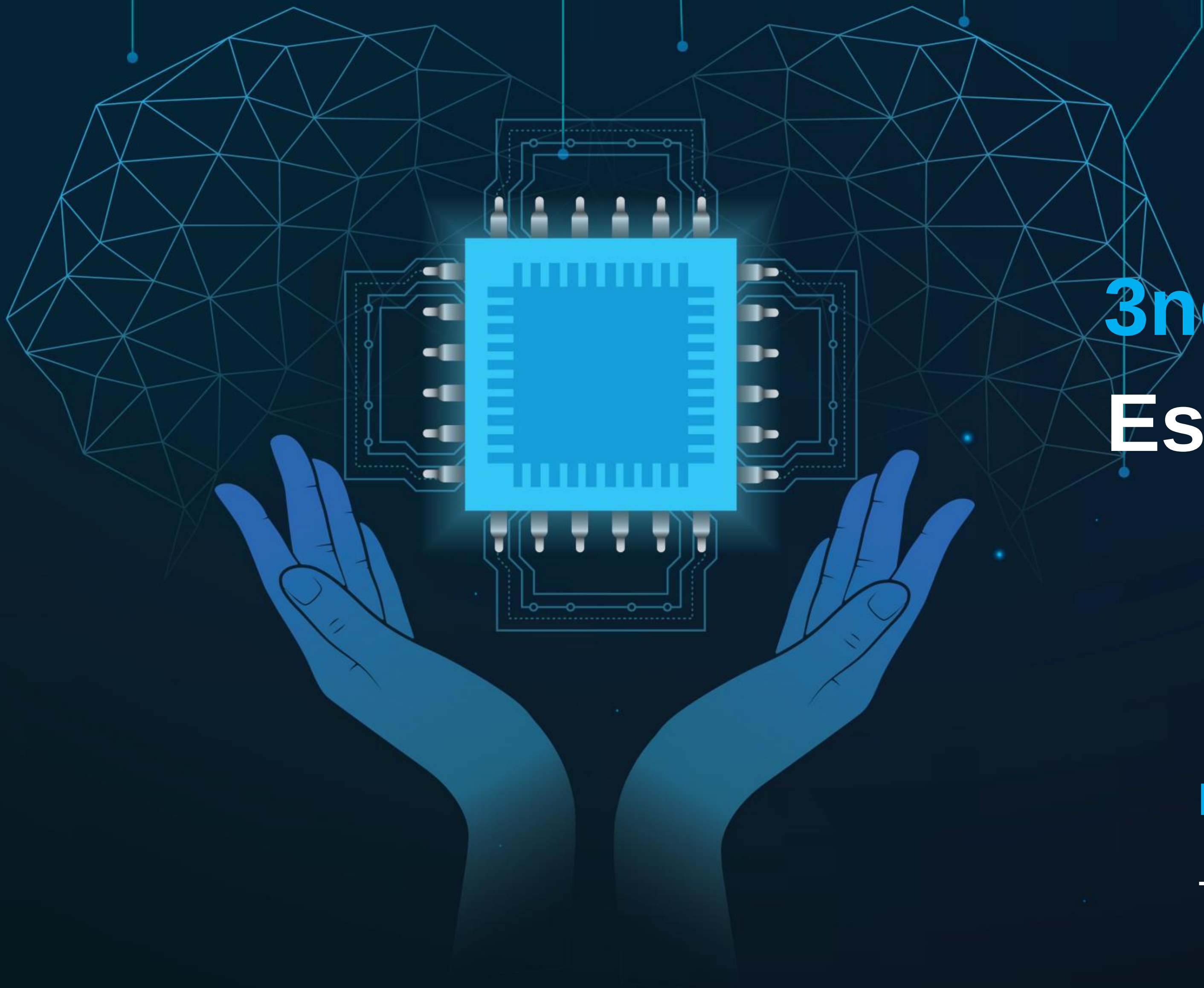


3nd Monocular Depth Estimation Challenge @ CVPR24

Mykola Lavreniuk

Team **EVP++**

Competition Journey:
Methods and Results

3rd Monocular Depth Estimation Challenge

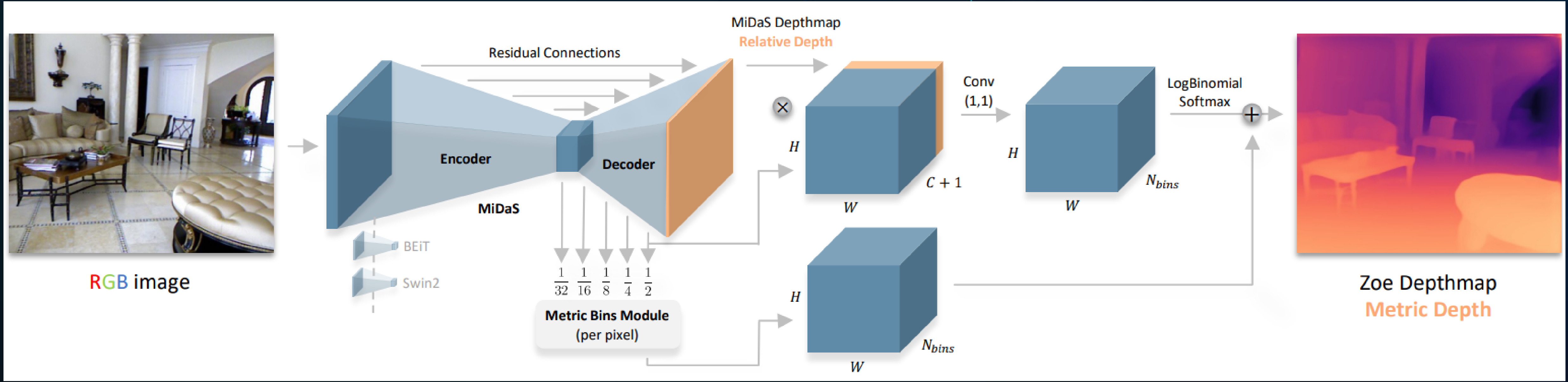
MDEC @ CVPR 2024

CVPR
SEATTLE, WA JUNE 17-21, 2024

Baseline: ZoeDepth

Zero-shot Transfer by Combining Relative and Metric Depth

