depth features by learning the corresponding feature of the high-resolution color image. An adaptive feature fusion module was used to fuse different level features adaptively. A network based on residual channel attention blocks was proposed in [20], where feature fusion blocks based on spatial attention were utilized to suppress texture-copying. Reference [21] proposed a progressive multi-branch aggregation design that gradually restores the degraded depth map. Reference [22] proposed separate branches for HR color image and LR depth map. A dual-skip connection structure, together with a multi-scale fusion strategy, allowed for more effective features to be learned. Reference [39] used a

transformer module to learn the useful content and structure information from the depth maps and the corresponding color images, respectively. Then, a multi-scale fusion strategy was used to improve the efficiency of color-depth fusion. Reference [43] proposed explicitly incorporating the depth gradient features in the DSR process. Reference [44] proposed PDR-Net, which incorporates an adaptive feature recombination module to adaptively recombine features from a HR color guidance image with features from the LR depth. Then, a multi-scale feature fusion module is used to fuse the recombined features using multi-scale feature distillation and joint attention mechanism. Finally, Reference [23] presented an upsampling method that incorporates the intensity image's high-resolution structural information into a multi-scale residual deep network via a cascaded transformer module.

However, the methods above mostly fuse the guidance features with the depth features using mere concatenation. Moreover, most of these methods rely on CNN for feature extraction, which operates on a limited receptive field and lacks the expressive power of transformers. At the same time, we propose using a CAGM module, which leverages transformers to fuse and extract meaningful features from HR color and LR depth images, resulting in superior results, as shown in Section 4.

2.2. Vision Transformers

Transformers [13] have gained great success across multiple domains recently. Contributing to this success was their inherent attention mechanism, which enables them to learn the long-range dependencies in the data. This success led many researchers to adopt transformers to computer vision tasks, where they have recently demonstrated promising results, specifically in image classification [45–47], segmentation [47,48], and object detection [49,50].

To allow transformers to handle 2D images, an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ is first divided into non-overlapping patches of size (P, P). Each patch is flattened and projected to a d-dimensional vector via a trainable linear projection, forming the patch embeddings $\mathbf{X} \in \mathbb{R}^{N \times d}$ where H, W are the height and width of the image, respectively, C is the number of channels, and $N = H \times W / P^2$ is the total number of patches. Finally, N is the effective input sequence length for the transformer encoder. Patch embeddings are enhanced with position embeddings to retain 2D image positional information.

In [13], a vanilla vision transformer encoder is constructed by stacking blocks of multi-head self-attention (MSA) and MLP layers. A residual connection is applied after every block, and layer normalization (LN) before every block. Given a sequence of embeddings $\mathbf{X} \in \mathbb{R}^{N \times d}$ with dimension d as input, a MSA block produces an output sequence $\bar{\mathbf{X}} \in \mathbb{R}^{N \times d}$ via

$$\begin{aligned} \mathbf{Q} &= \mathbf{X}\mathbf{W}_Q\,, \mathbf{K} = \mathbf{X}\mathbf{W}_K\,, \mathbf{V} = \mathbf{X}\mathbf{W}_V \\ \mathbf{A} &= \text{Softmax}(\mathbf{Q}\mathbf{K}^T / \sqrt{\mathbf{d}}) \\ \bar{\mathbf{X}} &= \mathbf{A}\mathbf{V} \end{aligned} \tag{1}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable matrices of size $d \times d$ that project the sequence $\mathbf{X}$ into keys, queries, and values, respectively. $\bar{\mathbf{X}}$ is a linear combination of all the values in $\mathbf{V}$ weighted by the attention matrix $\mathbf{A}$. In turn, $\mathbf{A}$ is calculated from similarities between the keys and query vectors.

Transformers derive their modeling capabilities from computing self-attention $\mathbf{A}$ and $\bar{\mathbf{X}}$. Since self-attention has a quadratic cost in time and space, it cannot be applied directly to images as N quickly becomes unmanageable. As a result of this inherent limitation, modality-aware sequence length restrictions have been applied to preserve the model's performance while restricting sequence length. Reference [45] showed that a transformer architecture could be directly applied to medium-sized image patches for different vision tasks. The aforementioned memory constraints are mitigated by this local self-attention.

Although the above self-attention module can effectively exploit intra-modality relationships in the input image, in a multi-modality setting, the inter-modality relationships, e.g., the relationships between different modalities, also need to be explored. Thus, a

cross-attention mechanism was introduced in which attention masks from one modality highlight the extracted features in another. Contrary to self-similarity, wherein query, key, and value are based on similarities within the same feature array, in cross-attention, keys, and values are calculated from features extracted from one modality, while queries are calculated from the other. Formally, a MSA block using cross-attention is given by

$$\mathbf{Q} = \hat{\mathbf{X}}\mathbf{W}_Q\,, \mathbf{K} = \mathbf{X}\mathbf{W}_K\,, \mathbf{V} = \mathbf{X}\mathbf{W}_V \tag{2}$$

where $\mathbf{X}$ is the input sequence of one modality and $\hat{\mathbf{X}}$ is the input sequence of the second modality. The calculation of attention matrix $\mathbf{A}$ and output sequence $\bar{\mathbf{X}}$ remains the same.

3. Proposed Method

3.1. Formulation

A guided DSR method aims to establish the nonlinear relation between corresponding LR and HR depth maps. The process of establishing this nonlinear relation is guided by a HR color image. We denote the LR depth map as $\mathbf{D}_{\mathrm{LR}} \in \mathbb{R}^{H/s \times W/s}$ and the HR guidance color image as $\mathbf{I}_{\mathrm{HR}} \in \mathbb{R}^{H \times W}$, where s is a given scaling factor. The corresponding HR depth map $\mathbf{D}_{\mathrm{HR}} \in \mathbb{R}^{H \times W}$ can be found from:

$$\mathbf{D}_{\mathrm{HR}} = \hat{\mathbf{F}}(\mathbf{D}_{\mathrm{LR}}, \mathbf{I}_{\mathrm{HR}}; \theta) \tag{3}$$

where $\hat{\mathbf{F}}$ represents mapping learned by the proposed architecture, and θ represents the parameters of the learned network. Although the scaling factor s is usually an exponent of 2, e.g., $s = 2, 4, 8, 16$, our upsampling module can perform upsampling for other scaling factors as well, making this architecture flexible enough for real applications.

3.2. Overall Network Architecture

Throughout the remainder of this paper, we denote the proposed architecture as the fully cross-attention transformer network (FCTN). As shown in Figure 1, the proposed architecture consists of three parts: a shallow feature extraction module, a deep feature extraction and guidance module called the cross-attention guidance module (CAGM), and an upsampling module. The CAGM extracts features from the LR depth image and guides the HR intensity image simultaneously.

Before we elaborate on the structure of each module, some significant challenges in leveraging transformers' performance for visual tasks, specifically SR, need to be addressed. First, in real-life scenarios, images can vary considerably in scale. Transformer-based models, however, work only with tokens of a fixed size. Furthermore, to maintain HR information, SR methods avoid downscaling the input as much as possible. Processing HR inputs of this magnitude would be unfeasible for vanilla transformers due to computational complexity as described in Section 2.2.

3.2.1. Shallow Feature Extraction Module

The proposed shallow feature extraction module extracts essential features to be fed to the CAGM. Shallow features are extracted from LR depth and HR color images via a single convolution layer with a kernel size of 3×3, followed by an activation function of a rectified linear unit (ReLU). In the experiments, we did not notice any noticeable improvement by using more than a single layer for shallow feature extraction. For shallow feature extraction, incorporating a convolution layer leads to more stable optimization and better results [51–53]. Moreover, the input space can also be mapped to a higher-dimensional feature space d easily.

Specifically, the shallow feature extraction module can be formulated as

$$\mathbf{I}_0 = \sigma(\mathbf{Conv}_3(\mathbf{I}_{\mathrm{HR}})) \tag{4}$$

$$\mathbf{D}_0 = \sigma(\mathbf{Conv}_3(\mathbf{D}_{\mathrm{LR}})) \tag{5}$$

where σ is a ReLU activation function and $\mathbf{Conv}_3(\cdot)$ is a 3×3 kernel.

3.2.2. Deep Feature Extraction and Guidance Module

While shallow features primarily contain low frequencies, deep features recover lost high frequencies. We propose a stacked transformer module that extracts deep features from the LR depth image based on the work of [47]. Self(cross)-attention is computed locally within non-overlapping windows, with complexity linear with image size. Working with large and variable-sized inputs is made feasible due to the aforementioned linear complexity. In addition, we shift the windows partitioning into consecutive layers. Overlapping of the shifted and preceding layer windows causes neighboring patches to gradually merge, and thus modeling power is significantly enhanced. Overall, the transformer-based module can efficiently extract and encode distant dependencies needed for dense tasks such as SR.

In addition, motivated by [54], we employ global and local skip connections. By using long skip connections, low-frequency information can be transmitted directly to the upsampling module, helping the CAGM focus on high-frequency information and stabilize training [51]. Furthermore, it allows the aggregation of features from different blocks by using such identity-based connections.

Besides deep feature extraction, a practical guidance module is also required to enhance the deep features extracted from LR depth and exploit the inter-modality information from the available HR color image. Traditionally, CNN-based methods extract features from the color image and concatenate them with features extracted from the depth image in a separate branch to obtain guidance from the color image. All features handled via this guidance scheme are treated equally in both the spatial and channel domains. Furthermore, CNN-derived feature maps have a limited receptive field, affecting guidance quality. In comparison, we propose providing guidance from the HR color image by incorporating a cross-attention mechanism to the aforementioned feature extraction transformer module. Cross-attention is a novel and intuitive fusion method in which attention masks from one modality highlight the extracted features in another. In this manner, both the inter-modality and intra-modality relationships are learned and optimized in a unified model. Thus, in the proposed CAGM, the feature extraction process from the LR depth and guidance from the HR image are seamlessly integrated into a single branch. Guidance from the HR image is injected into every block in the feature extraction module, providing multi-stage guidance. In particular, guidance provided to the lower-level features passed through the long skip connections ensures that high-resolution information is preserved and passed to the upsampling module. Lastly, by incorporating the guidance in the form of cross-attention, long-range dependencies between the LR depth patches and the guidance image patches can be exploited for better feature extraction.

