Multi-Modal Deep Neural Networks

With Applications to Voice Activity Detection *and* Guided Super Resolution

Ido Ariav, Supervised by Prof. Israel Cohen , May 2023

OUTLINE

- Introduction
- Voice Activity Detection (in a nutshell)
- Guided Super Resolution –
  - Background
  - Transformer based guidance
  - Cross-attention transformer
- Discussion & Future Work

INTRODUCTION

Why Multimodal?

“

if it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck

But if it only *looks* like a duck...

A grebe

A coot

A loon

But if it only *sounds* like a duck...

If It Sounds Like a Duck It Might Be a Frog

Something Wild

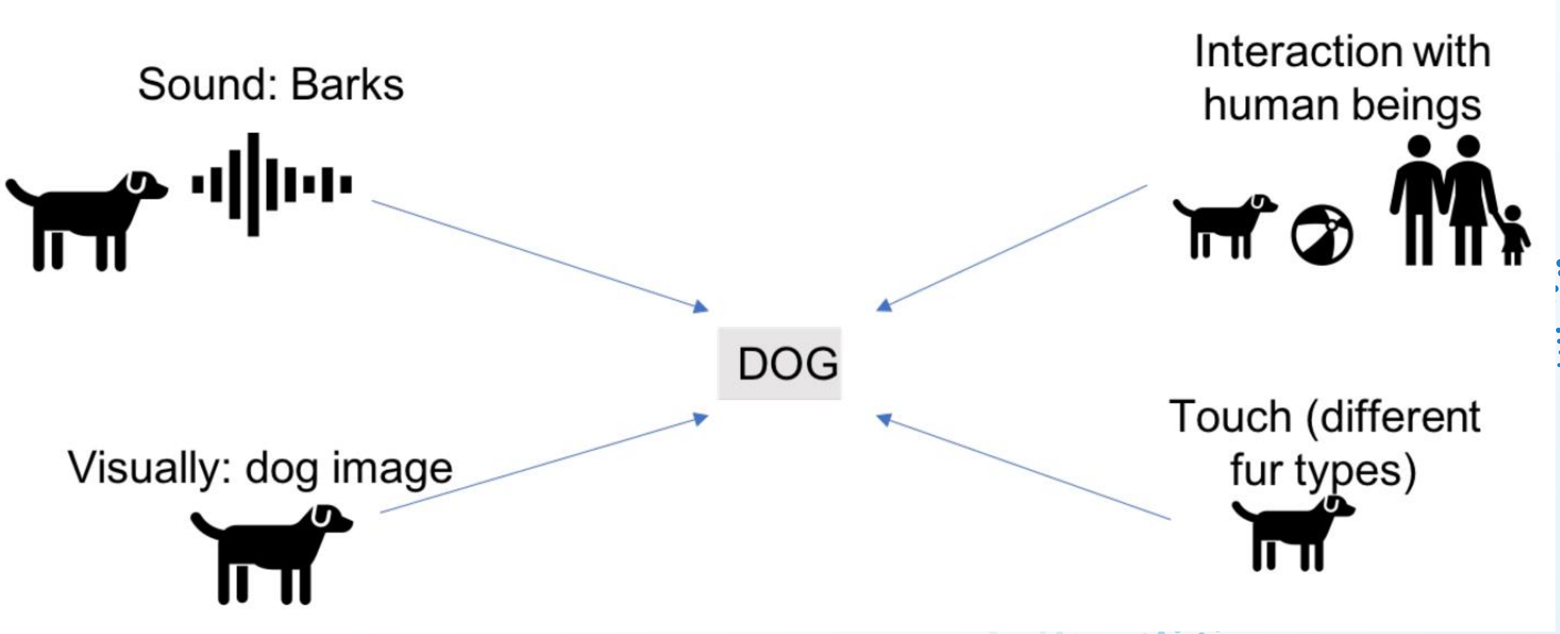
Chris Martin | April 3, 2015

Wildlife

The Wood frog chorus sounds like quacking ducks.

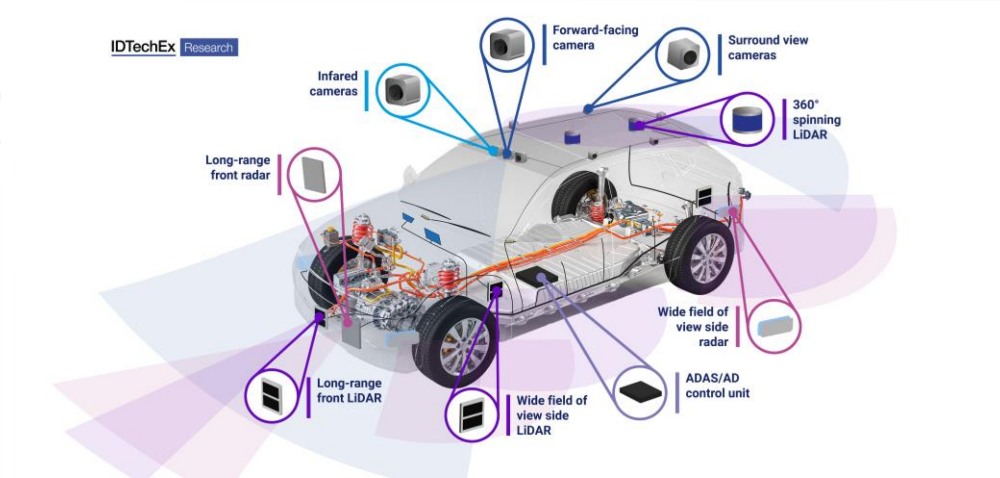
If you're out for a walk this month, and you hear something that sounds like ducks quacking, don't expect to see ducks. The call of a male wood frog fools a lot of people. The all-male frog chorus is revving up now, and wood frog males are the first to announce their availability to females.

We need multiple modalities



We need multiple modalities

Luckily, it's a Multimodal world



Voice Activity Detection

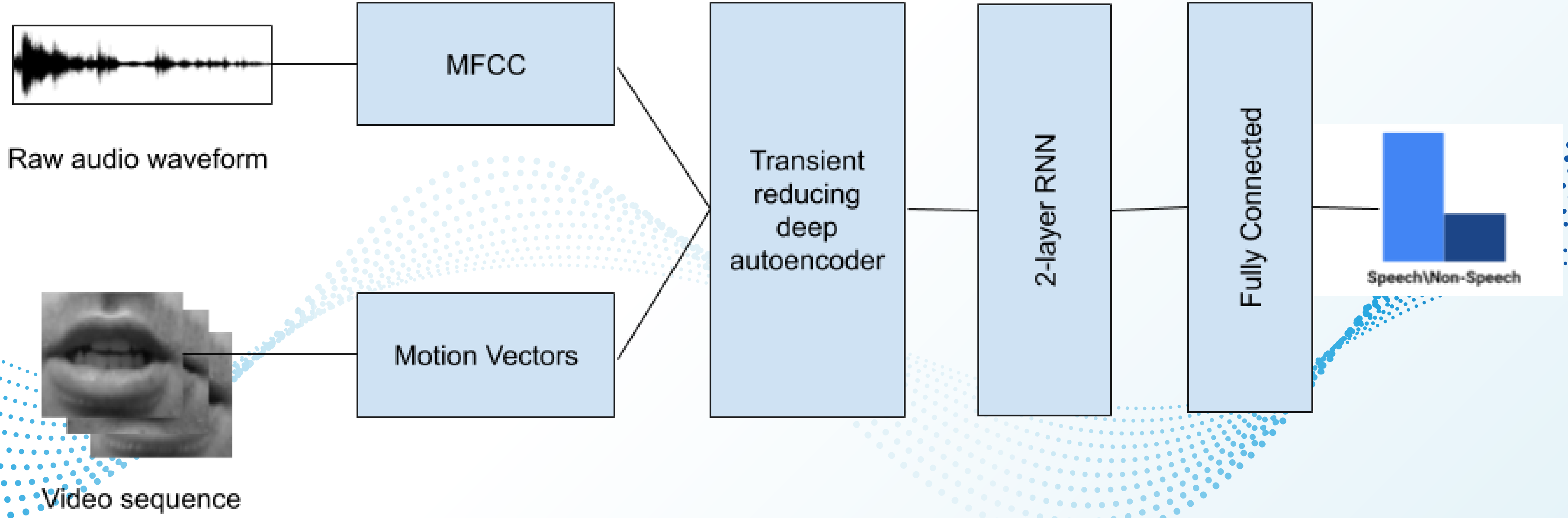
Publications

- “A deep architecture for audio-visual voice activity detection in the presence of transients”, Elsevier Signal Processing, 2017

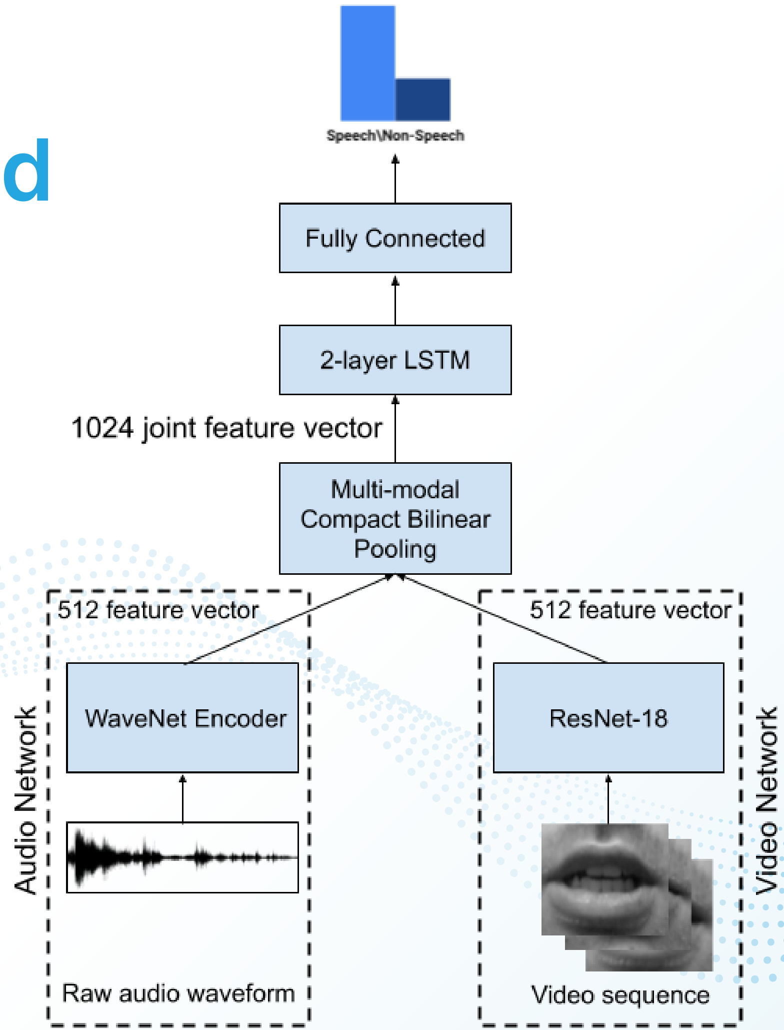
- “An End-to-End Multimodal Voice Activity Detection Using WaveNet Encoder and Residual Networks”, IEEE Journal of Selected Topics in Signal Processing, 2019