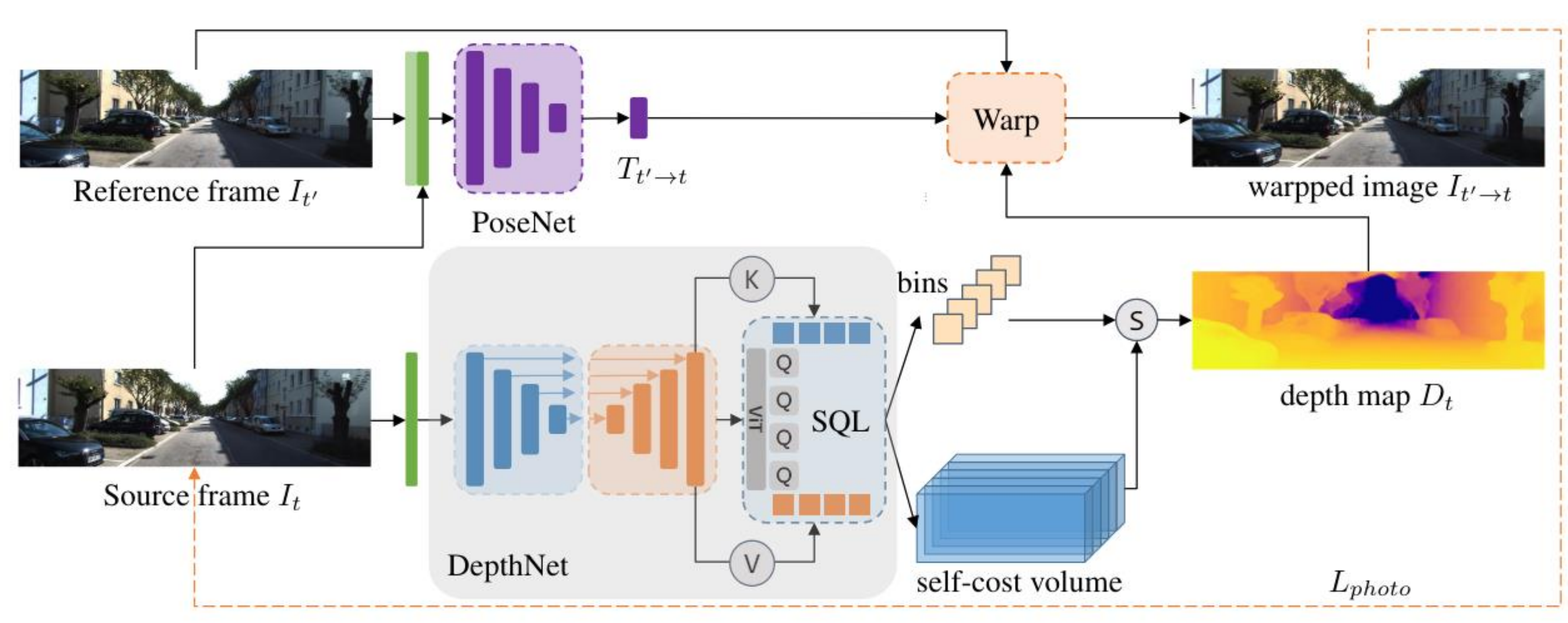
- ZoeDepth integrates relative and metric depth estimation through a two-stage process.
- First stage: Pre-trains an encoder-decoder on relative depth datasets.
- Second stage: Enhances the decoder with domain-specific heads based on new metric bins.



Exploring Alternative Models

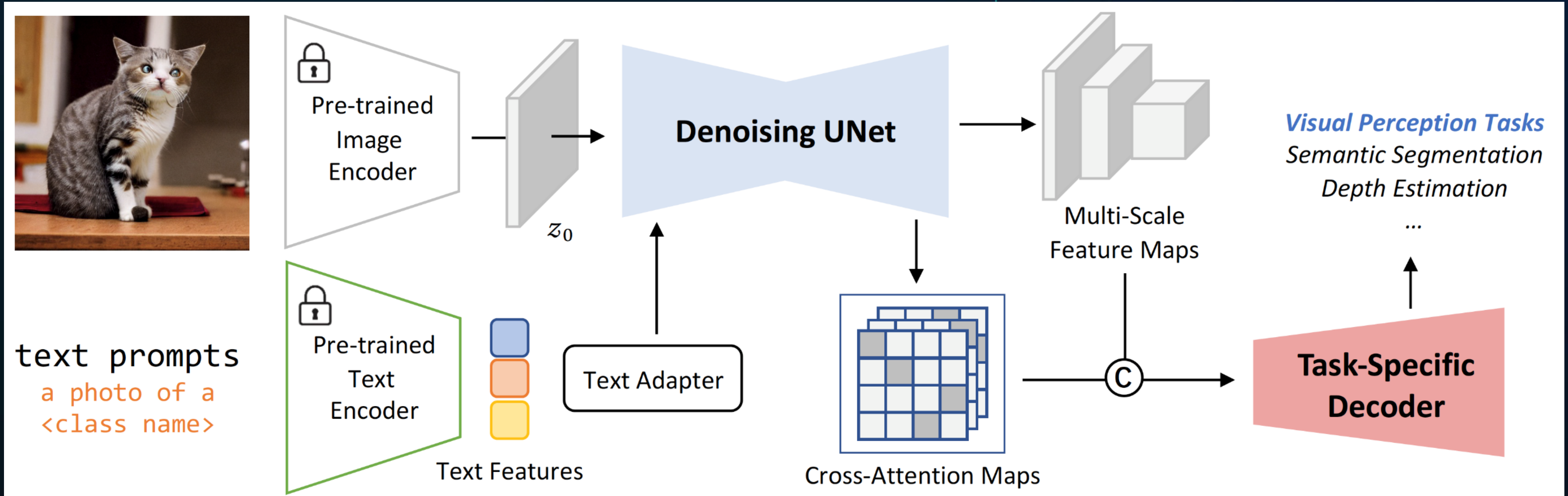
Self-Supervised Mono Depth Estimation: SQLdepth