To exploit the HR information further, we feed the HR intensity image to a second cascaded transformer module termed color feature guidance (CFG) to extract even more valuable HR information. The CFG is based on self-attention only and aims to encode distant dependencies in the HR image. These features are then used to scale the features extracted from the CAGM by element-wise multiplication before feeding them to the upsampling module.

We note that contrary to common practice in vision tasks, no downsampling of the input is done throughout the network. This way, our architecture preserves as much high-resolution information as possible, albeit at a higher computational cost.

Formally, given $\mathbf{I}_0$ and $\mathbf{D}_0$, provided by the shallow feature extraction module as input, the CAGM applies K cross-attention transformer blocks (CATB). Every CATB is constructed from L cross-attention transformer layers (CATL), and a convolutional layer and residual skip connection are inserted at the end of every such block. Finally, a 3×3 convolutional layer is applied to the output of the last CATB. This last convolutional layer improves the later aggregation of shallow and deep features by bringing the inductive bias of the convolution operation into the transformer-based network. Furthermore, the translational equivariance of the network is enhanced. In addition, $\mathbf{I}_0$ is fed to the CFG comprised of $\hat{L}$

transformer layers with self-attention. The CFG output is scaled to $[0, 1]$ using a sigmoid function and then used to scale the CAGM output before the upsampling module

The CFG module is formulated as

$$\hat{\mathbf{I}}_{\mathrm{i}} = \mathbf{TL}_i(\hat{\mathbf{I}}_{\mathrm{i}-1}), \quad i = 1 \dots \hat{L} \tag{6}$$

$$\mathbf{F}_{\mathrm{CFG}} = \mathbf{Conv}_3(\hat{\mathbf{I}}_{\hat{L}}) + \mathbf{I}_0 \tag{7}$$

where $\hat{\mathbf{I}}_0 = \mathbf{I}_0$ and TL stands for a vanilla transformer layer with self-attention. Finally, the entire CAGM can be formulated as

$$(\mathbf{I}_{\mathrm{i}}, \mathbf{D}_{\mathrm{i}}) = \mathbf{CATB}_i(\mathbf{I}_{\mathrm{i}-1}, \mathbf{D}_{\mathrm{i}-1}), \quad i = 1 \dots K \tag{8}$$

$$\mathbf{F}_{\mathrm{CAGM}} = \mathbf{Conv}_3(\mathbf{CATB}_{\mathrm{K}}) \otimes \hat{\sigma}(\mathbf{F}_{\mathrm{CFG}}) + \mathbf{D}_0 \tag{9}$$

where $\otimes$ is element-wise multiplication, $\mathbf{Conv}_3(\cdot)$ is a convolution layer with a 3×3 kernel and $\hat{\sigma}$ is a sigmoid function.

3.2.3. Cross-Attention Transformer Layer

The proposed cross-attention transformer layer (CATL) is modified from the standard MSA block presented in [13]. The two significant differences are; First, we use a cross-attention mechanism instead of self-attention. We demonstrate the effectiveness of using a cross-attention mechanism in Section 4.4. Second, cross-attention is computed locally for each window, ensuring linear complexity with image size, which makes it feasible for large inputs to be handled, as is often the case in SR.

Given as input feature map $\mathbf{F} \in \mathbb{R}^{\hat{H} \times \hat{W} \times d}$ extracted from either color or depth images, we first construct $\mathbf{F_{win}} \in \mathbb{R}^{\hat{H}\hat{W}/M^2 \times M^2 \times d}$ by partitioning $\mathbf{F}$ into $M \times M$ non-overlapping windows. Zero padding is applied during the partitioning process if necessary. Similarly to [55], relative position embeddings are added to $\mathbf{F_{win}}$ so that positional information can be retained. In a similar manner, the process is performed for both color and depth feature maps; we refer to this joint embedding as $\mathbf{Z}_I^0$ and $\mathbf{Z}_D^0$ for the color and depth, respectively.

In each CATL, the MSA module is replaced with a windows-based cross-attention MSA ($\mathrm{MSA_{ca}}$), while the other layers remain unchanged. By applying Equation (2) locally within each $M \times M$ window, we avoid global attention computations. Moreover, keys and values are calculated from the depth feature map, while the queries are calculated from the color feature map. Specifically, as illustrated in Figure 2b, our modified CATL consists of $\mathrm{MSA_{ca}}$ followed by a 2-layer MLP with GELU nonlinearity. Every MLP and $\mathrm{MSA_{ca}}$ module is preceded by an LN layer, and each module is followed by a residual connection.

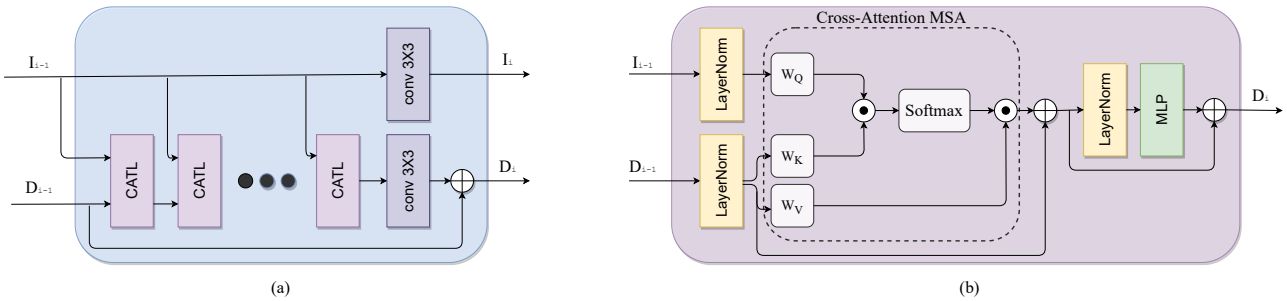

Figure 2. (**a**) Cross-Attention Transformer Block. (**b**) Cross-Attention Transformer Layer.

To enable cross-window connections in consecutive layers, regular and shifted window partitionings are used alternately. In shifted window partitioning, features are shifted by $M/2, M/2$ pixels. Finally, the CATL can be formalized as

$$\hat{\mathbf{Z}} = \mathbf{MSA_{ca}}(\mathbf{LN}(\mathbf{Z}_I^0, \mathbf{Z}_D^0)) + \mathbf{Z}_D^0 \tag{10}$$

$$\mathbf{Z} = \mathbf{MLP}(\mathbf{LN}(\hat{\mathbf{Z}}^1)) + \hat{\mathbf{Z}}^1 \tag{11}$$

where $\hat{\mathbf{Z}}$ and $\mathbf{Z}$ denote the output features of the MSA_{ca} and MLP modules, respectively.

3.2.4. Upsampling Module

The upsampling module operates on the CAGM output, scaled via the CFG module, as elaborated in Section 3.2.2. It aims to recover high-frequency details and reconstruct the HR depth successfully. The CAGM output is first passed through a convolution layer followed by a ReLU activation function to aggregate shallow and deep features from the CAGM. Then, we use a pixel shuffle module [56] to upsample the feature map to the HR resolution. Each pixel shuffle module can perform upsampling by a factor of two or three, and we concatenate such modules according to the desired scaling factor. Finally, the upsampled feature maps are passed through another convolution layer that outputs the reconstructed depth. The parameters of the entire upsampling module are learned in the training process to improve model representation.

Formally, given the output of the CAGM module $\mathbf{F}_{\text{CAGM}} \in \mathbb{R}^{H/s \times W/s}$, where s is the scaling factor, the upsampling module performs an upsampling by a factor s to reconstruct $\mathbf{D}_{\text{HR}} \in \mathbb{R}^{H \times W}$. The upsampling process for a given s can be formulated as follows:

$$\begin{aligned} \mathbf{F}_{\text{US},0} &= \mathbf{Conv}_3(\mathbf{F}_{\text{CAGM}}) \\ \mathbf{F}_{\text{US},i} &= \text{PixellShuffle}_i(\mathbf{F}_{\text{US},i-1}), \quad i = 1 \dots \log_2(s) \\ \mathbf{D}_{\text{HR}} &= \mathbf{Conv}_3(\mathbf{F}_{\text{US},i}). \end{aligned} \tag{12}$$

where $\mathbf{Conv}_3(\cdot)$ is a convolution layer with a 3×3 kernel. More implementation details are given in Section 4.1.

4. Experiments

4.1. Training Details

We constructed train and test data similarly to [3,4,23,25] using 92 pairs of depth and color images from the MPI Sintel depth dataset [57] and the Middlebury depth dataset [58–60]. The training and validation pairs used in this study are similar to the ones used in [4,23]. We refer the reader to [57,58] for further information on the data included in the aforementioned datasets.

During training, we randomly extracted patches from the full-resolution images and used these as input to the network. We used an LR patch size of 96×96 pixels to reduce memory requirements and computation time since using larger patches had no significant impact on training accuracy. Consequently, we used HR patches of 192×192 and 384×384 for upsampling factors of 2 and 4, respectively. Given that some full-scale images had a full resolution of < 400, we used LR patch sizes of 48×48 and 24×24 for upsampling factors 8 and 16, respectively. In order to generate the LR patches, each HR patch was downsampled with bicubic interpolation. As an augmentation method, we used a random horizontal flip while training.

4.2. Implementation Details

We construct the CAGM module in the proposed architecture by stacking $K = 6$ CATBs. Each CATB is constructed from $L = 6$ CATL modules as described in Section 3.2.2. These values for K and L provided the best performance to network size trade-off in the experiments, and Section 4.4, we report results with other configurations. All convolution layers have a stride of one with zero padding, so the features' size remains fixed. Throughout the network, in convolution and transformer blocks, we use a feature (embedding) dimension size of $d = 64$. We output depth values from the final convolution layer, which has only one filter. For window partitioning in the CATL, we use $M = 12$, and each MSA module has six attention heads.