Proposed Method

- A multimodal deep neural architecture

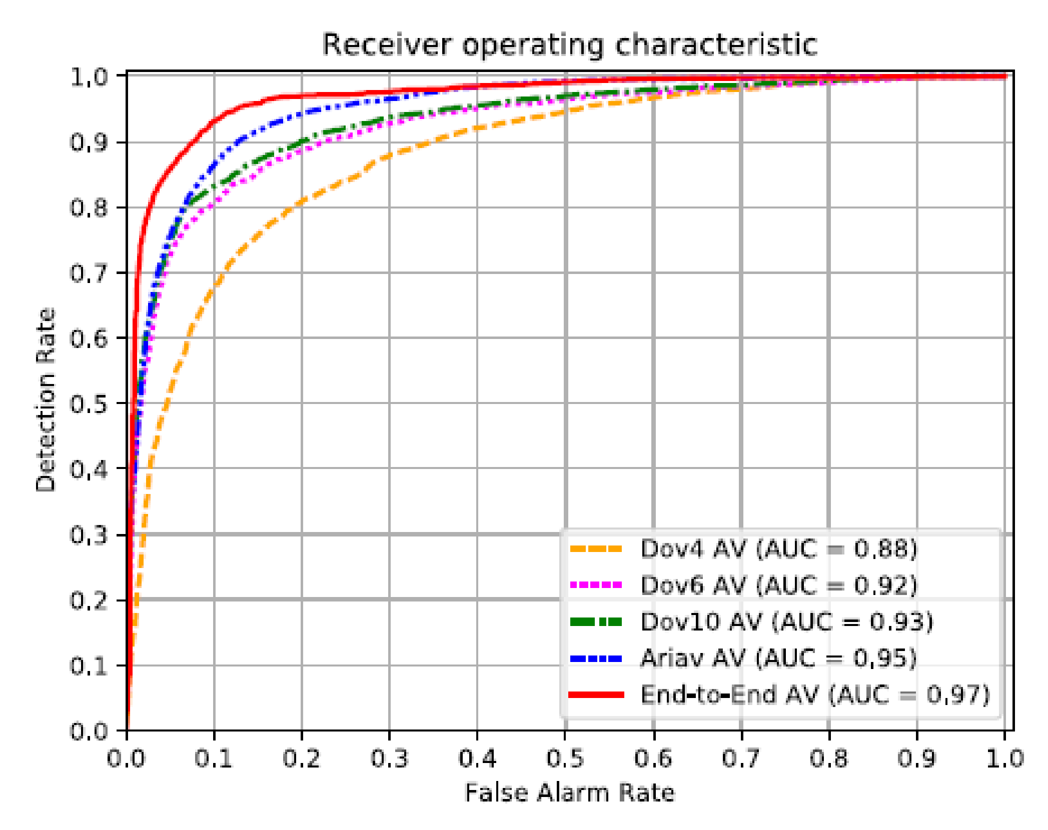


Proposed Method



Experimental Results

Comparison of our method to "Audio-Visual Voice Activity Detection Using Diffusion Maps" by Dov et al. and our previous work



Depth Super Resolution

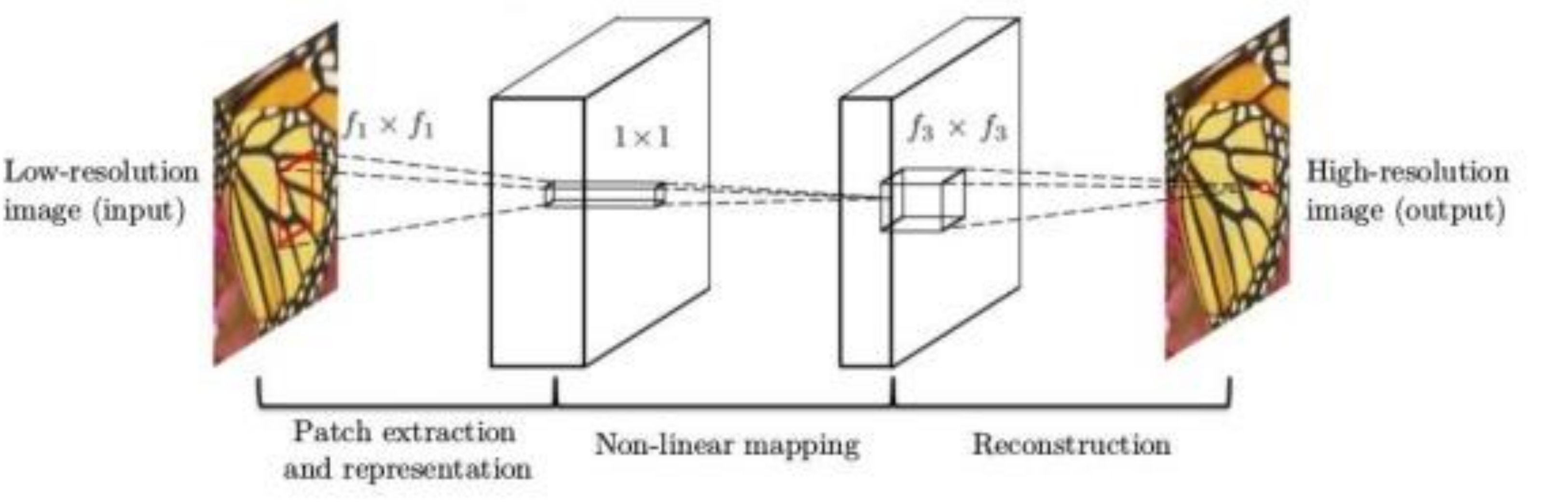
Publications

- “Depth Map Super-Resolution via Cascaded Transformers Guidance”, Frontiers in Signal Processing, 2022
- “Fully Cross-Attention Transformer for Guided Depth Super Resolution”, MDPI Sensors Special Issue on Deep Learning Technology and Image Sensing, 2023