Supervision comes from the consistency between the synthesis scene and source frame



Diffusion-based Mono Depth Estimation: VPD

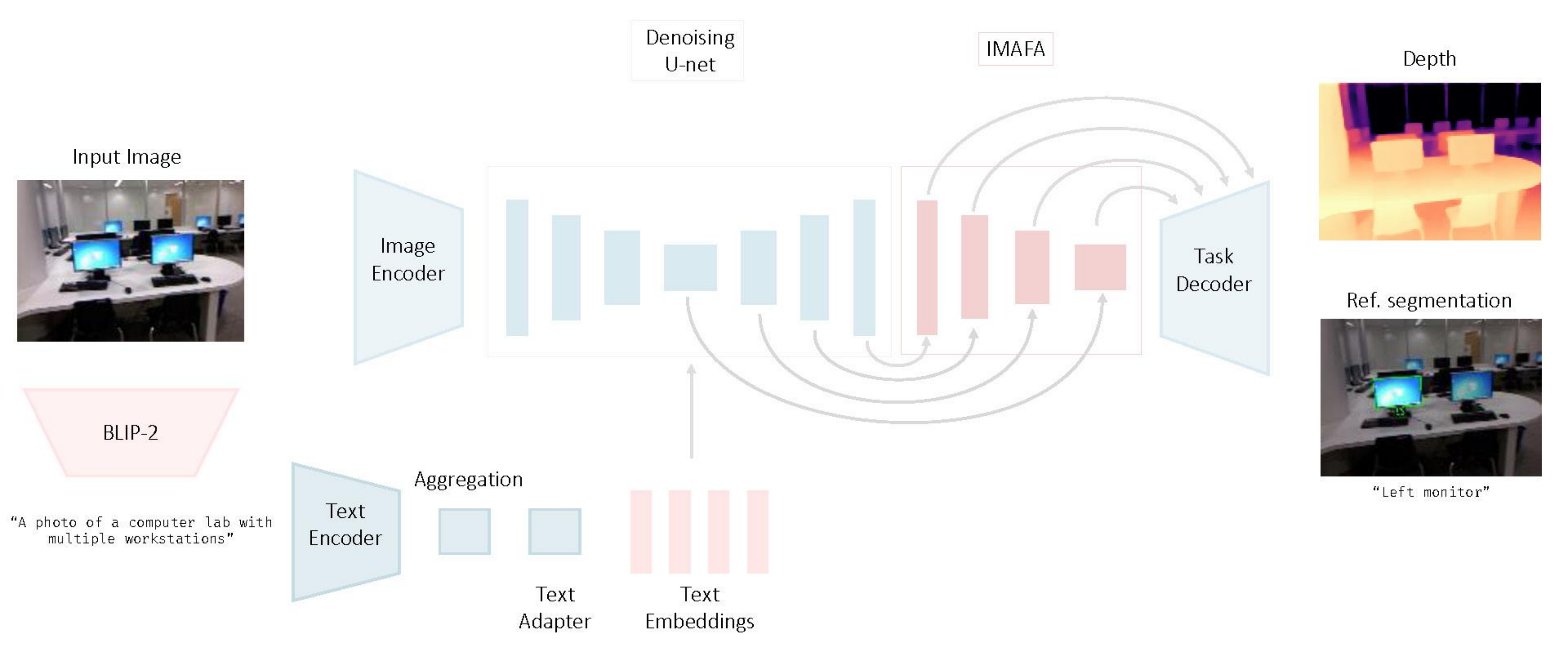
- Demonstrated that features learned by the denoising U-Net can be exploited for vision tasks.
- However, VPD could not be used for KITTI or SYNS due to relying on category names (templates).