We used the PyTorch framework [61] to train a dedicated network for each upsampling factor $s \in 2, 4, 8, 16$. Each network was trained for 3×10^5 iterations and optimized using the $\mathcal{L}_1$ loss and the ADAM optimizer [62] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. We used a learning rate of 10^{-4}, dividing the learning rate by 2 for every 1×10^5 iteration. All the models were trained on a PC with an i9-10940x CPU, 128GB RAM, and two Quadro RTX6000 GPUs.

4.3. Results

This section provides quantitative and qualitative evaluations of the proposed architecture for guided DSR. Our proposed architecture was evaluated on both the noise-free and the noisy Middlebury datasets. Further, we conduct experiments on the NYU Depth v2 dataset in order to demonstrate the generalization capabilities of the proposed architecture. We compare the results to other state-of-the-art methods, including global optimization-based methods [10,32], a sparse representation-based method [14], and mainly state-of-the-art deep learning-based methods [3,4,7,17,19–23,25,37,39,43,44]. We also report the results of naive bicubic interpolation as a baseline.

4.3.1. Noise-Free Middlebury Dataset

The Middlebury dataset provides high-quality depth and color image pairs from several challenging scenes. First, we evaluate the different methods for the noise-free Middlebury RGB-D datasets for different scaling factors. In Table 1, we report the obtained RMSE values. Boldface indicates the best RMSE for each evaluation, while the underline indicates the second best. In Table 1, all results are calculated from upsampled depth maps provided by the authors or generated by their code.

Clearly, from Table 1 we conclude that deep learning-based methods [3,4,7,17,19–23,25,37] outperform the more classic methods for DSR. In terms of RMSE values, the proposed architecture provides the best performance across almost all scaling factors. For large scaling factors, e.g., 8, 16, which are difficult for most methods, our method provides good reconstruction with the lowest RMSE error across all datasets. For scaling factors x4/x8/x16, our method obtained 0.48/0.99/1.55 as the average RMSE for the entire test set, respectively. Our results outperform the second-best results in terms of average RMSE values by 0.01/0.09/0.16, respectively.

In Figures 3 and 4, we provide upsampled depth maps on the "Art" and "Moebius" datasets and a scale factor of 8 for qualitative evaluation. Upsampled depth maps are generated from 5 state-of-the-art methods, which are MSG [4], DSR [3], RDGE [32], RDN [7] and CTG [23]. We also provide bicubic interpolation as a baseline for comparison. Compared with competing methods, the proposed architecture provides more detailed HR depth boundaries. Additionally, our approach mitigates the texture-copying effect evident in some other methods, as shown by the red arrow. A significant factor contributing to these results is the attention mechanism built into the transformer model. This attention mechanism transfers HR information from the guidance image to the upsampling process in a sophisticated manner. Moreover, the transformer's ability to consider both local and global information is key to improved performance at large scaling factors. Finally, these evaluations indicate that our CAGM contributes significantly to the success of depth map SR and enables accurate reconstruction even in complex scenarios with various degradations.

Table 1. An analysis of RMSE Values for different scaling factors on the noise-free Middlebury dataset. Boldface indicates the best RMSE for each evaluation, while the underline indicates the second best.

Method	Art			Books			Laundry			Dolls			Moebius			Reindeer		
	x4	x8	x16	x4	x8	x16	x4	x8	x16	x4	x8	x16	x4	x8	x16	x4	x8	x16
Bicubic	3.88	5.60	8.58	1.56	2.24	3.36	2.11	3.10	4.47	1.21	1.78	2.57	1.40	2.05	2.95	2.51	3.92	5.72
TGV [10]	4.06	5.08	7.61	2.21	2.47	3.54	2.20	3.92	6.75	1.42	2.05	4.44	2.03	2.58	3.50	2.67	4.29	8.80
JID [14]	1.92	2.76	5.74	0.71	1.01	1.93	1.10	1.83	3.62	0.92	1.26	1.74	0.89	1.27	2.13	1.41	2.12	4.64
RDGE [32]	3.26	4.31	6.78	1.53	2.18	2.92	2.06	2.87	4.22	1.49	1.94	2.45	1.44	2.21	2.79	2.58	3.24	4.90
MSG [4]	1.49	2.79	5.95	0.66	1.09	1.87	1.02	1.35	2.03	0.72	0.99	1.59	0.68	1.14	2.07	1.33	1.72	2.99
DSR [3]	1.21	2.23	3.95	0.60	0.89	1.51	0.75	1.21	1.89	0.81	1.10	1.60	0.67	0.96	1.57	0.96	1.57	2.54
PSR [25]	1.59	2.57	4.83	0.83	1.19	1.70	0.92	1.52	2.97	0.91	1.31	1.88	0.86	1.21	1.87	1.11	1.80	3.11
MFR [37]	1.54	2.71	4.35	0.63	1.05	1.78	1.11	1.75	3.01	0.89	1.22	1.74	0.72	1.10	1.73	1.23	2.06	3.74
RDN [7]	1.47	2.60	4.16	0.62	1.00	1.68	0.96	1.63	2.86	0.88	1.21	1.71	0.69	1.06	1.65	1.17	1.60	3.58
PMBA [21]	1.19	2.47	4.37	0.53	1.10	1.51	0.80	1.54	2.72	0.66	1.08	1.75	0.55	1.13	1.62	0.92	1.76	2.86
RYN [20]	0.98	2.04	3.37	0.36	0.73	1.37	0.64	1.21	2.01	0.59	0.97	**1.37**	0.50	0.81	1.37	0.74	1.41	2.22
CUN [22]	1.05	2.27	3.67	0.35	0.73	1.45	0.59	1.15	2.25	0.61	0.97	1.43	0.48	0.77	1.31	0.82	1.51	2.38
GDC [19]	1.09	2.04	3.58	0.38	0.68	1.41	0.64	1.13	2.13	0.63	0.97	1.44	0.49	0.79	1.37	0.84	1.51	2.43
TDTN [39]	1.24	2.45	–	0.48	0.86	–	0.68	1.29	–	0.76	1.15	–	0.61	0.91	–	0.95	1.75	–
MIG [43]	1.46	2.74	4.26	0.58	0.95	1.67	0.93	1.57	2.85	0.87	1.21	1.75	0.66	1.04	1.66	1.17	2.11	3.81
PDR [44]	1.59	2.57	4.83	0.83	1.19	1.70	0.92	1.52	2.97	0.91	1.31	1.88	0.86	1.21	1.87	1.11	1.80	3.11
CTG [23]	0.73	1.89	2.76	0.35	**0.66**	1.22	**0.43**	0.87	1.62	0.50	0.90	1.49	**0.46**	0.76	1.31	**0.43**	1.19	1.84
FCTN (Proposed)	**0.71**	**1.71**	**2.56**	**0.34**	0.68	**1.12**	0.47	**0.79**	**1.43**	**0.45**	**0.81**	1.40	**0.46**	**0.68**	**1.18**	0.47	**1.12**	**1.64**

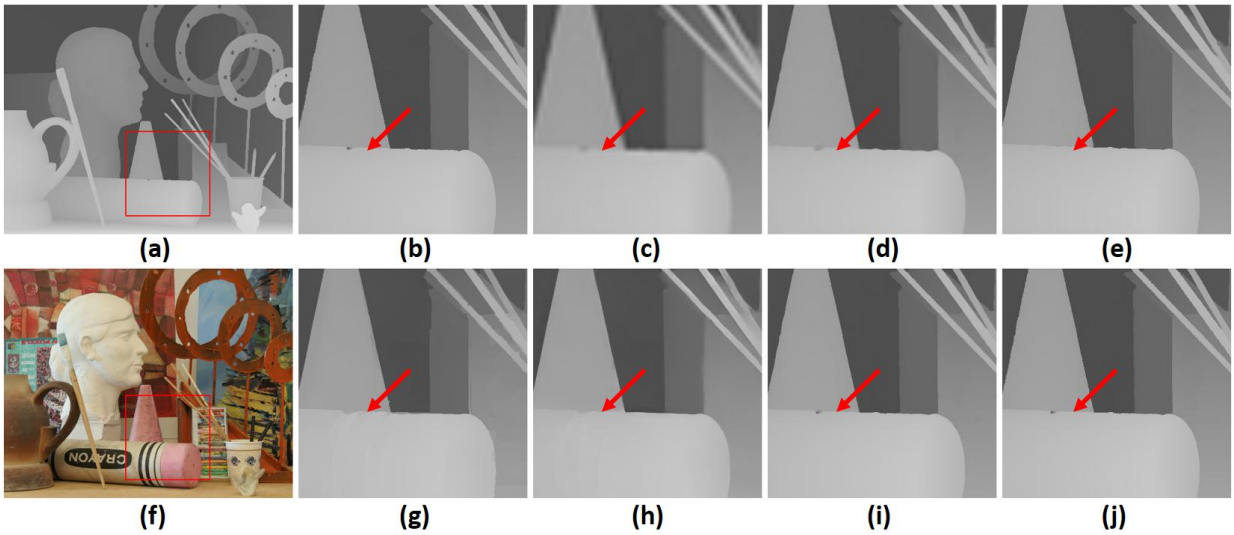

Figure 3. A visual quality comparison for depth map SR at a scale factor of 8 on the noise-free "art" dataset. (**a**) HR depth image, (**f**) HR color image, (**b**) extracted ground truth patch (marked by a red square), and upsampled patches by (**c**) Bicubic, (**d**) MSG [4], (**e**) DSR [3], (**g**) RDGE [32], (**h**) RDN [7], (**i**) CTG [23], (**j**) the proposed FCTN method (best viewed on the enlarged electronic version).