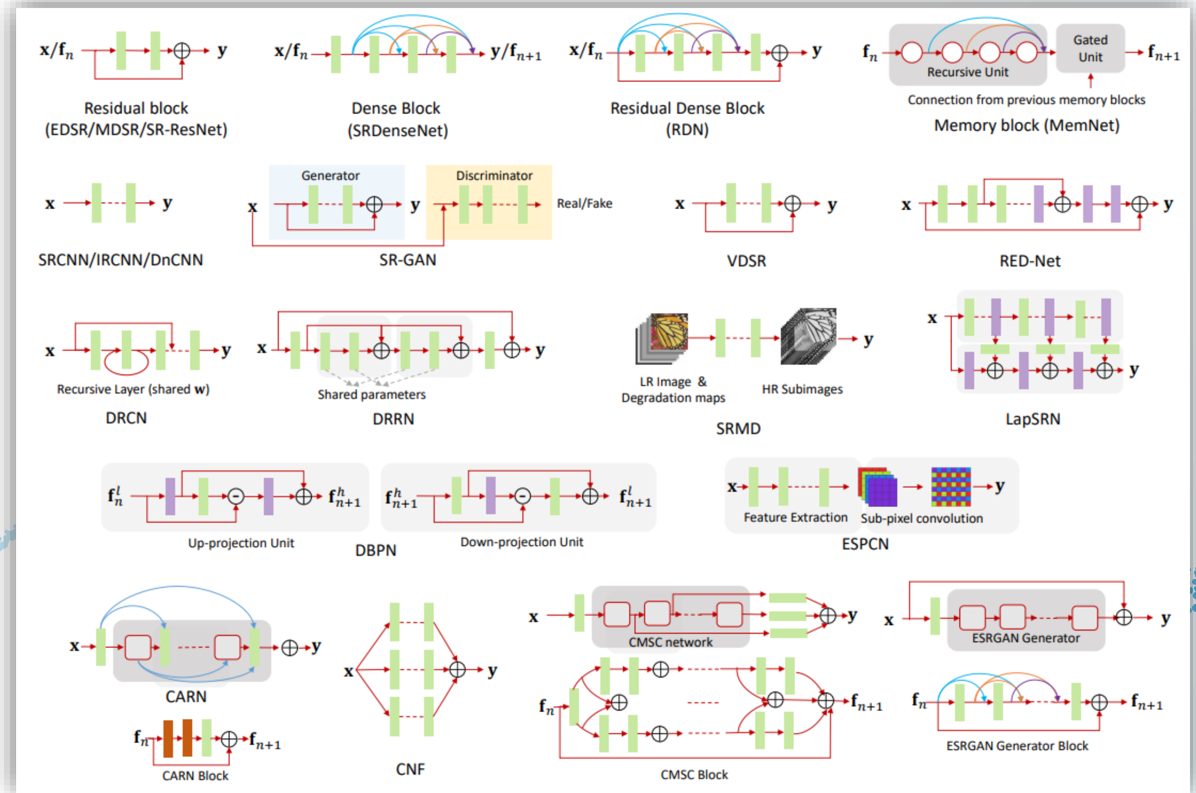
Super Resolution

Super Resolution

- Since 2015 (SRCNN), deep learning took over the field of SR

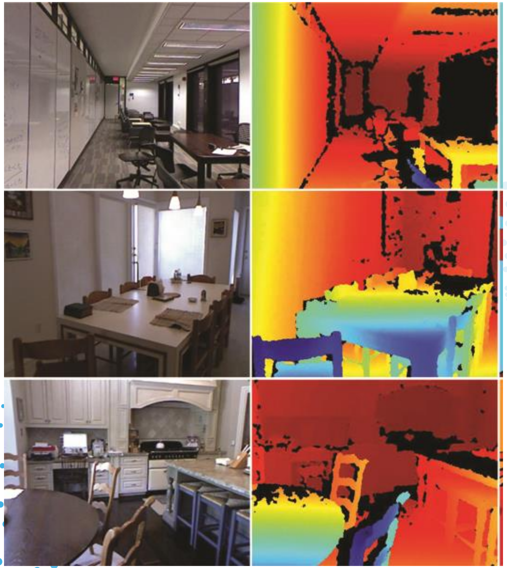


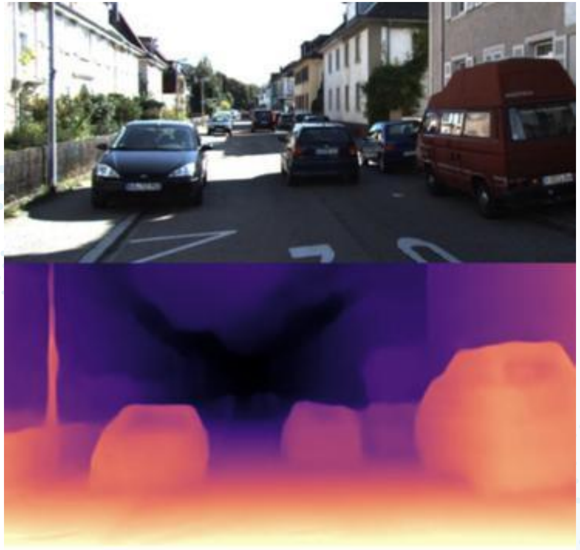
Super Resolution

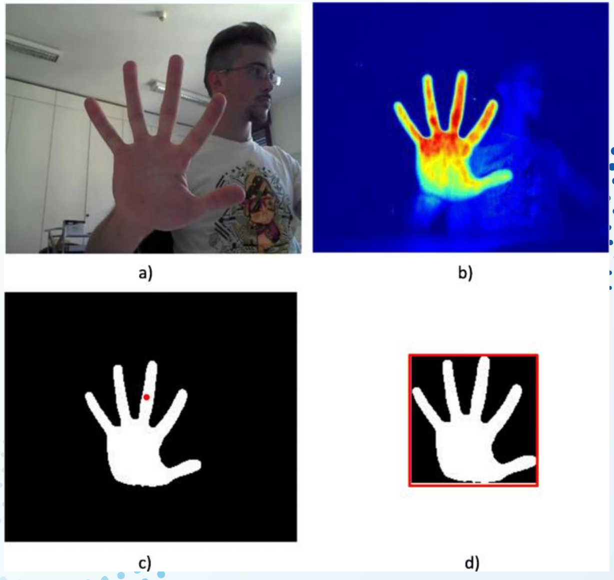


Depth Super Resolution

- Depth plays a vital role in many real-life scenarios -

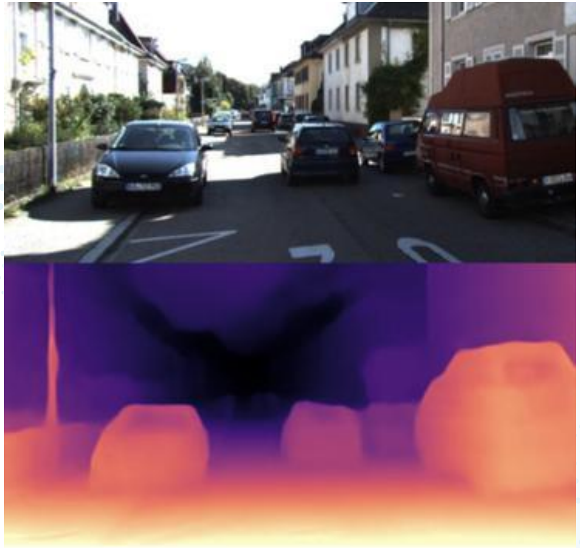


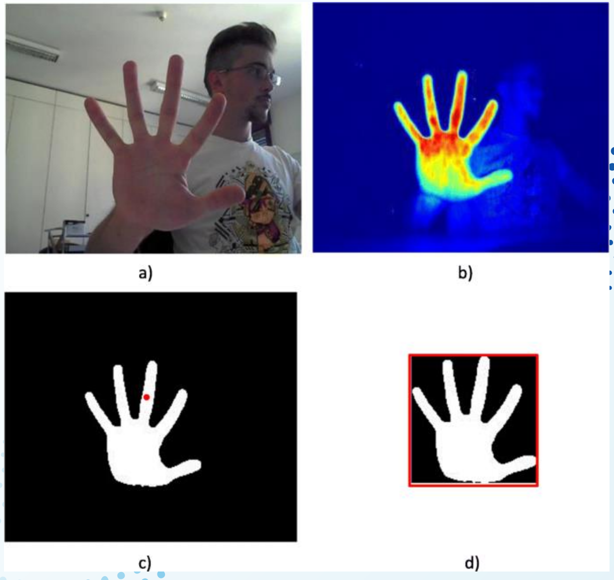




Depth Super Resolution

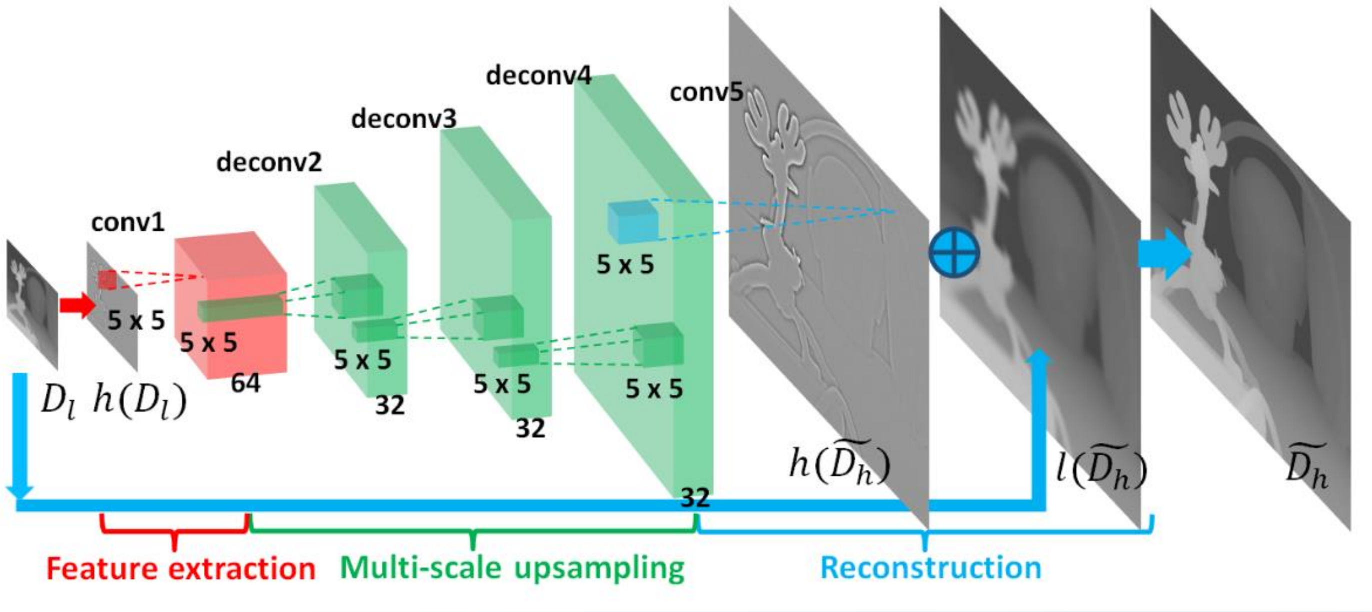
- However, depth sensors usually have a low spatial resolution





Depth Super Resolution

- Existing SR methods gave limited results when applied to DSR



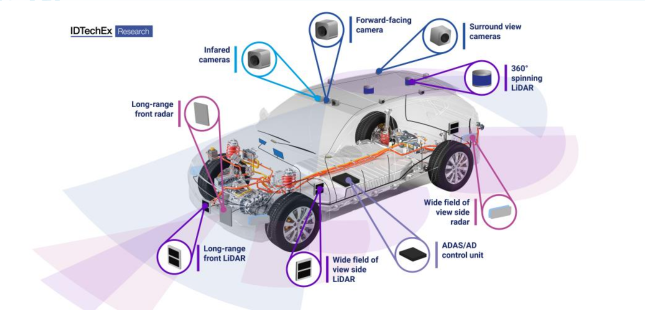
Depth Super Resolution

- Why? - intrinsic differences between color and depth images. depth maps:
    - generally contain less textures and more sharp boundaries
    - are usually degraded by noise due to the imprecise acquisition sensors
- The difficulty in capturing HR depth maps further increases the challenge

Depth Super Resolution

- Solution - Adding Another Modality

- incorporate HR image as guidance, since they contain plenty of useful high-frequency components which assist the process of DSR