Diffusion-based: EVP

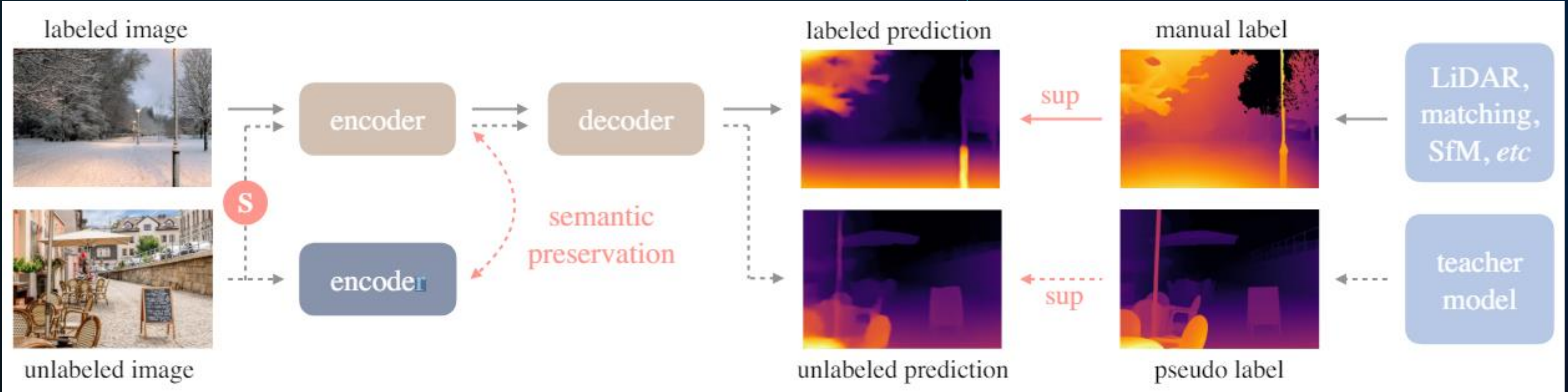
by Mykola Lavreniuk, Shariq Farooq Bhat, Matthias Müller, Peter Wonka

- ✓ Automatically generate free-form image captions using BLIP-2.
- ✓ Alternatively, use image embeddings (from CLIP) instead of text embeddings.
- ✓ A novel Inverse Multi-Attentive Feature Alignment module improves the accuracy.



Large-Scale Unlabeled Data: Depth Anything

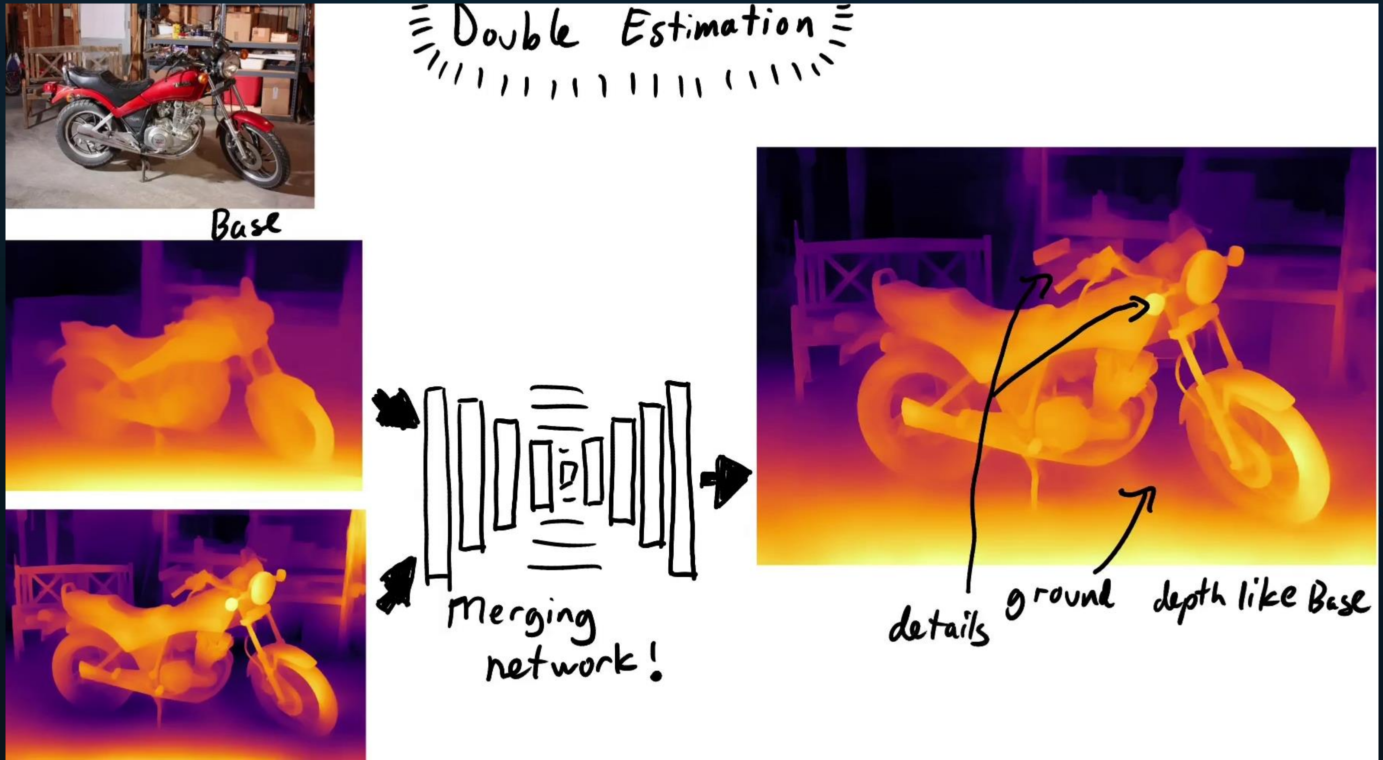
- DINOv2 pretrained VIT-L encoder.
- ZoeDepth metric bins module as a decoder.
- Pretrained on 1.5M labeled and 62M+ unlabeled images.