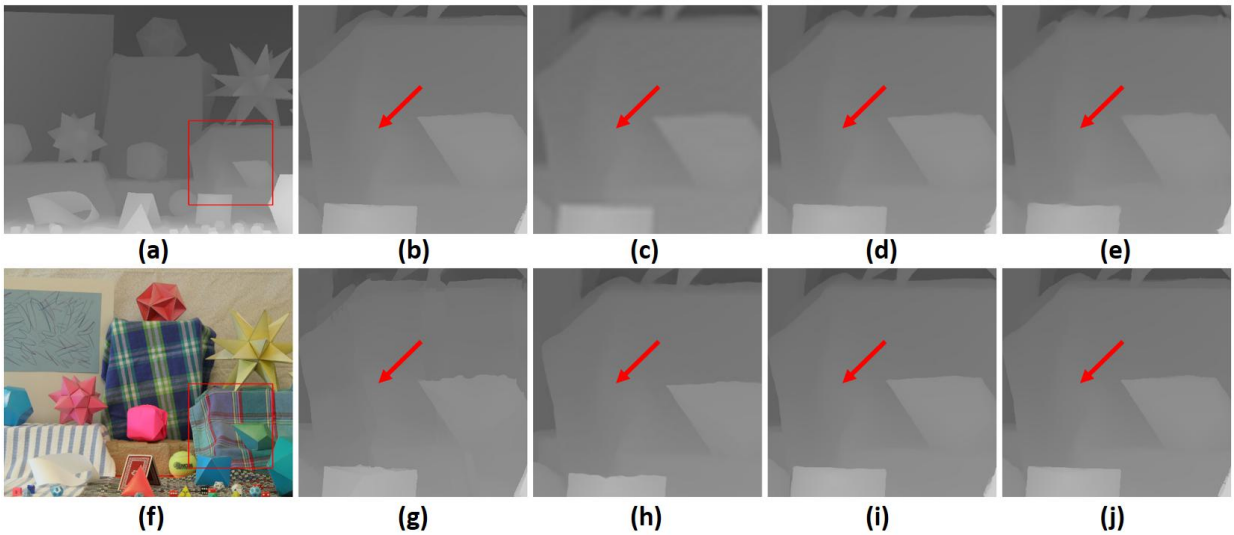

Figure 4. A visual quality comparison for depth map SR at a scale factor of 8 on the noise-free "Moebius" dataset. (**a**) HR depth image, (**f**) HR color image, (**b**) extracted ground truth patch (marked by a red square, and upsampled patches by (**c**) Bicubic, (**d**) MSG [4], (**e**) DSR [3], (**g**) RDGE [32], (**h**) RDN [7], (**i**) CTG [23], (**j**) the proposed FCTN method (best viewed on the enlarged electronic version).

4.3.2. Noisy Middlebury Dataset

We further demonstrate the robustness of the proposed architecture on the noisy Middlebury dataset. We added Gaussian noise to the LR training data, simulating the case where depth maps are corrupted during acquisition, in the same way as [3,7,23,37]. All the models were retrained and evaluated on a test set corrupted with the same Gaussian noise. For the noisy dataset, we report the RMSE values in Table 2.

Table 2. An analysis of RMSE values for different scaling factors on the noisy Middlebury dataset. Boldface indicates the best RMSE for each evaluation, while the underline indicates the second best.

Method	Art		Books		Laundry		Dolls		Moebius		Reindeer	
	x8	x16	x8	x16	x8	x16	x8	x16	x8	x16	x8	x16
Bicubic	6.74	9.04	4.68	5.30	5.35	6.53	4.51	4.90	4.54	5.02	5.71	7.12
TGV [10]	7.26	12.05	2.88	4.73	4.45	8.06	2.82	5.14	3.01	6.11	4.65	9.03
MSG [4]	4.24	7.42	2.48	4.19	3.31	4.88	2.53	3.41	2.47	3.76	3.36	4.95
MFR [37]	3.97	6.14	2.13	3.17	2.82	4.57	2.25	3.30	2.13	3.33	3.01	4.86
RDN [7]	4.09	6.62	2.11	3.36	2.88	5.11	2.33	3.59	2.18	3.69	3.09	4.93
DSR [3]	–	6.96	–	5.66	–	7.54	–	4.28	–	3.39	–	5.25
RYN [20]	3.47	–	1.88	–	2.47	–	1.97	–	1.87	–	2.68	–
GDC [19]	3.31	4.77	1.69	**2.46**	2.20	3.36	1.89	2.59	1.72	2.68	2.57	3.44
JIIF [40]	3.87	7.14	1.75	2.47	–	–	–	–	2.03	3.18	–	–
MIG [43]	3.95	6.15	2.10	3.17	3.00	4.88	2.21	3.51	2.12	3.51	3.04	4.97
CTG [23]	3.26	4.72	1.61	2.96	1.63	3.47	1.64	**2.16**	1.63	2.24	**1.79**	3.59
FCTN (Proposed)	**3.01**	**4.55**	**1.54**	2.66	**1.61**	3.15	**1.59**	2.32	**1.27**	**2.09**	1.81	**3.17**

Our first observation is that noise added to the LR depth maps significantly affects the reconstructed HR depth maps regardless of the method or scaling factor used. However,

the proposed architecture still generates clean and sharp reconstructions and outperforms competing methods in terms of RMSE.

An even more realistic scenario is that data acquired by both the depth and color sensors are corrupted by noise. Our method was further tested by adding Gaussian noise with a mean of 0 and a standard deviation of 5 to the HR guidance images. This was done both in training and in testing. We again retrained the models and report the obtained average RMSE values in Table 3. In Table 3, we observe that the added noise in the HR guidance image did not significantly affect the performance of our method, compared to only adding noise to LR depth. According to our results, the proposed CAGM is somewhat insensitive to noise added to the guidance image.

Table 3. An Analysis of the average RMSE values for different noise schemes.

Middlebury Dataset Version	x4	x8	x16
Noise-Free	0.48	0.99	1.55
Depth Noise	1.17	1.80	2.99
Depth and Color Noise	1.35	2.01	3.19

4.3.3. NYU Depth v2 Dataset

In this section, the proposed architecture is tested on the challenging public NYU Depth v2 [63] dataset as a means of demonstrating its generalization ability. There are 1449 high-quality RGB-D images of natural indoor scenes in this dataset, with apparent misalignments between depth maps and color images. We note that data from NYU Depth v2 are very different from the Middlebury Dataset and were not included in the training data of our models.

We report the average RMSE value across the entire dataset in Table 4. Boldface indicates the best RMSE value. As a baseline, we report the results of Bicubic interpolation as well as the results of competing guided SR approaches; ATGV-Net [5], MSG [4], DSR [3], RDN [7], RYN [20], PMBA [21], DEAF [42], JIIF [40], DCT [17], and CTG [23]. The proposed architecture achieves the lowest average RMSE, demonstrating the proposed method's generalization ability and robustness.

Table 4. Quantitative comparisons of the ablation experiments. Reported results are average RMSE on the noise-free Middlebury dataset for scaling factors 4, 8, and 16. Boldface indicates the best RMSE for each evaluation, while the underline indicates the second best.

Method	Average RMSE on NYU Depth v2 Dataset
Bicubic	2.36
ATGV-Net [5]	1.28
MSG [4]	1.31
RDN [7]	1.21
DSR [3]	1.34
RYN [20]	1.06
PMBA [21]	1.06
DEAF [42]	1.12
JIIF [40]	1.37
DCT [17]	1.59
CTG [23]	<u>0.95</u>
FCTN (Proposed)	**0.91**

4.3.4. Inference Time

For a DSR method to applyto real-world applications, it is often required to work in a close-to-real-time performance. Thus, we report the inference time of the proposed architecture compared to other competing approaches. Inference times were measured

using an image of size 1320×1080 pixels and the setup described in Section 4.2. We report our results in milliseconds in Table 5

Table 5 shows that compared to traditional methods, the proposed architecture, as well as other deep learning-based methods, provide significantly faster inference times. Moreover, the proposed method is comparable to competing methods and achieves lower RMSE values. In contrast, References [10,12,32] require multiple optimization iterations to obtain accurate reconstructions, leading to slower inference times. Some methods, such as [3,32], upsample the LR depth as an initial preprocessing step before the image is fed to the model. As a result, they show very similar inference times regardless of the scaling factor.

Table 5. Average inference times (milliseconds) for different scaling factors.

Method	x2	x4	x8	x16
Bicubic	10	10	10	10
TGV [10]	45,730	49,780	46,340	46,200
AR [12]	158,010	157,730	157,950	158,770
RDGE [32]	68,070	67,690	68,450	68,170
MSG [4]	260	300	380	420
DSR [3]	220	230	230	230
RYN [20]	460	630	720	880
CTG [23]	150	380	480	530
FCTN (Proposed) [23]	140	304	420	490

4.4. Ablation Study

In the ablation study, we test the effects of the CATB number in the CAGM and CATL number in each CATB on model performance. Results are shown in Figure 5a,b, respectively. It is observed that the RMSE of the reconstructed depth is positively correlated with both hyperparameters until it becomes eventually saturated. As we increase either hyperparameter, model size becomes increasingly prominent, and training\inference time and memory requirements are negatively impacted. Thus, to balance the performance and model size, we choose 6 for both hyperparameters as described in Section 4.2. CATL numbers were evaluated with a configuration of $K = 6$ CATBs.