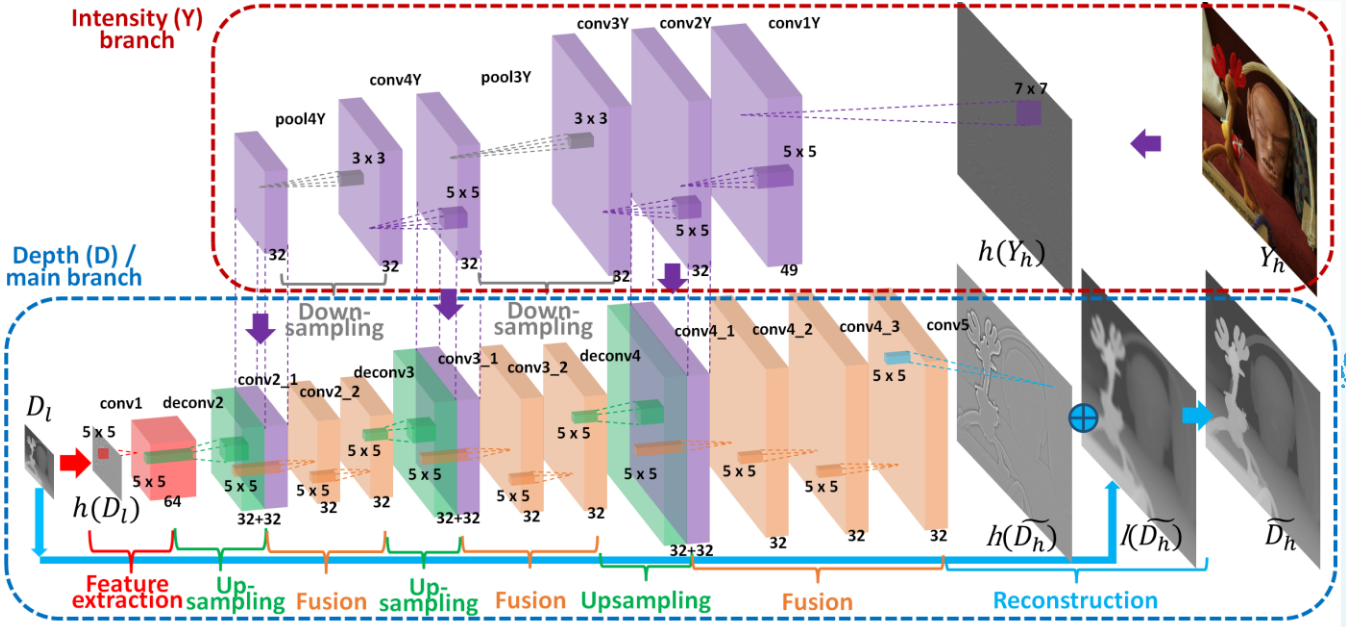


Depth Super Resolution

- **Solution - Adding Another Modality**

Depth Super Resolution

- **Adding Another Modality –**



Depth Super Resolution

- Drawbacks -
  - Texture copying -

  - Naïve guidance
  - Limited receptive field

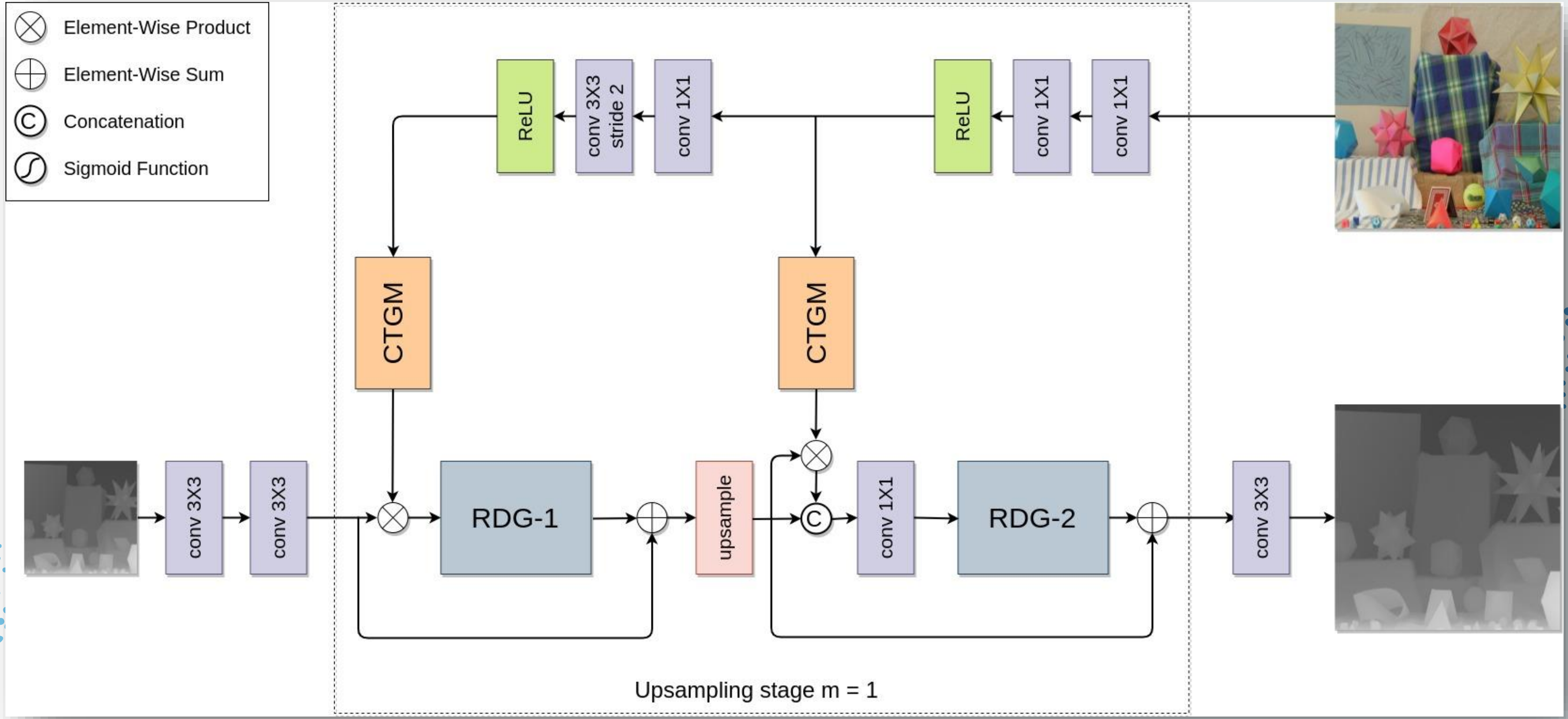
Proposed Method

"Depth Map Super-Resolution via Cascaded Transformers Guidance", Frontiers in Signal Processing, 2022

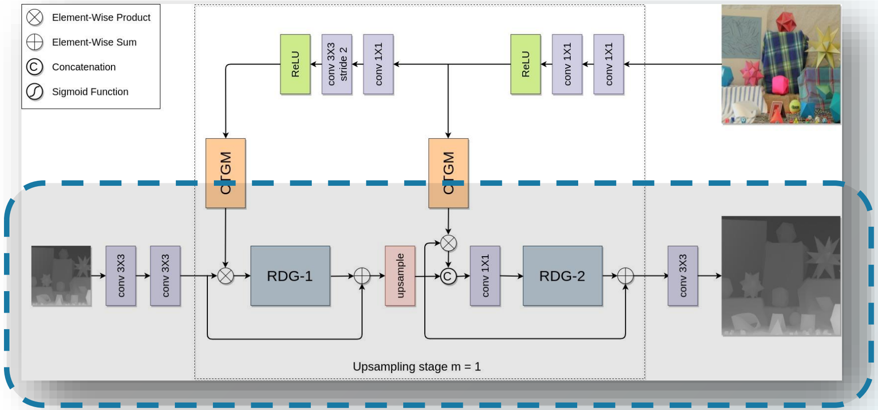
Proposed Method

- Depth upsampling via *Residual Dilated Groups*
- a *cascaded transformer-based guidance* mechanism from the intensity branch
- linear memory constraints - applicable even for very large images
- can handle different input resolutions - applicable to real-world tasks

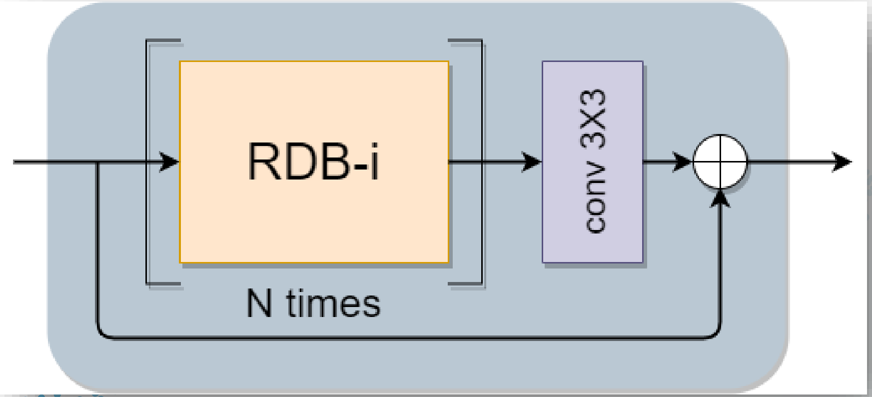
Proposed Method

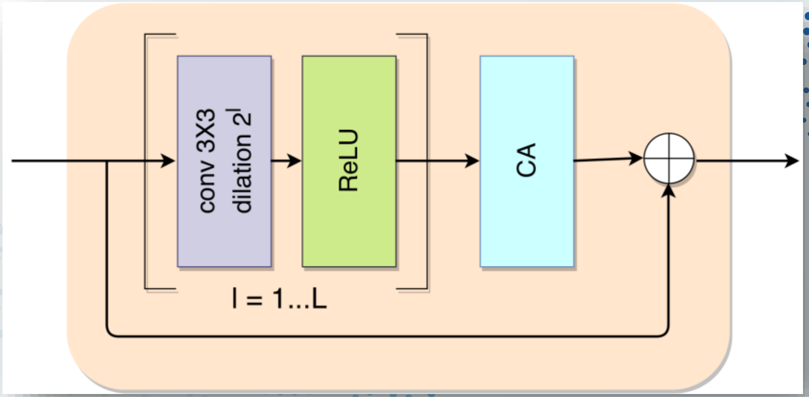


Proposed Method



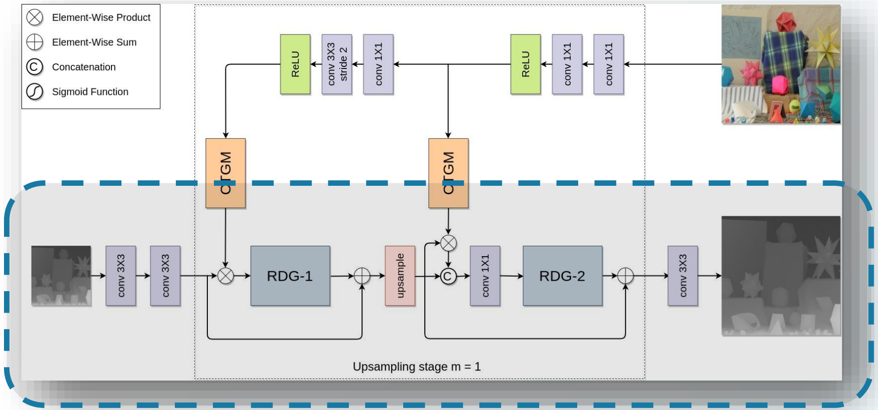
- **Depth Upsampling Branch -**

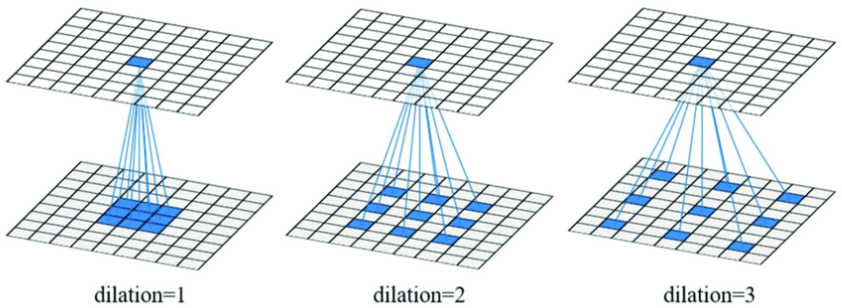


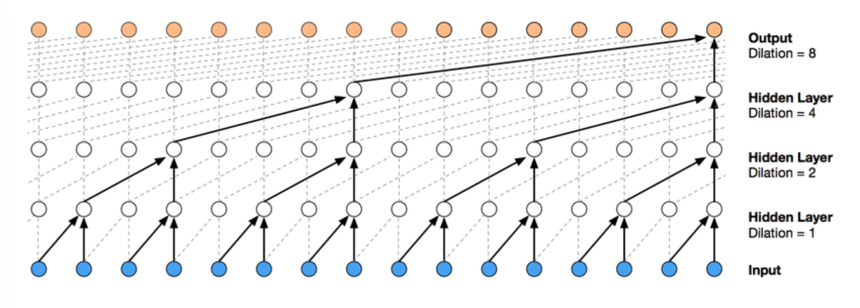


Proposed Method

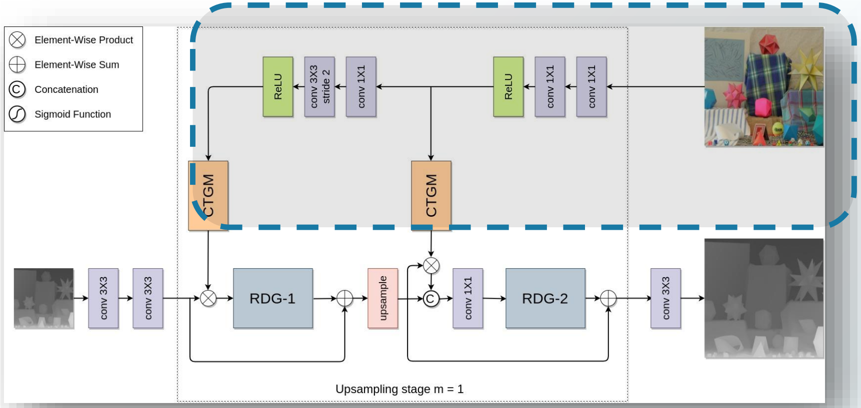
- dilated convolutions - increase the receptive field