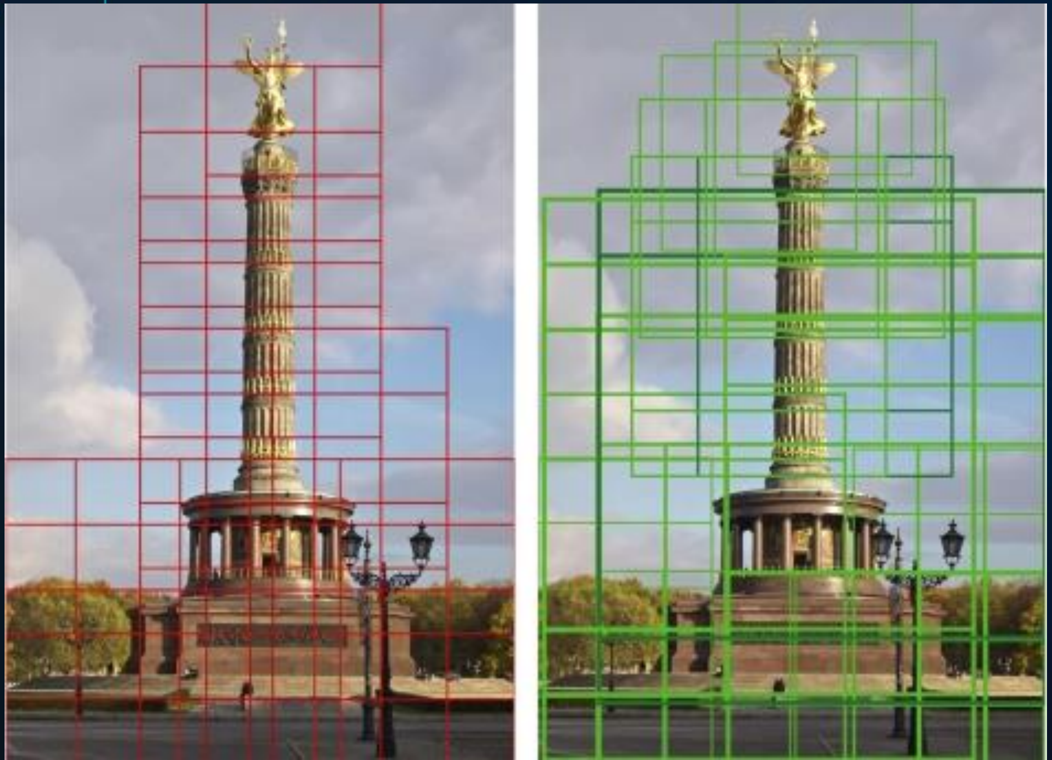


Approaches for Local Details Enhancement

BoostingDepth

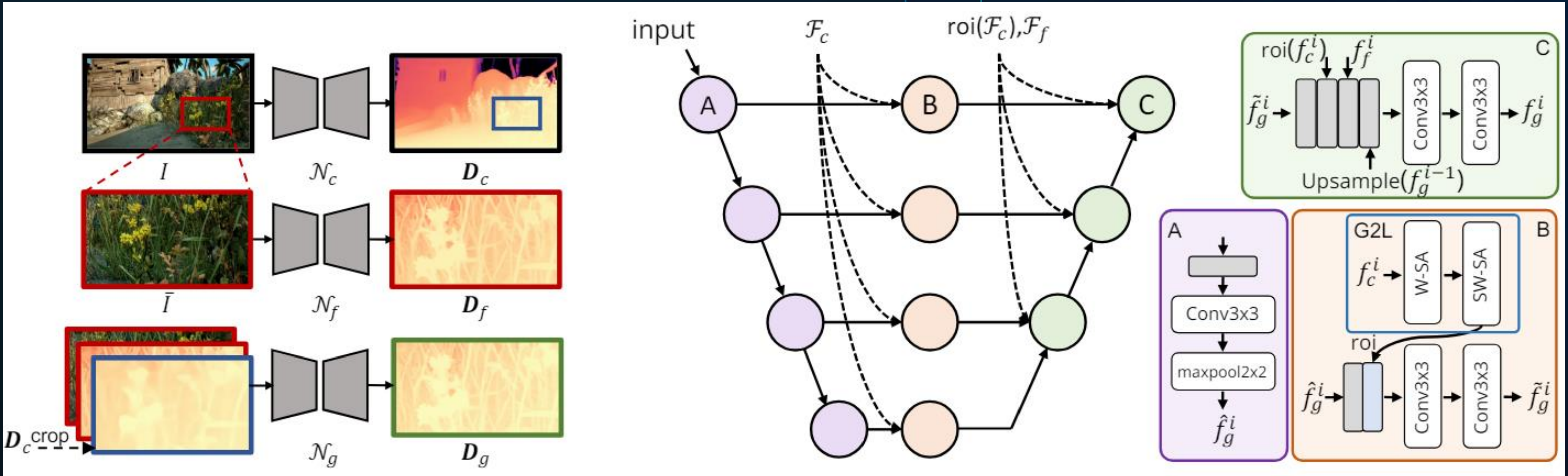
- Network provides depth for the whole image.
- Network inferred using a sliding window approach to preserve finer details.
- GAN combines the best of both.





PatchFusion

- A coarse network provides globally consistent depth.
- A fine network offers detailed refinements.
- A guided fusion network combines the best of both.