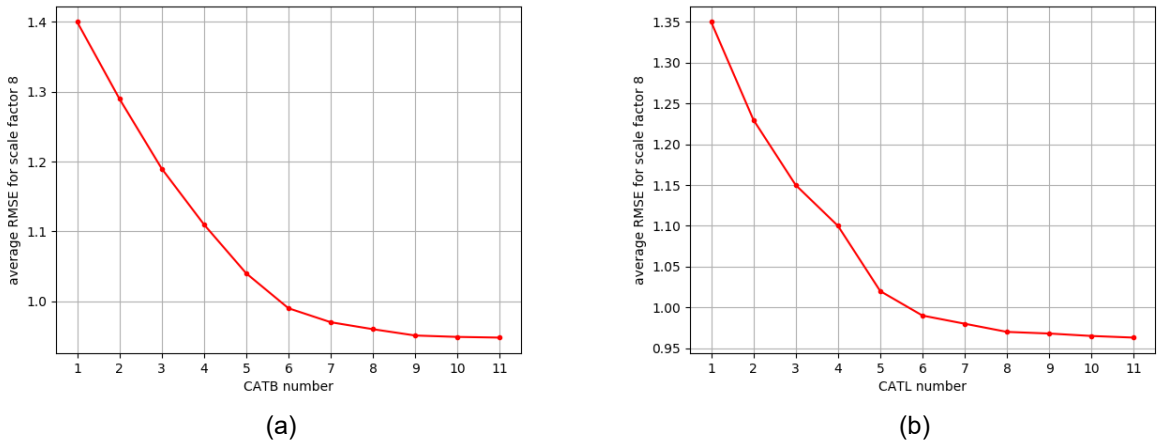

Figure 5. Ablation study on different configurations of the proposed CAGM. Results are the average RMSE on the noise-free Middlebury dataset for scaling factor 8. (**a**) The effect of the CATB number in the CAGM, and (**b**) the effect of CATL number in each CATB.

The impact of each component in our design is evaluated via the following experiments: (1) Our architecture without any guidance from the color image, denoted as "Depth-Only". (2) Our architecture without shifted windows in the CATL, denoted "w/o

shift". (3) Our architecture without the CFG module, denoted "w/o CFG". (4) Our architecture without the use of cross-attention for guidance. In this setting, we replaced the CATL with a similar design using only self-attention with depth features as input. Features from the color image were concatenated after every modified CATL to provide guidance. We denote this setting as "w/o cross-attention".

We evaluate the different designs on the Middlebury test set at scaling factors 4, 8, and 16. We use the same CATB and CATL configuration as described in Section 4.2 in these experiments. We summarize the results in Table 6 and observe that: (1) As expected, using only the LR depth for DSR without guidance from a color image provides inferior results. (2) As also observed in [47], incorporating shifted window partitioning into our CATL improves the performance. Using shifted windows partitioning enables connections among windows in the preceding layers, improving the representation capability of each CATL. (3) Our CFG module provides additional high-frequency information directly to the upsampling module. As a result, the upsampling module can reconstruct a higher quality HR depth, and we observe that performance improves slightly. (4) We observe that using a simple concatenation of features instead of the proposed cross-attention guidance leads to inferior results. Incorporating the guidance from the color image via cross-attention allows the color feature to interact elaborately with the depth features and to encode long-distant dependencies between the two modalities.

Table 6. An analysis of the average RMSE values for different ablation experiments on the noise-free Middlebury dataset. Boldface indicates the best RMSE for each evaluation.

Design	**Depth-Only**			**w/o Shift**			**w/o CFG**			**w/o Cross-Attention**			**FCTN (Proposed)**		
Scale Factor	x4	x8	x16	x4	x8	x16	x4	x8	x16	x4	x8	x16	x4	x8	x16
RMSE	0.65	1.39	3.01	0.52	1.14	1.90	0.51	1.06	1.79	0.59	1.28	2.17	**0.48**	**0.99**	**1.55**

5. Conclusions

We introduce a novel transformer-based architecture with cross-attention for guided DSR. First, a shallow feature extraction module extracts meaningful features from LR depth and HR color images. These features are fed to a cascaded transformer module with cross-attention, which extracts more elaborate features while simultaneously incorporates guidance from the color features via the cross-attention mechanism. The cascaded transformer module is constructed by stacking transformer layers with shifted window partitioning, which enables interactions between windows in consecutive layers. Using such a design, the proposed architecture achieves state-of-the-art results on the DSR benchmarks. At the same time, model size and inference time remain comparably small, making our architecture usable for real-world applications.

Our future work will explore more realistic depth artifacts (e.g., sparse depth values, misalignment between guidance and depth images, etc.). Moreover, we will examine the proposed architecture on additional real-world continuous data acquired from sensors mounted, e.g., on an autonomous robot.

Author Contributions: Conceptualization, I.A.; Methodology, I.A. and I.C.; Writing—original draft, I.A.; Writing—review & editing, I.C.; Supervision, I.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the PMRI—Peter Munk Research Institute-Technion.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: MPI Sintel Dataset—[http://sintel.is.tue.mpg.de/, accessed on 15 February 2023]. Middlebury Stereo Datasets—[https://vision.middlebury.edu/stereo/data/, accessed on 15 February 2023].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Izadi, S.; Kim, D.; Hilliges, O.; Molyneaux, D.; Newcombe, R.; Kohli, P.; Shotton, J.; Hodges, S.; Freeman, D.; Davison, A.; et al. KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera. In Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, 16–19 October 2011; pp. 559–568.
2. Schamm, T.; Strand, M.; Gumpp, T.; Kohlhaas, R.; Zollner, J.M.; Dillmann, R. Vision and ToF-based driving assistance for a personal transporter. In Proceedings of the 2009 International Conference on Advanced Robotics, Munich, Germany, 22–26 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 1–6.
3. Guo, C.; Li, C.; Guo, J.; Cong, R.; Fu, H.; Han, P. Hierarchical features driven residual learning for depth map super-resolution. *IEEE Trans. Image Process.* **2018**, *28*, 2545–2557. [CrossRef]
4. Hui, T.W.; Loy, C.C.; Tang, X. Depth map super-resolution by deep multi-scale guidance. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 353–369.
5. Riegler, G.; Rüther, M.; Bischof, H. Atgv-net: Accurate depth super-resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 268–284.
6. Song, X.; Dai, Y.; Qin, X. Deeply supervised depth map super-resolution as novel view synthesis. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 2323–2336. [CrossRef]
7. Zuo, Y.; Fang, Y.; Yang, Y.; Shang, X.; Wang, B. Residual dense network for intensity-guided depth map enhancement. *Inf. Sci.* **2019**, *495*, 52–64. [CrossRef]
8. He, K.; Sun, J.; Tang, X. Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1397–1409. [CrossRef] [PubMed]
9. Yang, Q.; Yang, R.; Davis, J.; Nistér, D. Spatial-depth super resolution for range images. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, 17–22 June 2007; IEEE: Piscataway, NJ, USA, 2007; pp. 1–8.
10. Ferstl, D.; Reinbacher, C.; Ranftl, R.; Rüther, M.; Bischof, H. Image guided depth upsampling using anisotropic total generalized variation. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 993–1000.
11. Jiang, Z.; Hou, Y.; Yue, H.; Yang, J.; Hou, C. Depth super-resolution from RGB-D pairs with transform and spatial domain regularization. *IEEE Trans. Image Process.* **2018**, *27*, 2587–2602. [CrossRef]
12. Yang, J.; Ye, X.; Li, K.; Hou, C.; Wang, Y. Color-guided depth recovery from RGB-D data using an adaptive autoregressive model. *IEEE Trans. Image Process.* **2014**, *23*, 3443–3458. [CrossRef]
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
14. Kiechle, M.; Hawe, S.; Kleinsteuber, M. A joint intensity and depth co-sparse analysis model for depth map super-resolution. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1545–1552.
15. Kwon, H.; Tai, Y.W.; Lin, S. Data-driven depth map refinement via multi-scale sparse representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 159–167.
16. Park, J.; Kim, H.; Tai, Y.W.; Brown, M.S.; Kweon, I. High quality depth map upsampling for 3d-tof cameras. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 1623–1630.
17. Zhao, Z.; Zhang, J.; Xu, S.; Lin, Z.; Pfister, H. Discrete cosine transform network for guided depth map super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5697–5707.
18. Lutio, R.d.; D'aronco, S.; Wegner, J.D.; Schindler, K. Guided super-resolution as pixel-to-pixel transformation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8829–8837.
19. Kim, J.Y.; Ji, S.; Baek, S.J.; Jung, S.W.; Ko, S.J. Depth Map Super-Resolution Using Guided Deformable Convolution. *IEEE Access* **2021**, *9*, 66626–66635. [CrossRef]
20. Li, T.; Dong, X.; Lin, H. Guided depth map super-resolution using recumbent y network. *IEEE Access* **2020**, *8*, 122695–122708. [CrossRef]
21. Ye, X.; Sun, B.; Wang, Z.; Yang, J.; Xu, R.; Li, H.; Li, B. Pmbanet: Progressive multi-branch aggregation network for scene depth super-resolution. *IEEE Trans. Image Process.* **2020**, *29*, 7427–7442. [CrossRef]
22. Cui, Y.; Liao, Q.; Yang, W.; Xue, J.H. RGB Guided Depth Map Super-Resolution with Coupled U-Net. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
23. Ariav, I.; Cohen, I. Depth Map Super-Resolution via Cascaded Transformers Guidance. *Front. Signal Process.* **2022**, *3*.. [CrossRef]
24. Zhang, K.; Zuo, W.; Gu, S.; Zhang, L. Learning deep CNN denoiser prior for image restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3929–3938.
25. Huang, L.; Zhang, J.; Zuo, Y.; Wu, Q. Pyramid-Structured Depth Map Super-Resolution Based on Deep Dense-Residual Network. *IEEE Signal Process. Lett.* **2019**, *26*, 1723–1727. [CrossRef]