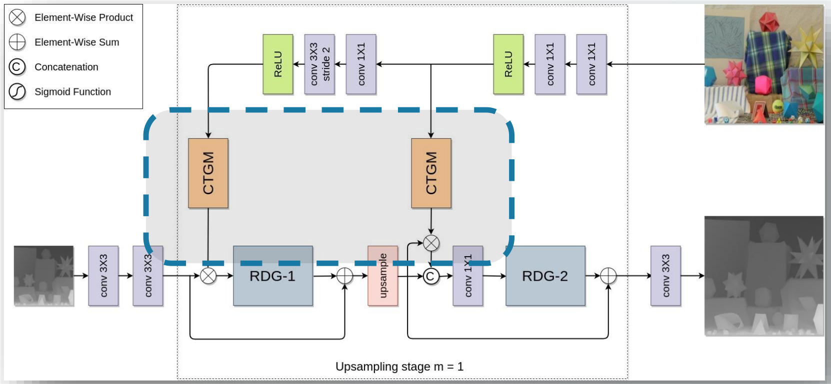
Proposed Method



- **RGB Branch -**

  simple convolutions with appropriate strides to do downsampling + ReLU

Proposed Method



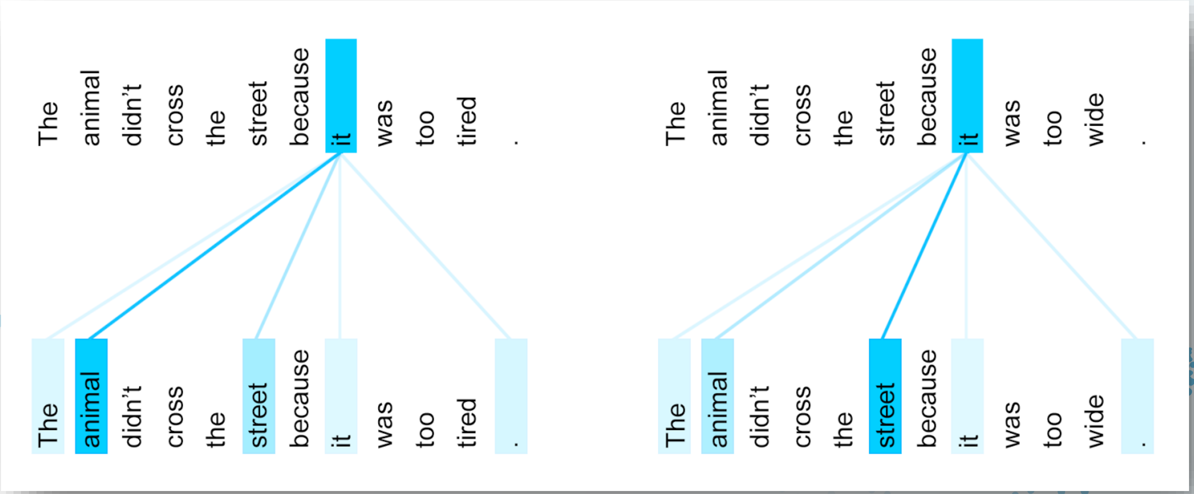
- **Cascaded Transformer Guidance Module -**
  - **used to scale the corresponding depth features in the depth branch by element-wise multiplication**

A Short intro on Vision transformers

- **What is Attention?**
  - **Transformers originated in text\NLP regimes**
  - **The Attention mechanism enables the transformers to have extremely long-term memory**
  - **A transformer model can “attend” or “focus” on all previous tokens**