(a) Overall Pipeline

(b) Guided Fusion Network

Solution summary

Backbone: VIT-L encoder

Decoder: ZoeDepth metric bins module

Pretrain: Depth Anything (1.5M labeled and 62M+ unlabeled images)

Finetune: indoor (NYUv2) and outdoor (KITTI, Virtual KITTI 2, DIODE)

Loss: SILog loss

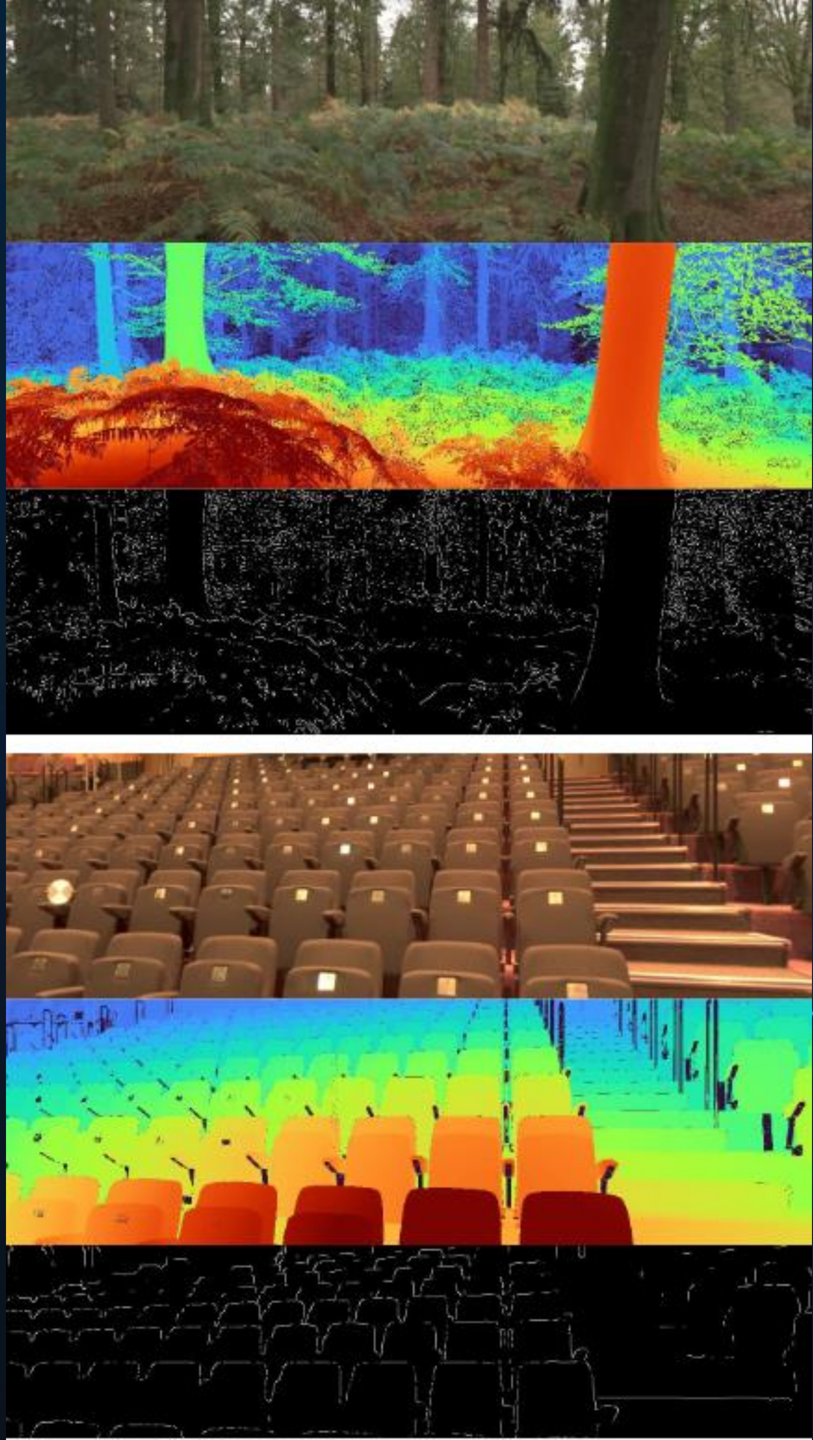
Augmentations: standard from Depth Anything

Training:

indoor - lr = 0.000161, epochs = 5, batch_size = 16, max_d = 10
outdoor - lr = 0.000020, epochs = 5, batch_size = 1, max_d = 80

Other tricks:

- ✓ Sliding window.
- ✓ Turn off TTA.
- ✓ Turn off padding of the input image.



What did not work for me

Models: models with higher accuracy on KITTI

Train and test images:

larger image resolution,

longer training

Augmentations: random crop, CutFlip (URCDC-Depth paper)

Other tricks:

- Dedicated models for edge or local details enhancement.
- TTA or model soup improve RMSE, AbsRel, but decrease F1.