26. He, K.; Sun, J.; Tang, X. Guided image filtering. In Proceedings of the European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 1–14.
27. Liu, M.Y.; Tuzel, O.; Taguchi, Y. Joint geodesic upsampling of depth images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 169–176.
28. Lu, J.; Forsyth, D. Sparse depth super resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2245–2253.
29. Dong, W.; Shi, G.; Li, X.; Peng, K.; Wu, J.; Guo, Z. Color-guided depth recovery via joint local structural and nonlocal low-rank regularization. *IEEE Trans. Multimed.* **2016**, *19*, 293–301. [CrossRef]
30. Ham, B.; Cho, M.; Ponce, J. Robust image filtering using joint static and dynamic guidance. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4823–4831.
31. Ham, B.; Min, D.; Sohn, K. Depth superresolution by transduction. *IEEE Trans. Image Process.* **2015**, *24*, 1524–1535. [CrossRef] [PubMed]
32. Liu, W.; Chen, X.; Yang, J.; Wu, Q. Robust color guided depth map restoration. *IEEE Trans. Image Process.* **2016**, *26*, 315–327. [CrossRef] [PubMed]
33. Park, J.; Kim, H.; Tai, Y.W.; Brown, M.S.; Kweon, I.S. High-quality depth map upsampling and completion for RGB-D cameras. *IEEE Trans. Image Process.* **2014**, *23*, 5559–5572. [CrossRef]
34. Yang, J.; Ye, X.; Li, K.; Hou, C. Depth recovery using an adaptive color-guided auto-regressive model. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 158–171.
35. Riegler, G.; Ferstl, D.; Rüther, M.; Bischof, H. A deep primal-dual network for guided depth super-resolution. *arXiv* **2016**, arXiv:1607.08569.
36. Zhou, W.; Li, X.; Reynolds, D. Guided deep network for depth map super-resolution: How much can color help? In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1457–1461.
37. Zuo, Y.; Wu, Q.; Fang, Y.; An, P.; Huang, L.; Chen, Z. Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 297–306. [CrossRef]
38. de Lutio, R.; Becker, A.; D'Aronco, S.; Russo, S.; Wegner, J.D.; Schindler, K. Learning Graph Regularisation for Guided Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1979–1988.
39. Yao, C.; Zhang, S.; Yang, M.; Liu, M.; Qi, J. Depth super-resolution by texture-depth transformer. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
40. Tang, J.; Chen, X.; Zeng, G. Joint implicit image function for guided depth super-resolution. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, China, 20–24 October 2021; pp. 4390–4399.
41. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
42. Liu, P.; Zhang, Z.; Meng, Z.; Gao, N. Deformable Enhancement and Adaptive Fusion for Depth Map Super-Resolution. *IEEE Signal Process. Lett.* **2021**, *29*, 204–208. [CrossRef]
43. Zuo, Y.; Wang, H.; Fang, Y.; Huang, X.; Shang, X.; Wu, Q. MIG-net: Multi-scale Network Alternatively Guided by Intensity and Gradient Features for Depth Map Super-resolution. *IEEE Trans. Multimed.* **2021**, *24*, 3506–3519. [CrossRef]
44. Liu, P.; Zhang, Z.; Meng, Z.; Gao, N.; Wang, C. PDR-Net: Progressive depth reconstruction network for color guided depth map super-resolution. *Neurocomputing* **2022**, *479*, 75–88. [CrossRef]
45. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
46. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv* **2021**, arXiv:2102.12122.
47. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
48. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Nashville, TN, USA, 19–25 June 2021; pp. 6881–6890.
49. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
50. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
51. Haris, M.; Shakhnarovich, G.; Ukita, N. Deep back-projection networks for super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1664–1673.

52. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2472–2481.
53. Xiao, T.; Singh, M.; Mintun, E.; Darrell, T.; Dollár, P.; Girshick, R. Early convolutions help transformers see better. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 30392–30400.
54. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
55. Hu, H.; Zhang, Z.; Xie, Z.; Lin, S. Local relation networks for image recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3464–3473.
56. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
57. Butler, D.J.; Wulff, J.; Stanley, G.B.; Black, M.J. A naturalistic open source movie for optical flow evaluation. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 611–625.
58. Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; Westling, P. High-resolution stereo datasets with subpixel-accurate ground truth. In Proceedings of the German Conference on Pattern Recognition, Munster, Germany, 2–5 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 31–42.
59. Scharstein, D.; Pal, C. Learning conditional random fields for stereo. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, 17–22 June 2007; IEEE: Piscataway, NJ, USA, 2007; pp. 1–8.
60. Scharstein, D.; Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42. [CrossRef]
61. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 8024–8035.
62. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
63. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 746–760.

Chapter 7

Discussion and Conclusions

7.1 Discussion and Conclusions

This research thesis introduces different deep neural network architectures for the multimodal tasks of audio-visual voice activity detection and guided super resolution of depth maps. These approaches are advantageous because they combine the different modalities and extract complementary information from each, resulting in improved performance in each task compared to existing methods.

7.1.1 audio-visual VAD

Traditionally, in silent acoustic environments, speech can be successfully distinguished from silent regions using methods based on simple acoustic features. These methods, however, perform significantly worse in the presence of noise, even at moderate SNR ratios. Other methods assume statistical models for the noisy signal, focusing on the estimation of the parameters of the model. The main drawback of such methods is that they cannot properly model highly non-stationary noise and transient interferences. Transients, like speech, often show rapid variations in their spectrum over time, which makes them difficult to distinguish.

To try and overcome these limitations, more recent studies address the problem of voice activity detection from a machine learning point of view, in which the goal is to classify segments of the noisy signal into speech and non-speech classes. Instead of assuming explicit distributions for the noisy signal, learning-based methods learn

implicit models from training data. Specifically, deep neural networks have gained popularity in recent years in a variety of machine learning tasks. For useful signal representations, these models use deep neural networks with multiple layers, and their potential for voice activity detection has been partially explored. Despite this, these methods still struggle with frames that contain both speech and transients. Transients, which are characterized by fast variations in time and high energy values, often appear more dominant than speech. Consequently, frames containing only transients appear similar to frames containing both transients and speech and are mistakenly identified as speech frames.

Another school of studies suggests improving speech detection's robustness to noise and transients by incorporating a video signal, which is independent of the acoustic environment. Often, the video captures the mouth region of the speaker, which is represented by specifically designed features that model the shape and movement of the mouth. Several approaches exist in the literature for fusing audio and video signals, called early fusion and late fusion. In early fusion, video, and audio features are concatenated into a single feature vector and processed as single-modal data. In late fusion, measures of speech presence and absence are constructed separately from each modality and then combined using statistical models.

To address the aforementioned problems, this thesis proposes two multimodal audio-visual VAD approaches:

1) We propose a deep architecture for speech detection, based on specifically designed auto-encoders providing a new representation of the audio-visual signal, in which the effect of transients is reduced. The new representation is fed into a deep RNN, trained in a supervised manner to generate voice activity estimation while exploiting the differences in the dynamics between speech and the transients. Experimental results have demonstrated that the proposed architecture outperforms competing state-of-the-art detectors providing accurate detections even in low SNR and in the presence of challenging types of transients.

2) The study described in the former part is extended and we propose a deep multimodal end-to-end architecture for speech detection, which utilizes residual connections and dilated convolutions and operates on raw audio and video signals to extract mean-

ingful features in which the effect of transients is reduced. The features from each modality are fused into a joint representation using MCB which allows higher-order relations between the two modalities to be explored. In order to further exploit the differences in the dynamics between speech and the transients, the joint representation is fed into a deep LSTM. A fully connected layer is added, and the entire network is trained in a supervised manner to perform voice activity detection. Experimental results have demonstrated that our multimodal end-to-end architecture outperforms unimodal variants while providing accurate detections even under low SNR conditions and in the presence of challenging types of transients. Furthermore, the use of MCB for modality fusion has also been shown to outperform other methods for modality fusion.

7.1.2 guided super resolution of depth maps

The upsampling of depth information is not a trivial task. A naive upsampling of the LR image, e.g., bicubic interpolation, usually produces unsatisfactory results with blurred and unsharp edges.

Traditional methods for SR of depth maps can be categorized as filter-based, energy-minimization-based„ and learning-based. Filter-based methods have relatively low computational complexity but often result in apparent artifacts in HR depth maps. In contrast, energy minimization methods require cumbersome and time-consuming computations. Many of them rely heavily on regularization from a statistic or prior, which may not be available in all scenarios. Finally, learning-based methods, specifically ones incorporating deep neural networks, have proliferated in recent years, providing the best results regarding upsampled depth map quality.

In recent methods, the SR of the depth map is guided or enhanced by a corresponding HR intensity image, if available. These methods assume that correspondence can be established between an edge in the intensity image and the matching edge in the depth map. Since the intensity image has a higher resolution, its edges can determine depth discontinuities in the super-resolved HR depth map. Nevertheless, edges in the intensity image may not correspond to depth discontinuities, e.g., smooth surfaces with highly textured textures in the intensity image. In these cases, color textures are over-transferred from textured regions to the super-resolved depth map, resulting in texture

copying. Therefore, a more sophisticated guidance scheme is needed.

As with the unimodal SR methods, these multimodal methods can also be classified into four subcategories: filtering-based methods, global optimization-based methods, sparse representation-based methods, and deep learning-based methods. The use of deep learning methods for guided SR of depth images has gained increasing attention and has demonstrated promising results. However, the guidance features and depth features are mostly combined using mere concatenation, even in recent methods. Furthermore, most of these methods rely on CNNs for feature extraction, which are limited in their receptive fields and lack the expressive power of transformers.

To address these problems, this thesis proposes two guided super resolution approaches:

1) We present a generalized framework to address the problem of depth map upsampling by using a cascaded transformer module for guided depth SR. An LR depth map is progressively upsampled using residual dilated blocks and a novel guidance module, based on the cascaded transformer that operates on shifted window partitioning of the image, scales the intermediate feature maps of the network. The proposed architecture achieves state-of-the-art performance for super-resolving depth maps using such a design.

2) We extend the work described above and introduce a novel transformer-based architecture with cross-attention for guided depth SR. First, a shallow feature extraction module extracts meaningful features from LR depth and HR color images. These features are fed to a cascaded transformer module with cross-attention, which extracts more elaborate features while simultaneously incorporating guidance from the color features via the cross-attention mechanism. The cascaded transformer module is constructed by stacking transformer layers with shifted window partitioning, which enables interactions between windows in consecutive layers. Using such a design, the proposed architecture achieves state-of-the-art results on depth SR benchmarks. At the same time, model size and inference time remain comparably small, making our architecture usable for real-world applications.