A Short intro on Vision transformers

- **What is Attention?**

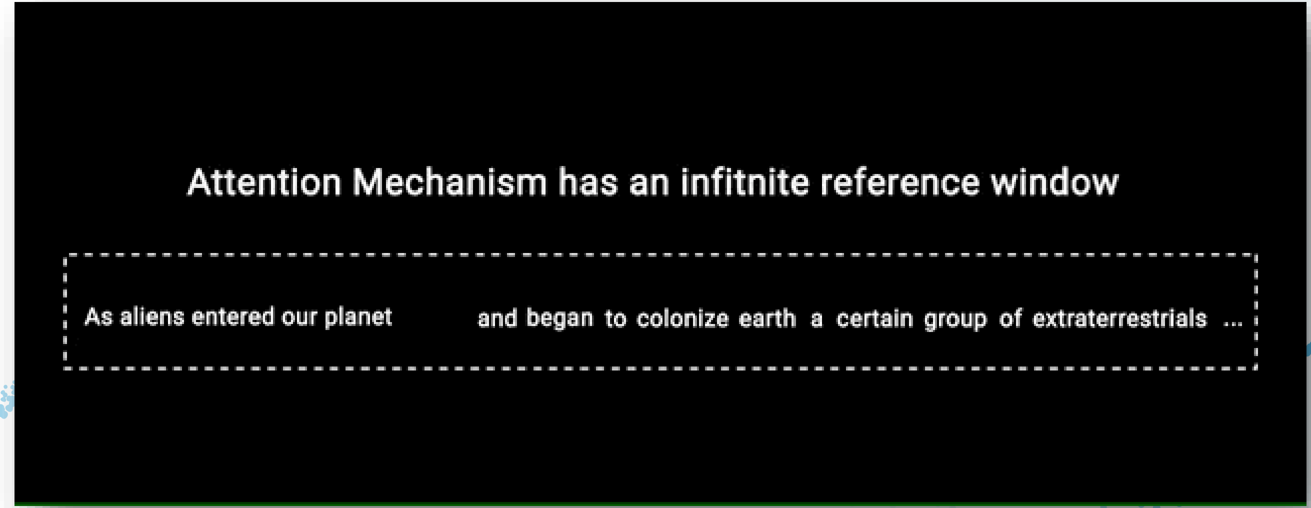


A Short intro on Vision transformers

- **What is Attention?**

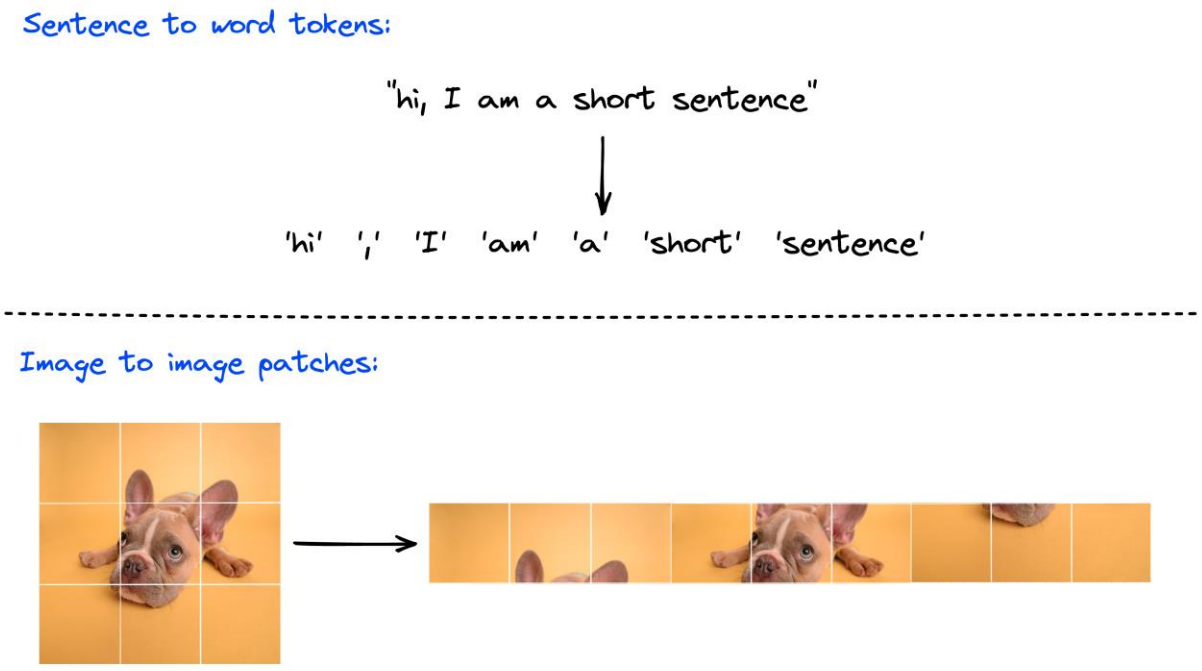
A Short intro on Vision transformers

- **What is Attention?**



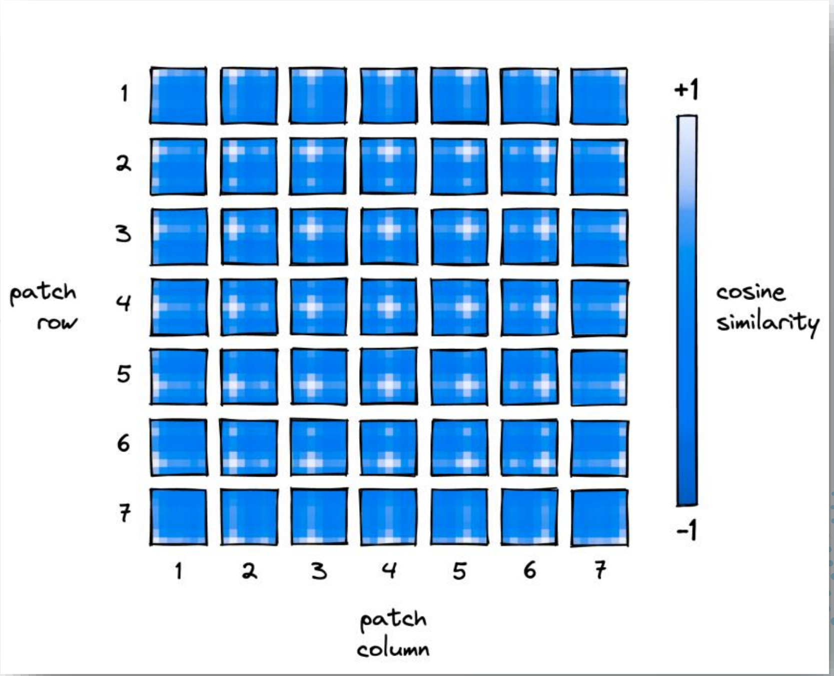
A Short intro on Vision transformers

- **Transformer for images**



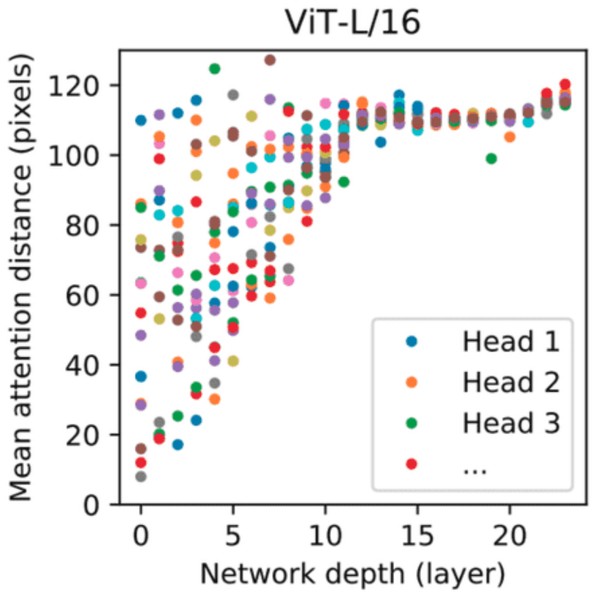
A Short intro on Vision transformers

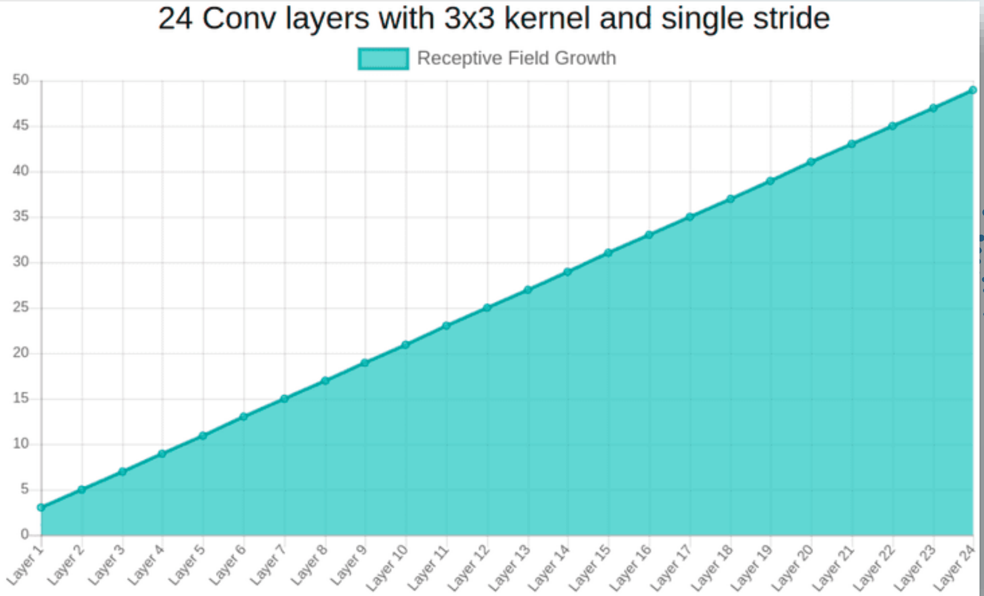
- **Transformer for images – positional embeddings**



A Short intro on Vision transformers

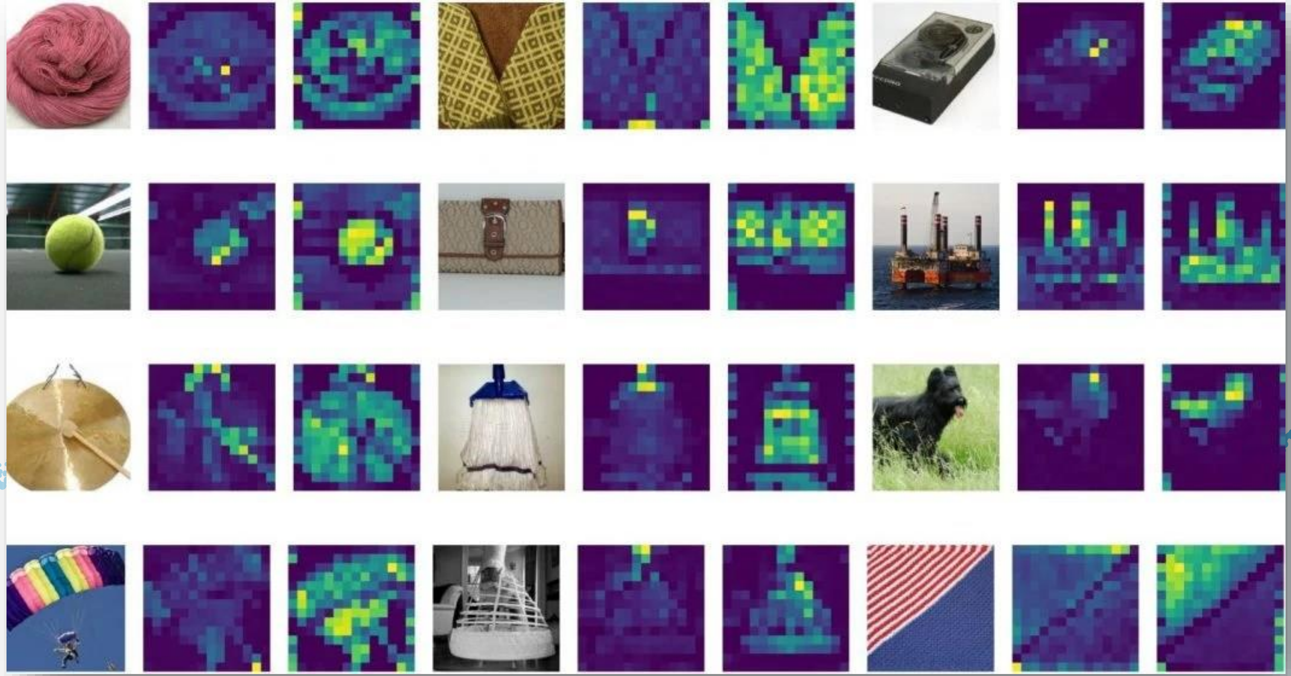
- **Increased receptive field**



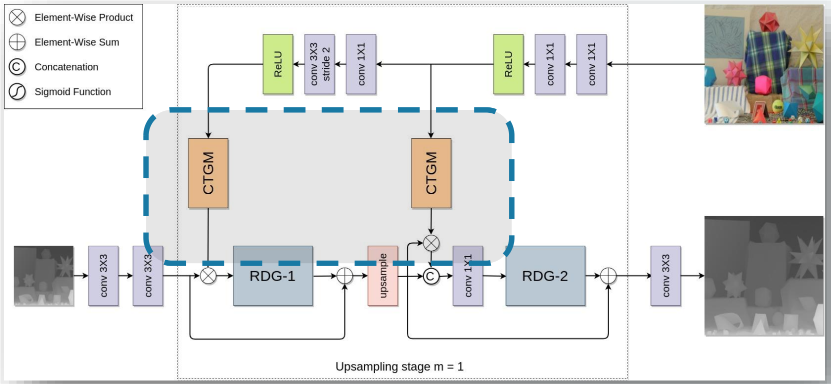


A Short intro on Vision transformers

- **Attention assigns different weights to different parts of the input**



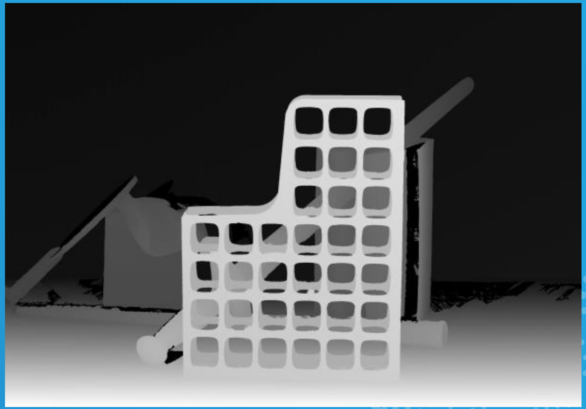
Proposed Method

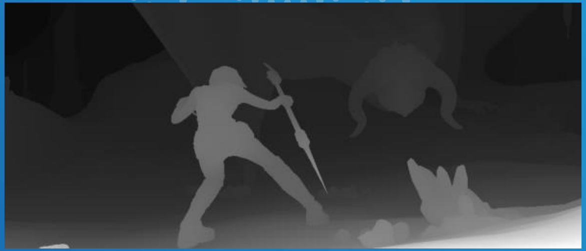


- **Cascaded Transformer Guidance Module -**
  - **learns structural and content information from a large receptive field**
  - **encode distant dependencies leveraging both local and global information for guidance**

Datasets -

Middlebury & MPI Sintel





Used LR patches of size 96/96/48/24 for scaling factors 2/4/8/16

Experimental Results

SOTA in terms of RMSE on almost all datasets & scaling factors

Method	Art x2	Art x4	Art x8	Art x16	Books x2	Books x4	Books x8	Books x16	Laundry x2	Laundry x4	Laundry x8	Laundry 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**

Method	Dolls x2	Dolls x4	Dolls x8	Dolls x16	Moebius x2	Moebius x4	Moebius x8	Moebius x16	Reindeer x2	Reindeer x4	Reindeer x8	Reindeer 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**

Experimental Results

Robust under various noises in both depth & guidance image

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

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

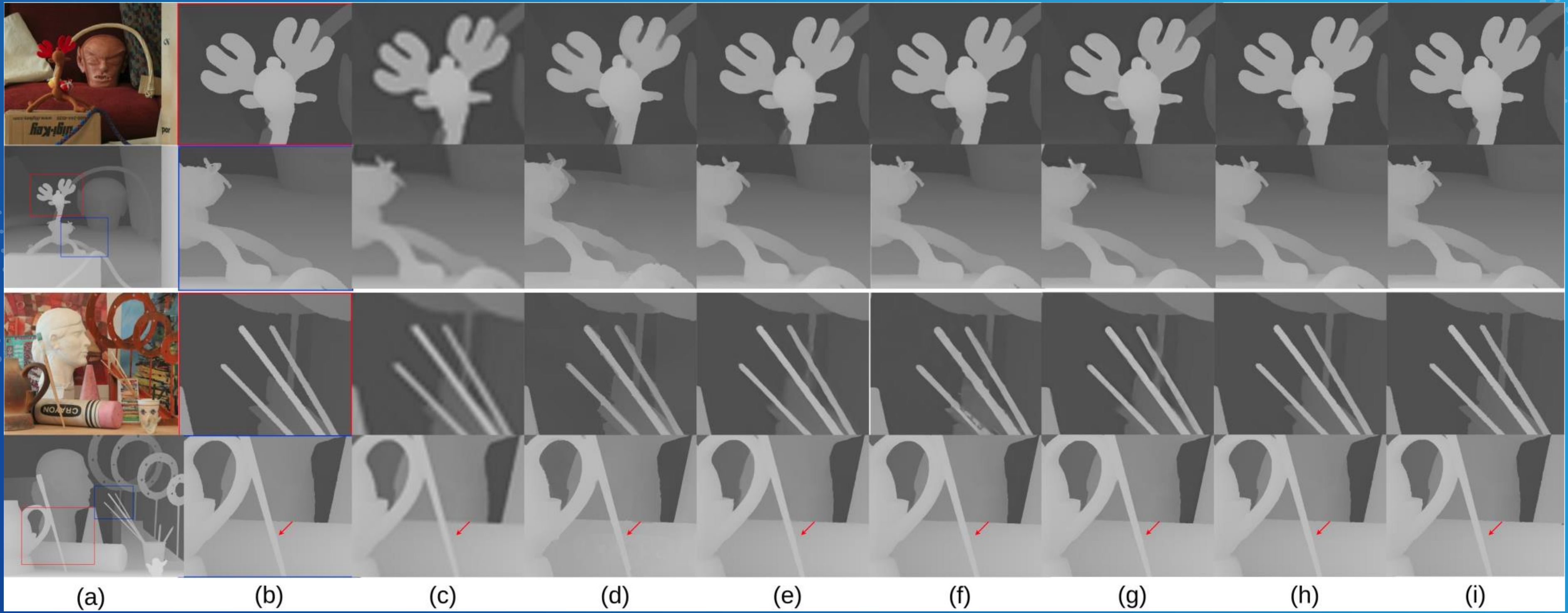
Experimental Results

Generalizes to unseen datasets (NYU_Depth_V2)