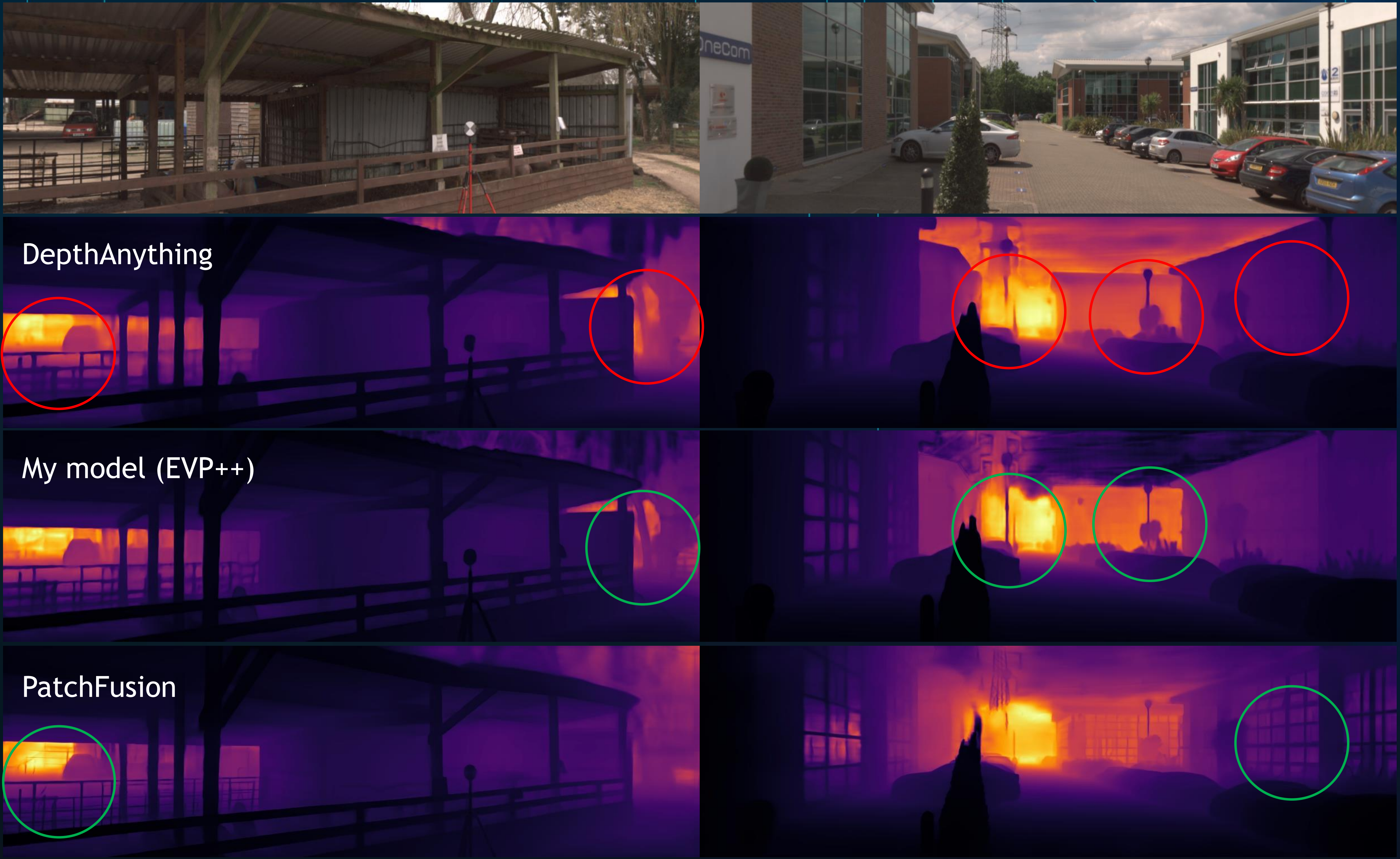
Quantitative results

#	User	Team Name	Total Rank↓	F-Score↑	F-Score (Edges)↑	MAE↓	RMSE↓	AbsRel↓	Edge Accuracy↓	Edge Completion↓
1	**lavreniuk**	**EVP++**	**18**	**20.8658 (4)**	**10.9224 (3)**	**3.7086 (1)**	**6.5308 (1)**	**19.0214 (1)**	**2.8835 (3)**	**6.7700 (5)**
2	zhouguangyuan	PICO-MR	31	23.7210 (1)	11.0130 (2)	3.7787 (2)	6.6115 (2)	21.2386 (4)	3.8974 (17)	4.4517 (3)
3	moyushin		38	23.2499 (2)	10.7783 (4)	3.8725 (3)	6.7038 (3)	21.6987 (5)	3.5911 (14)	9.8569 (7)
4	inso-13		41	18.5956 (8)	9.4273 (9)	3.9224 (4)	7.1649 (4)	20.1185 (2)	2.8921 (4)	15.6460 (10)
5	pihai		51	17.8328 (9)	9.1387 (12)	4.1147 (5)	7.7332 (6)	21.2310 (3)	2.9465 (5)	17.8145 (11)
6	Depth_3DV		57	20.4244 (6)	10.1868 (5)	4.4079 (9)	7.8909 (8)	23.9416 (9)	3.6105 (16)	5.7953 (4)
7	surajiitd	visioniitd	57	19.0651 (7)	9.9200 (7)	4.5318 (10)	7.9626 (9)	23.2745 (7)	3.2596 (11)	7.9953 (6)
8	erdosv001		58	20.7673 (5)	9.9617 (6)	4.3302 (8)	7.8348 (7)	27.7973 (11)	3.4458 (13)	13.2527 (8)
9	jsk24	RGA Inc.	65	22.7924 (3)	11.5192 (1)	5.2061 (13)	9.2339 (13)	28.8613 (12)	4.1541 (21)	0.8980 (2)
10	weijianing		67	17.8121 (10)	9.7525 (8)	5.0386 (12)	8.9196 (11)	24.0101 (10)	3.1615 (7)	14.1550 (9)
11	luo0207		67	16.9120 (12)	9.0666 (14)	4.1357 (6)	7.3481 (5)	22.0509 (6)	3.2377 (10)	18.5220 (14)
12	qing		73	17.5704 (11)	9.1273 (13)	4.2759 (7)	8.3635 (10)	23.3455 (8)	3.1765 (8)	20.6621 (16)
13	hitcslj	HIT-AIIA	89	16.7148 (13)	9.2525 (10)	5.4767 (15)	11.0510 (19)	34.2035 (19)	2.5703 (1)	18.0436 (12)
14	dagouqin		95	16.4478 (14)	8.8896 (15)	5.2907 (14)	10.5310 (17)	33.6741 (18)	2.5965 (2)	18.7283 (15)
15	al	ReadingLS	98	14.8093 (16)	8.1357 (16)	5.0099 (11)	8.9448 (12)	29.3938 (13)	3.2837 (12)	30.2778 (18)
16	hyc123		107	15.9223 (15)	9.1679 (11)	8.2542 (20)	13.8783 (20)	43.8823 (20)	4.1054 (20)	0.7403 (1)
17	yogurts		110	13.7089 (18)	7.5505 (19)	5.4867 (16)	9.4419 (14)	30.7377 (15)	3.6072 (15)	18.3600 (13)
18	SmartHust		112	11.8998 (19)	8.0770 (17)	6.3256 (19)	10.8861 (18)	30.4607 (14)	2.9862 (6)	33.6279 (19)
19	mdec		118	13.7211 (17)	7.7630 (18)	5.5645 (17)	9.7169 (15)	32.0420 (16)	3.9712 (18)	21.6256 (17)
20	journey2japan		132	11.3561 (20)	6.6000 (21)	5.9075 (18)	9.9886 (16)	33.4098 (17)	3.9832 (19)	54.6467 (21)
21	smhh		133	11.0444 (21)	7.0866 (20)	8.7645 (21)	15.8637 (21)	63.3160 (21)	3.2209 (9)	40.6098 (20)