7.2 Future Research Directions

The subject of multimodal neural networks was explored in this thesis. A number of such architectures have been developed and demonstrated for use cases involving audio-visual VAD and guided super resolution of depth images. As a result of these developments, further research can be conducted in the following areas:

1) **Extension to other temporal multimodal tasks**. In this thesis, we discuss the representation and fusion of audio and video signals for VAD in a noisy environment. Research directions in the future may include applying these multimodal architectures to speech recognition or speech enhancement tasks. The proposed architecture could also be used with altogether different temporal modalities, e.g., replace the audio signal with an electrocardiogram (ECG) signal and train the network to analyze ECG data. It is possible because the architecture is based on raw signals and is not reliant on any audio or image-specific features.

2) **More realistic noise and artifacts**. In this thesis, we considered an audio-visual setting in which the audio signal was contaminated by background noises and transient interferences. In future research directions, additional methods for noise injection, in which the video signal is augmented and the speakers' voices are modified according to the injected noise levels (Lombard effect), can be explored. In the setting of guided DSR, we intend to explore even more realistic noise and artifacts in our test sets (e.g., missing depth values, misalignment between guidance and depth images, etc.).

3) **Super-Resolution of Dynamic Elevation Model (DEM)**. In this thesis, we presented several deep architectures for guided super resolution of depth information, with promising results. Our future research will investigate the applicability of our work to the challenging case of aerial imagery SR in which both a Raster (color) image and a Dynamic Elevation Model (DEM) are available. DEMs for most of Earth's surface are low-resolution and cannot accurately reflect the terrain's morphology. Our objective would be to improve the DEM resolution using both the LR DEM and raster image as inputs. Many applications require HR DEM, such as off-road path planning, line-of-sight analysis, digital twins, etc. The more detailed the DEM, the more accurate these

analyses become.

This task differs from depth SR as investigated in this thesis. The depth data tends to be smoother than DEM data, with apparent discontinuities in depth. Hence, the SR of such data and, more specifically, the guidance of such an up-sampling process is more challenging.

Bibliography

[AAL+15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proc. International Conference on Computer Vision (ICCV)*, 2015.

[AHC10] AJ Aubrey, YA Hicks, and JA Chambers. Visual voice activity detection with optical flow. *IET Image Processing*, 4(6):463–472, 2010.

[AM08] Ibrahim Almajai and Ben Milner. Using audio-visual features for robust voice activity detection in clean and noisy speech. In *Proc. 16th European Signal Processing Conference (EUSIPCO)*, 2008.

[ARH+07] Andrew Aubrey, Bertrand Rivet, Yulia Hicks, Laurent Girin, Jonathon Chambers, and Christian Jutten. Two novel visual voice activity detectors based on appearance models and retinal filtering. In *Proc. 15th European Signal Processing Conference (EUSIPCO)*, pages 2409–2413, 2007.

[BA18] Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018.

[BL+07] Yoshua Bengio, Yann LeCun, et al. Scaling learning algorithms towards AI. *Large-scale kernel machines*, 34(5):1–41, 2007.

[Boa20] George Boateng. Towards real-time multimodal emotion recognition among couples. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 748–753, 2020.

[CCFC02] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *Proc. International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

[CK11] Namgook Cho and Eun-Kyoung Kim. Enhanced voice activity detection using acoustic event detection and classification. *IEEE Transactions on Consumer Electronics*, 57(1):196–202, 2011.

[CKM06] Joon-Hyuk Chang, Nam Soo Kim, and Sanjit K. Mitra. Voice activity detection based on multiple statistical models. *IEEE Transactions on Signal Processing*, 54(6):1965–1976, 2006.

[CL06] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[CLYX21] Yingjie Cui, Qingmin Liao, Wenming Yang, and Jing-Hao Xue. Rgb guided depth map super-resolution with coupled u-net. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[CMS+20] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[DBK+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Dim22] Giovanna Maria Dimitri. A short survey on deep learning for multimodal integration: Applications, future perspectives and challenges. *Computers*, 11(11):163, 2022.

[DTC15] David Dov, Ronen Talmon, and Israel Cohen. Audio-visual voice activity detection using diffusion maps. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4):732–745, 2015.

[DTC16] David Dov, Ronen Talmon, and Israel Cohen. Kernel method for voice activity detection in the presence of transients. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12):2313–2326, 2016.

[FPY+16] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[FRR+13] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 993–1000, 2013.

[GBZD] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 317–326.

[GJ] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *Proc. International Conference on Machine Learning (ICML)*, pages 1764–1772.

[GLG+18] Chunle Guo, Chongyi Li, Jichang Guo, Runmin Cong, Huazhu Fu, and Ping Han. Hierarchical features driven residual learning for depth map super-resolution. *IEEE Transactions on Image Processing*, 28(5):2545–2557, 2018.

[GMH13] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6645–6649, 2013.

[HDY+12] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[HL13] Wei-Tyng Hong and Chien-Cheng Lee. Voice activity detection based on noise-immunity recurrent neural networks. *International Journal of Advancements in Computing Technology (IJACT)*, 5(5):338–345, 2013.

[HLT16] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *European conference on computer vision*, pages 353–369. Springer, 2016.

[HM13] Thad Hughes and Keir Mierle. Recurrent neural networks for voice activity detection. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7378–7382, 2013.

[hoc] Long short-term memory (neural computation 9 (8): 1735 {1780, 1997), author=Hochreiter, JS Sepp and Schmidhuber, J, journal=Fakultät für Informatik, Technische Universität München, Licence details: https://www. bioinf. jku. at/publications/older/2604. pdf, year=1997.

[HS06] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[HST12] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012.

[HWL+18] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):117–128, 2018.

[HZRS] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[IKH+11] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011.

[JHY+18] Zhongyu Jiang, Yonghong Hou, Huanjing Yue, Jingyu Yang, and Chunping Hou. Depth super-resolution from rgb-d pairs with transform and spatial domain regularization. *IEEE Transactions on Image Processing*, 27(5):2587–2602, 2018.

[JMR94] J.-C. Junqua, Brian Mak, and Ben Reaves. A robust algorithm for word boundary detection in the presence of noise. *IEEE Transactions on Speech and Audio Processing*, 2(3):406–412, 1994.

[KHK13] Martin Kiechle, Simon Hawe, and Martin Kleinsteuber. A joint intensity and depth co-sparse analysis model for depth map super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 1545–1552, 2013.

[KJB+21] Joon-Yeon Kim, Seowon Ji, Seung-Jin Baek, Seung-Won Jung, and Sung-Jea Ko. Depth map super-resolution using guided deformable convolution. *IEEE Access*, 9:66626–66635, 2021.

[KN91] David A. Krubsack and Russel J. Niederjohn. An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech. *IEEE Transactions on Signal Processing*, 39(2):319–329, 1991.

[KTL15] HyeokHyen Kwon, Yu-Wing Tai, and Stephen Lin. Data-driven depth map refinement via multi-scale sparse representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 159–167, 2015.

[LDL20] Tao Li, Xiucheng Dong, and Hongwei Lin. Guided depth map super-resolution using recumbent y network. *IEEE Access*, 8:122695–122708, 2020.

[LDWS19] Riccardo de Lutio, Stefano D'aronco, Jan Dirk Wegner, and Konrad Schindler. Guided super-resolution as pixel-to-pixel transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8829–8837, 2019.

[LHB15] Simon Leglaive, Romain Hennequin, and Roland Badeau. Singing voice detection with deep recurrent neural networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 121–125, 2015.

[LJL16] Wootaek Lim, Daeyoung Jang, and Taejin Lee. Speech emotion recognition using convolutional and recurrent neural networks. In *Proc. Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–4. IEEE, 2016.

[LLC+21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *International Conference on Computer Vision (ICCV)*, 2021.

[LRM] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1457.

[LWJ11] Qingju Liu, Wenwu Wang, and Philip Jackson. A visual voice activity detection method with adaboosting. *Proc. Sensor Signal Processing for Defence (SSPD)*, pages 1–5, 2011.

[MLS+13] Vicente P. Minotto, Carlos BO Lopes, Jacob Scharcanski, Claudio R. Jung, and Bowon Lee. Audiovisual voice activity detection based on microphone arrays and color information. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):147–156, 2013.

[NKK+11] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proc. 28th International Conference on Machine Learning (ICML)*, pages 689–696, 2011.

[PKT+11] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S Brown, and Inso Kweon. High quality depth map upsampling for 3d-tof cameras. In *2011 International Conference on Computer Vision*, pages 1623–1630. IEEE, 2011.

[PP13] Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proc. 19th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 239–247. ACM, 2013.

[PSM+18] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. End-to-end audiovisual speech recognition. *arXiv preprint arXiv:1802.06424*, 2018.

[RHW88] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

[RRB16] Gernot Riegler, Matthias Rüther, and Horst Bischof. Atgv-net: Accurate depth super-resolution. In *European conference on computer vision*, pages 268–284. Springer, 2016.

[RSB+04] Javier Ramírez, José C. Segura, Carmen Benítez, Angel De La Torre, and Antonio Rubio. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, 42(3):271–287, 2004.

[SCK10] Jong Won Shin, Joon-Hyuk Chang, and Nam Soo Kim. Voice activity detection based on statistical models and machine learning approaches. *Computer Speech & Language*, 24(3):515–530, 2010.

[SDQ18] Xibin Song, Yuchao Dai, and Xueying Qin. Deeply supervised depth map super-resolution as novel view synthesis. *IEEE Transactions on circuits and systems for video technology*, 29(8):2323–2336, 2018.

[SKS99] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, 1999.

[SRG+06] David Sodoyer, Bertrand Rivet, Laurent Girin, J.-L. Schwartz, and Christian Jutten. An analysis of visual speech information applied to voice activity detection. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 601–604, 2006.

[SRG+09] David Sodoyer, Bertrand Rivet, Laurent Girin, Christophe Savariaux, Jean-Luc Schwartz, and Christian Jutten. A study of lip movements during spontaneous dialog and its application to voice activity detection. *Journal of the Acoustical Society of America*, 125(2):1184–1196, 2009.