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**

Experimental Results

Reduces "texture copying" effect



Proposed Method

- **Discussion –**
  - Guidance based on cascaded transformer with large receptive field
  - Linear memory constraints – applicable to large images and real-life scenarios
  - Good generalization abilities and insensitivity to noise

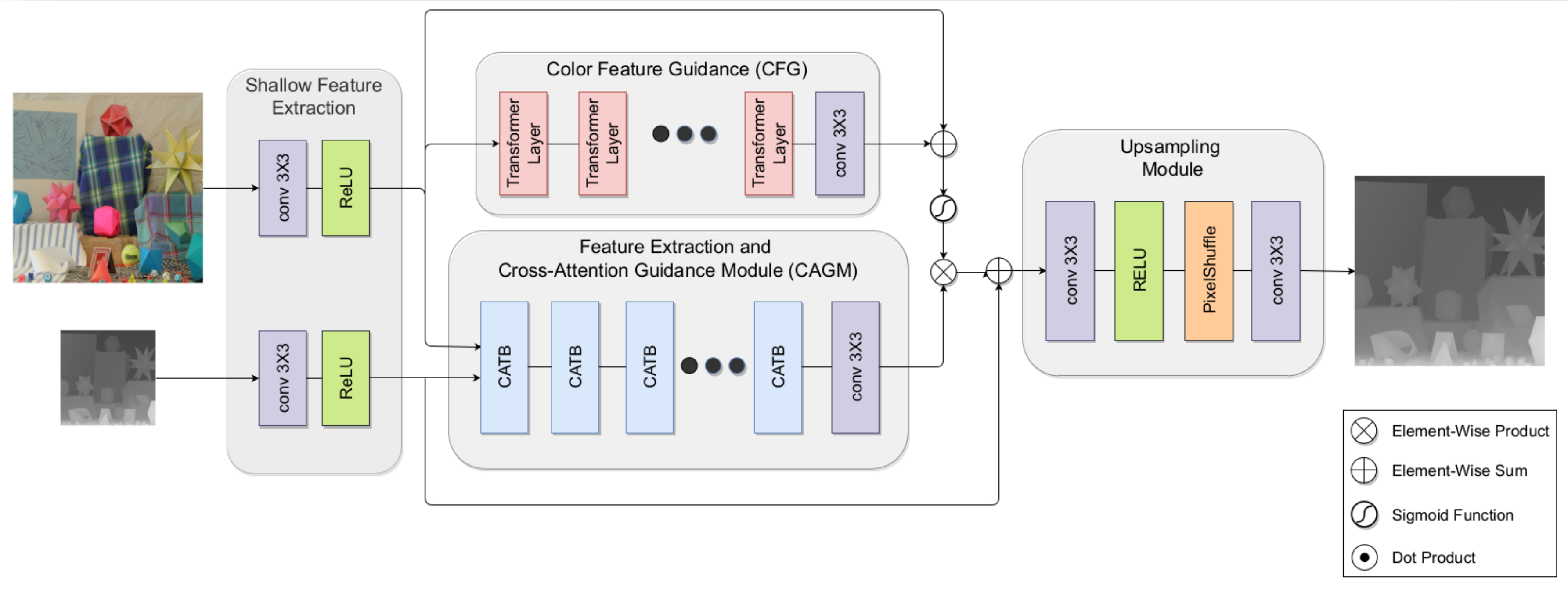
Proposed Method

“Fully Cross-Attention Transformer for Guided Depth Super Resolution”, MDPI Sensors Special Issue on Deep Learning Technology and Image Sensing, 2023

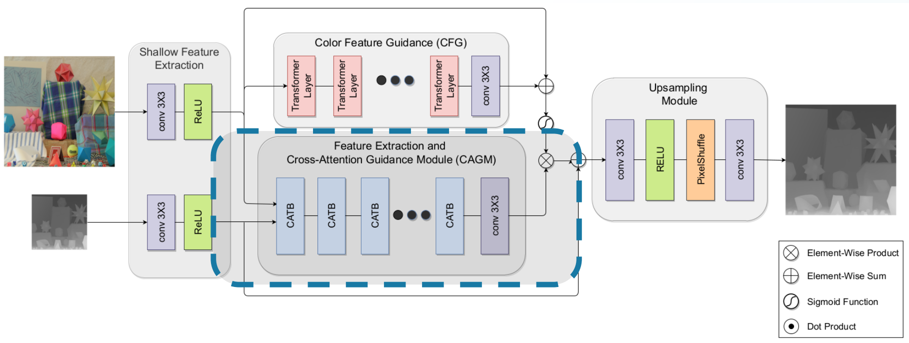
Proposed Method

- Fully transformer based architecture
- Guidance via cross-attention in a single branch
- Same linear memory constraints as in previous work - applicable for very large images and different resolutions

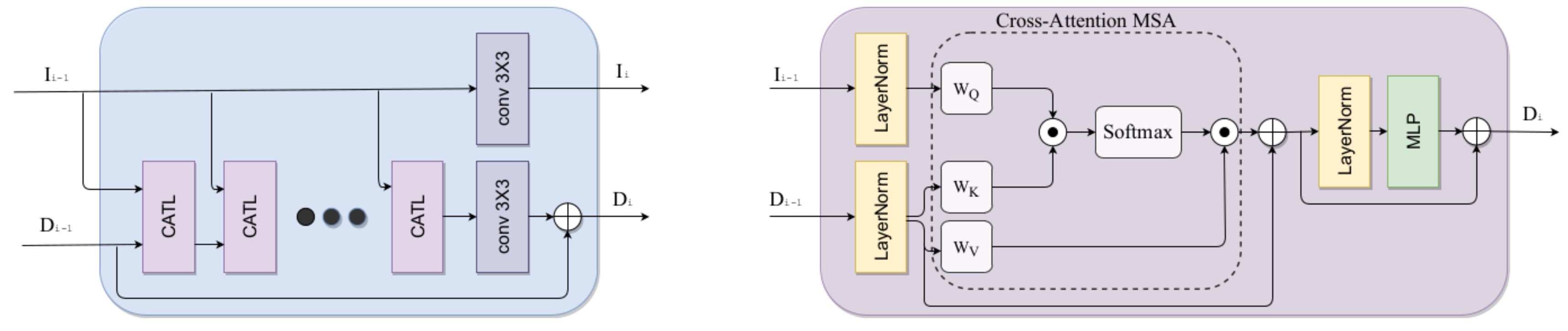
Proposed Method



Proposed Method



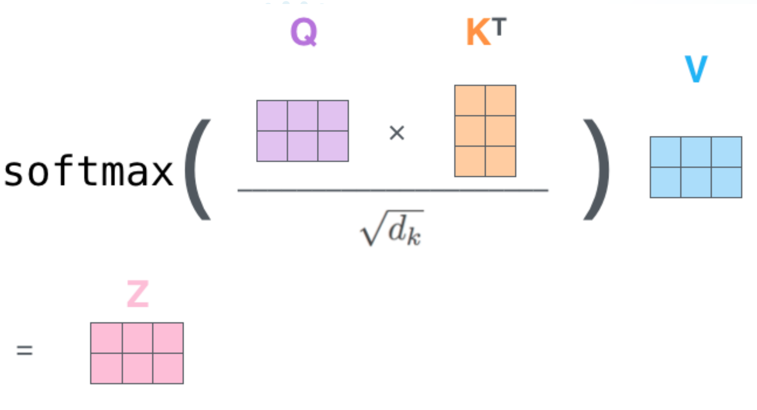
- **Main branch – combines feature extraction with guidance from the RGB image via a cross-attention design**

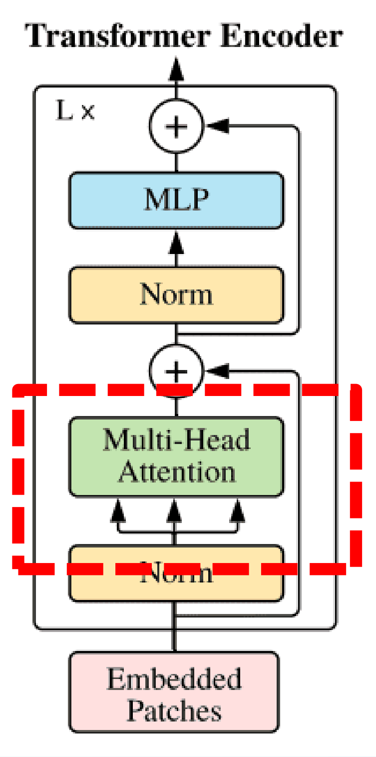


(a) (b)

Another short drill down on transformers

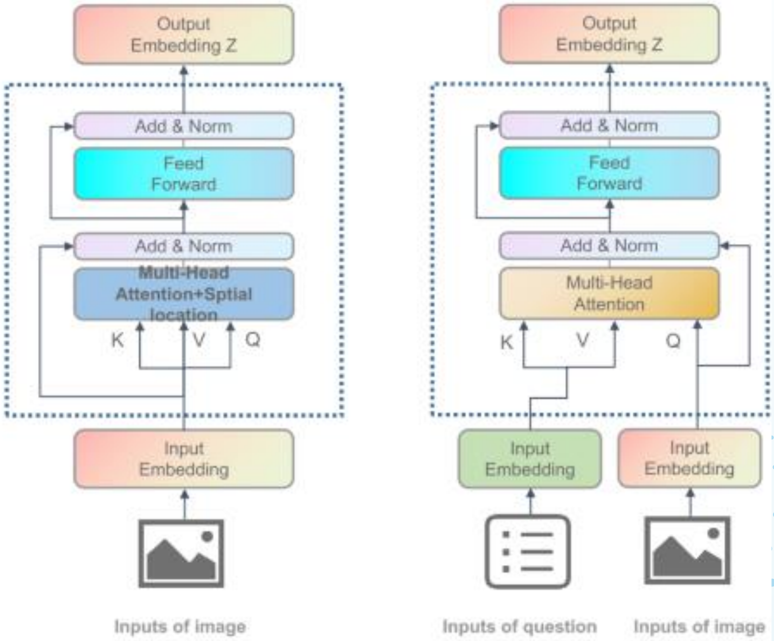
- **In a transformer encoder, attention is calculated via dot product between 3 matrices – Q, K, V**