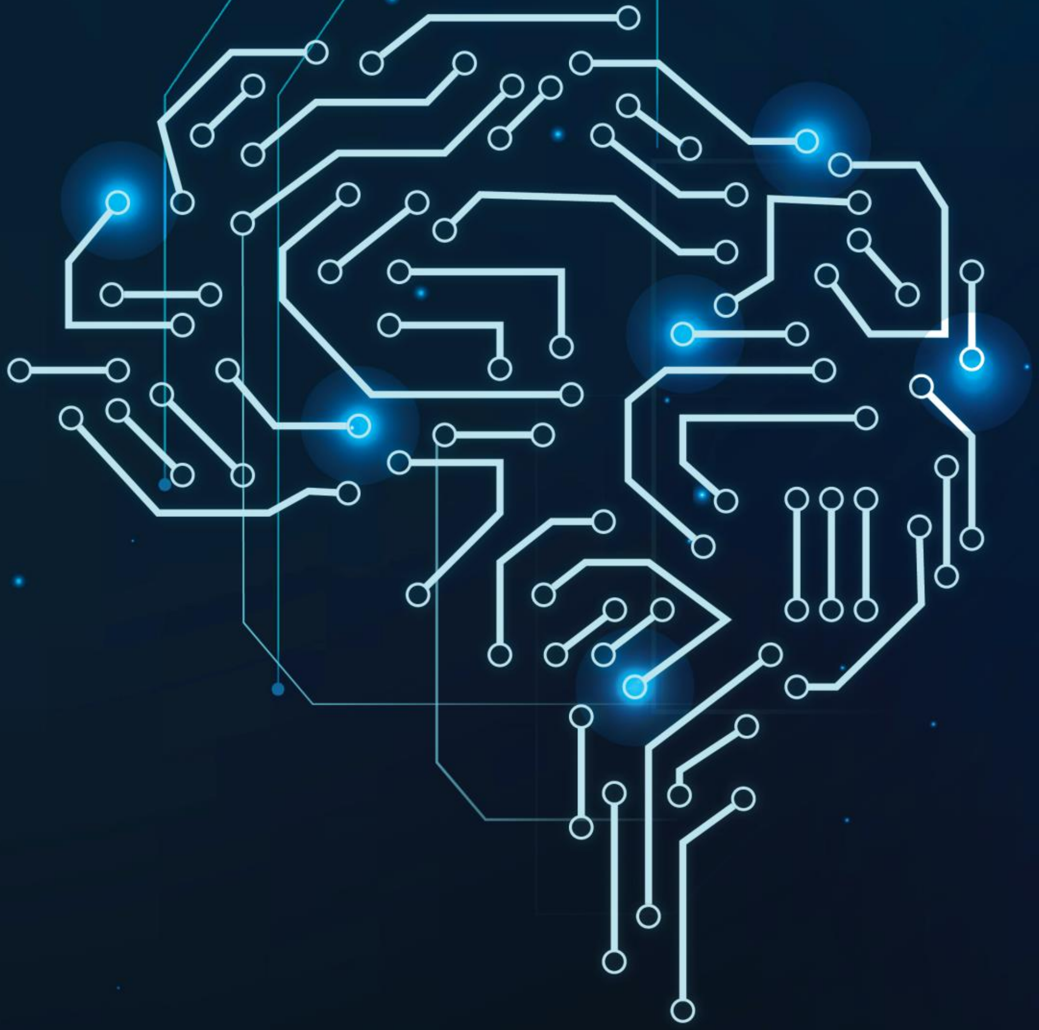
Green - top 1 score, yellow - top 5 scores, total rank - the sum of ranks across all metrics.

Only EVP++ gets more than 1 top rank, including first place in 3 traditional metrics and the best rank sum.

Qualitative results


DepthAnything
My model (EVP++)
PatchFusion

Thank you!



Link to the code