[SSG+09] Thomas Schamm, Marcus Strand, Thomas Gumpp, Ralf Kohlhaas, J Marius Zollner, and Rudiger Dillmann. Vision and tof-based driving assistance for a personal transporter. In *2009 International Conference on Advanced Robotics*, pages 1–6. IEEE, 2009.

[TF00] Joshua B. Tenenbaum and William T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

[TGSS14] Samuel Thomas, Sriram Ganapathy, George Saon, and Hagen Soltau. Analyzing convolutional neural networks for speech activity detection in mis-

matched acoustic conditions. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2519–2523, 2014.

[TJY+12] P. Tiawongsombat, Mun-Ho Jeong, Joo-Seop Yun, Bum-Jae You, and Sang-Rok Oh. Robust visual speakingness detection using bi-level HMM. *Pattern Recognition*, 45(2):783–793, 2012.

[TR07] Rasool Tahmasbi and Sadegh Rezaei. A soft voice activity detection using garch filter and variance gamma distribution. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1129–1134, 2007.

[TRB+16] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204. IEEE, 2016.

[TTN+17] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017.

[VDODZ+] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *Proc. SSW*, page 125.

[VGX] Stefaan Van Gerven and Fei Xie. A comparative study of speech detection methods. In *Proc. 5th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1095 – 1098, 1997.

[VLL+10] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[WXL+21] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

[WZ11] Ji Wu and Xiao-Lei Zhang. Maximum margin clustering based statistical VAD with multiple observation compound feature. *IEEE Signal Processing Letters*, 18(5):283–286, 2011.

[YGS89] Ben P Yuhas, Moise H Goldstein, and Terrence J Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11):65–71, 1989.

[YK16] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. 2016.

[YNO10] Takami Yoshida, Kazuhiro Nakadai, and Hiroshi G. Okuno. An improvement in audio-visual voice activity detection for automatic speech recognition. In *Proc. 23rd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 51–61, 2010.

[YSW+20] Xinchen Ye, Baoli Sun, Zhihui Wang, Jingyu Yang, Rui Xu, Haojie Li, and Baopu Li. Pmbanet: Progressive multi-branch aggregation network for scene depth super-resolution. *IEEE Transactions on Image Processing*, 29:7427–7442, 2020.

[YYDN07] Qingxiong Yang, Ruigang Yang, James Davis, and David Nistér. Spatial-depth super resolution for range images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[YYL+14] Jingyu Yang, Xinchen Ye, Kun Li, Chunping Hou, and Yao Wang. Color-guided depth recovery from rgb-d data using an adaptive autoregressive model. *IEEE transactions on image processing*, 23(8):3443–3458, 2014.

[ZFY+19] Yifan Zuo, Yuming Fang, Yong Yang, Xiwu Shang, and Bin Wang. Residual dense network for intensity-guided depth map enhancement. *Information Sciences*, 495:52–64, 2019.

[ZLL+18] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[ZPZ+20] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049, 2020.

[ZSL+20] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[ZW13] Xiao-Lei Zhang and Ji Wu. Deep belief networks based voice activity detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(4):697–710, 2013.

[ZZGZ17] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017.

[ZZHG16] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. Multimodal deep convolutional neural network for audio-visual emotion recognition. In *Proc. ACM on International Conference on Multimedia Retrieval (ICMR'16)*, pages 281–284. ACM, 2016.

[ZZX+21] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Chunxia Zhang, and Junmin Liu. Discrete cosine transform network for guided depth map super-resolution. *arXiv preprint arXiv:2104.06977*, 2021.

נושא המחקר השני שתזה זו עוסקת בו הינו סופר רזולוציה מונחית של מידע עומק. מידע עומק שנלכד על ידי חיישני עומק לרוב מאופיין ברזולוציה מרחבית נמוכה, מה שמגביל את היישומים הפוטנציאליים. בעבר הוצעו מספר שיטות לסופר רזולוציה מונחית של מפות עומק באמצעות רשתות קונבולוציה עמוקות כדי להתגבר על מגבלה זו. במערכות של סופר רזולוציה מונחית, מפות עומק ברזולוציה גבוהה מוסקות ממפות ברזולוציה נמוכה עם הנחיה נוספת של תמונת צבע מתאימה ברזולוציה גבוהה. עם זאת, שיטות אלה עדיין מועדות לבעיות של העתקת יתר של טקסטורה מתמונת הצבע, הנובאות מהנחיה לא אופטימלית ידי תמונת הצבע. ספציפית, ברוב השיטות הקיימות, הנחיה מתמונת הצבע מושגת על ידי שרשור נאיבי של וקטורי מאפיינים מתמונות הצבע והעומק. אנו מציעים רשת עמוקה לסופר רזולוציה של מפות עומק אשר בה מודול טרנספורמר היררכי משלב מידע מבני ברזולוציה גבוהה מתמונת הצבע לתוך תהליך הגדלת תמונת העומק.

בנוסף, אנו מציעים רשת נוספת המבוססת כולה על טרנספורמרים לפתרון בעיית הסופר רזולוציה של מפת עומק. מודול טרנספורמר היררכי מחלץ מאפיינים משמעותיים מתמונת העומק. הוא משלב מנגנון חדשני של תשומת לב מוצלבת (cross-attention) כדי לספק הנחיה רציפה ומתמשכת מתמונת הצבע אל תהליך העלאת הדגימה של תמונת העומק. באמצעות חלוקת תמונת הקלט לחלונות בגודל קבוע, ניתן להשיג סיבוכיות ליניארית ברזולוציית התמונה, כך שניתן ליישם שיטה זו גם על תמונות ברזולוציה גבוהה. השיטות המוצעות לסופר רזולוציה מונחית לתמונות עומק השיגו ביצועים עדיפים על פני שיטות מתקדמות אחרות, כפי שמודגם באמצעות ניסויים נרחבים.

תקציר

התחום של למידה רב-מודאלית, למידה ממספר אופנים, התקדם רבות בעשור האחרון, וספציפית בתחומים של ראייה ממוחשבת ועיבוד אודיו. המחקר המוצג במסגרת תזה זו כולל תכן ופיתוח של רשתות נוירונים עמוקות ורב-מודאליות עם יישומים בתחומים של זיהוי פעילות דיבור אודיו-ויזואלית וסופר רזולוציה מונחית של תמונות עומק.

נושא המחקר הראשון שבה מתמקדת תזה זו היא זיהוי פעילות דיבור אודיו-ויזואלית. ספציפית, אנו חוקרים את הנושא של זיהוי דיבור בסביבות אקוסטיות מאתגרות, כגון רמות רעש גבוהות והפרעות חולפות (כגון נקישות מקלדת) הנפוצים בתרחישים רבים. התרחיש בו אנו מתמקדים הוא זה של אות אודיו ווידאו הנקלט על ידי מצלמה המכוונת אל פני הדובר בסביבה מולטי-מודאלית. בהתאם לכך, זיהוי דיבור מתורגם לשאלה כיצד למזג נכון את אותות האודיו והווידאו ולקבל החלטה על הימצאות/אי הימצאות אות דיבור על סמך הסיגנל האחוד. אנו מציעים להתמודד עם סוגיה זו במסגרת של למידה עמוקה רב-מודאלית. אנו מציגים ארכיטקטורה המבוססת על וריאציה של מקודד אוטומאטי ,(autoencoder) המשלבת את שני האופנים, ומספקת ייצוג חדש של האות ששבו מופחתים הרעשים וההפרעות. הייצוג החדש מוזן לרשת עצבית זמנית מסוג RNN לזיהוי דיבור. רשת זו מאומנת באופן מפוקח לשם קידוד נוסף של ההבדלים בין דינמיקה של דיבור ורעשים חולפים.

בעבודת המשך, אנו מציעים לשלב את אותות האודיו והוידאו באמצעות רשת עצבית עמוקה המאומנת מקצה לקצה לזיהוי דיבור. בתנאים רועשים, יש לחלץ מאפיינים משמעותיים ורלוונטים משני האופנים כדי להבחין במדויק בין דיבור לרעש. אנו משתמשים ברשת שיורית עמוקה כדי לחלץ מאפיינים אלה מאות הווידאו, בעוד שמקודד מסוג WaveNet משמש לחילוץ מאפיינים מאות האודיו. כדי ליצור ייצוג משותף של שני האותות, וקטורי המאפיינים משני האותות מאוחדים באמצעות רכיב מסוג pooling. bilinear compact multimodal לאחר מכן אנו מאמנים את המערכת בצורה מפוקחת מקצה לקצה באמצעות רשת זמנית מסוג LSTM לקידוד המידע הזמני.

המחקר בוצע בהנחייתו של פרופסור ישראל כהן בפקולטה להנדסת חשמל ומחשבים על שם אנדרו וארנה ויטרבי.

מחבר חיבור זה מצהיר כי המחקר, כולל איסוף הנתונים, עיבודם והצגתם, התייחסות והשוואה למחקרים קודמים וכו', נעשה כולו בצורה ישרה, כמצופה ממחקר מדעי המבוצע לפי אמות המידה האתיות של העולם האקדמי. כמו כן, הדיווח על המחקר ותוצאותיו בחיבור זה נעשה בצורה ישרה ומלאה, לפי אותן אמות מידה.

תודות

ברצוני להביע את תודתי הכנה למנחה האחראי שלי, פרופ' ישראל כהן, על ההנחיה הצמודה לכל אורך הדרך. העצות שלו, התמיכה הבלתי-פוסקת ורעיונותיו המועילים תרמו בצורה רבה ומשמעותית למחקר שלי.

לבסוף, ברצוני להודות למשפחתי על ההבנה ועל העזרה במהלך המסע של לימודי הדוקטורט שלי.

רשתות נוירונים רב מודאליות עם יישומים לזיהוי דיבור וסופר רזולוציה מונחית

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר

דוקטור לפילוסופיה

עידו אריאב

הוגש לסנט הטכניון — מכון טכנולוגי לישראל

תמוז תשפ"ג חיפה יולי 2023

רשתות נוירונים רב מודאליות עם יישומים לזיהוי דיבור וסופר רזולוציה מונחית

עידו אריאב