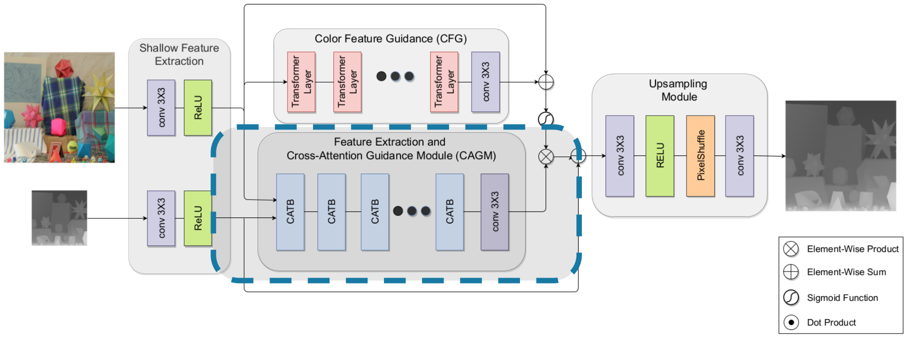


Another short drill down on transformers

- **In cross attention – K & V come from one modality, while Q comes from the other**

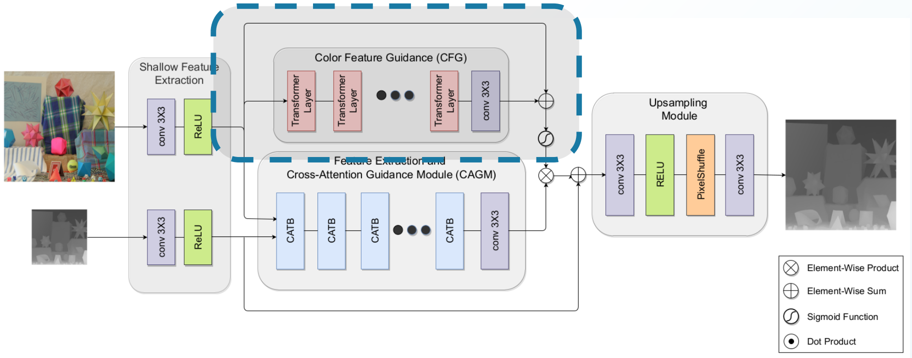


Proposed Method



- **Main branch –**
  - **the cross attention allows continues guidance from the guidance image**
  - **All elements of the guidance features can interact with all elements of the depth upsampling**

Proposed Method



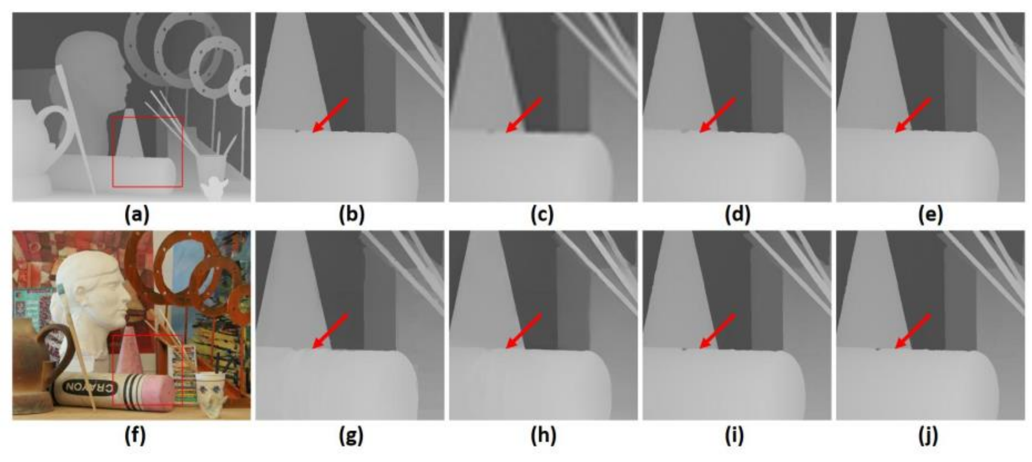
- **Color Feature Guidance -**
- **Similar design to the previous work (cascaded self-attention transformer)**
- **Scales the features before the final upsampling, incorporating more HR information**

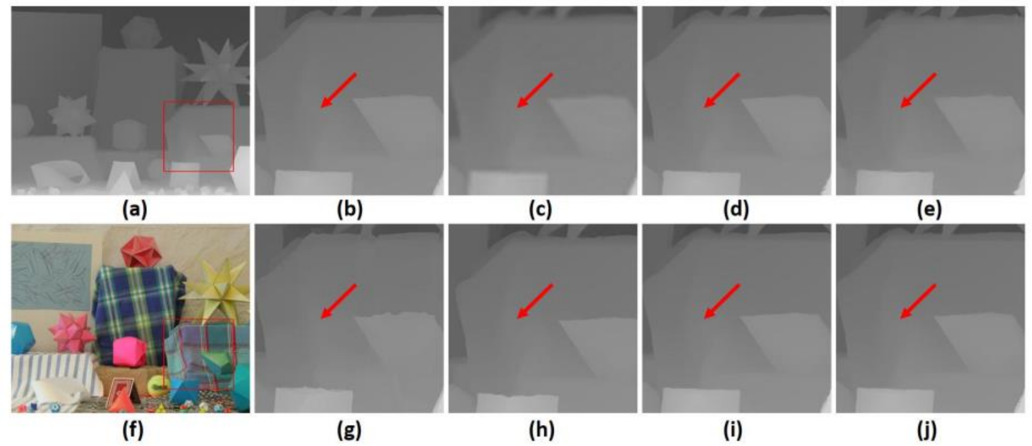
Experimental Results

improves upon previous work in all parameters –

Better reconstruction (RMSE), better generalization, faster (~20%)

Experimental Results





Experimental Results

Ablation study demonstrates the importance of cross-attention, and CFG

Table 6. Quantitative Comparisons of the Ablation Experiments. Reported Results are Average RMSE on the Noise-free Middlebury Dataset for Scaling Factors 4, 8, and 16.

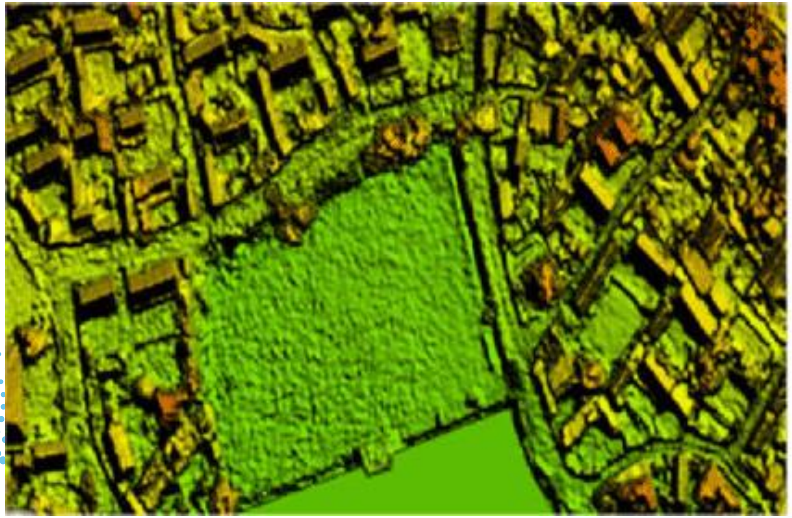
Design	Depth-Only			w/o shift			w/o CFG			w/o cross-attention			proposed		
Scale Factor	x4	x8	x16	x4	x8	x16	x4	x8	x16	x4	x8	x16	x4	x8	x16
RMSE	0.65	1.39	3.01	0.52	1.14	1.90	0.51	1.06	1.79	0.59	1.28	2.17	**0.48**	**0.99**	**1.55**

Future work

- **Apply to a real-world use case –**
  - aerial imagery SR in which a Raster (color) image and a Dynamic Elevation Model (DEM) are available.
  - DEMs are mostly low-resolution whereas Raster images are HR
  - Our objective would be to improve the DEM resolution using both the LR DEM and raster image as inputs.

Future work

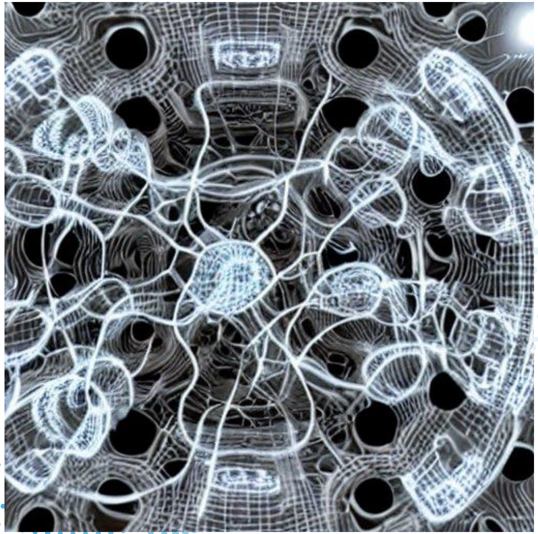
- **Apply to a real-world use case –**

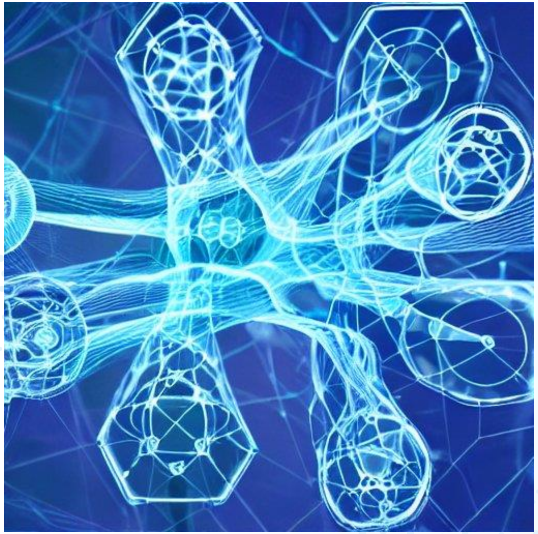


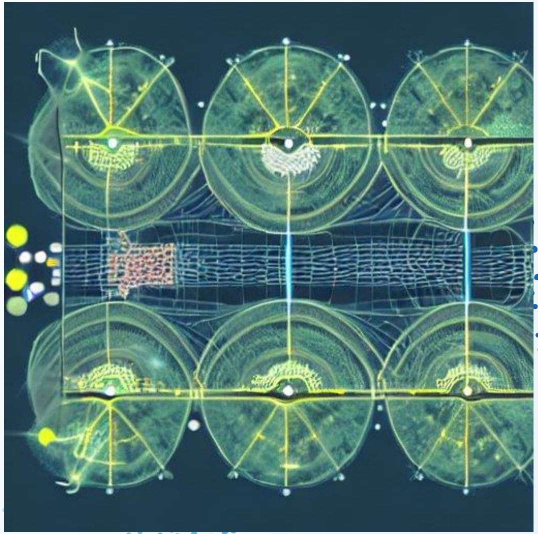
Discussion

- deep multi-modal networks for voice activity detection and depth images SR
- Transformer based architectures for guided SR
- Fusion\guidance attention & cross-attention mechanisms
- SOTA results in both tasks (at time of publication)

Questions?







**multimodal neural networks according to StableDiffusion..*