NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

360 Vision in the Foundation AI Era: Principles, Methods, and Future Directions

Asst Prof Addison, Wang Lin

BioRAI Lab, School of Electrical and Electronic Engineering, NTU-Singapore

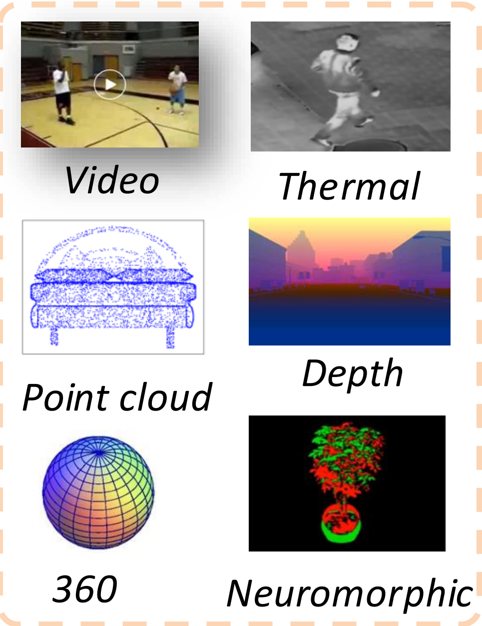
linwang@ntu.edu.sg

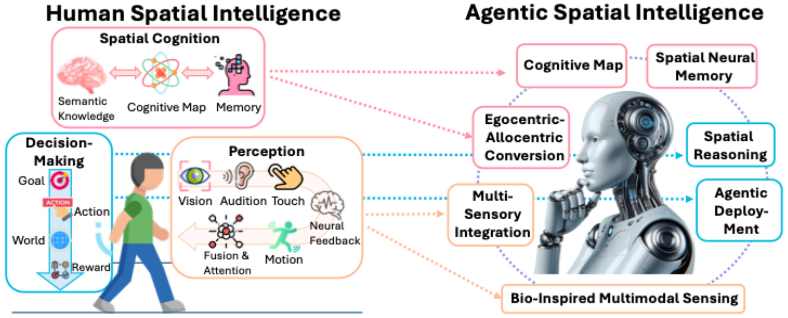
https://dr.ntu.edu.sg/cris/rp/rp02550

Overview of Our Research

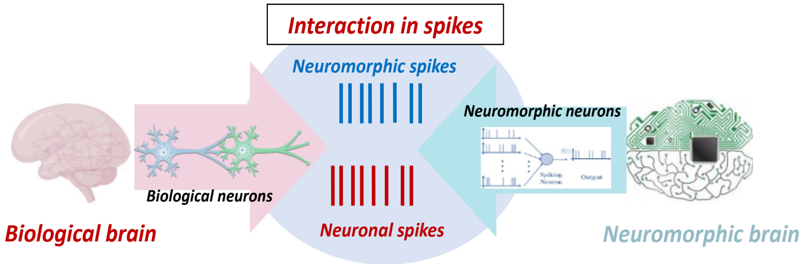
Sensor Fusion

Visual Sensing

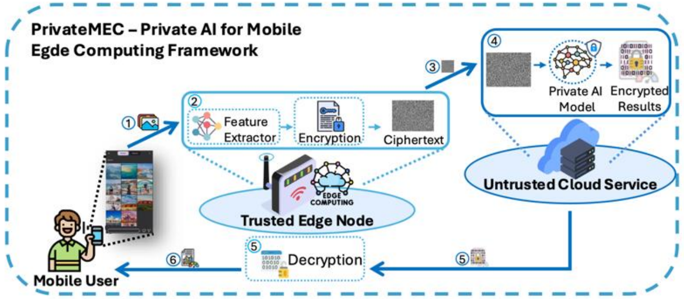




Neuro-inspired AI

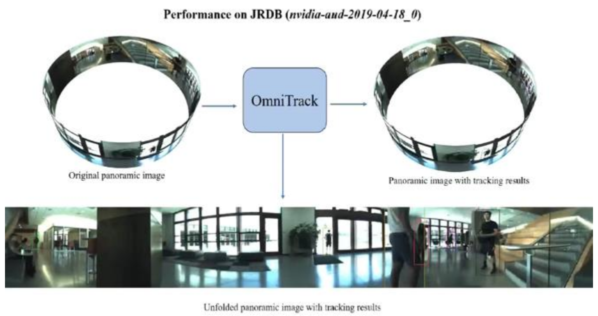


Brain-inspired computing

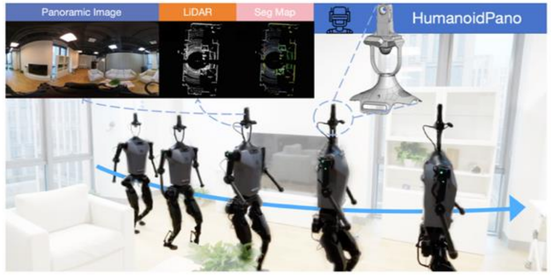


Edge AI

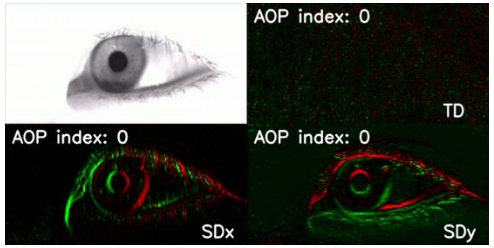
Applications to Embodied AI



Quadruple scene navigation



Sensor design for humanoid



Sensation/Emotion for Robot

360 Cameras

(a) ONE X2

(b) Titan

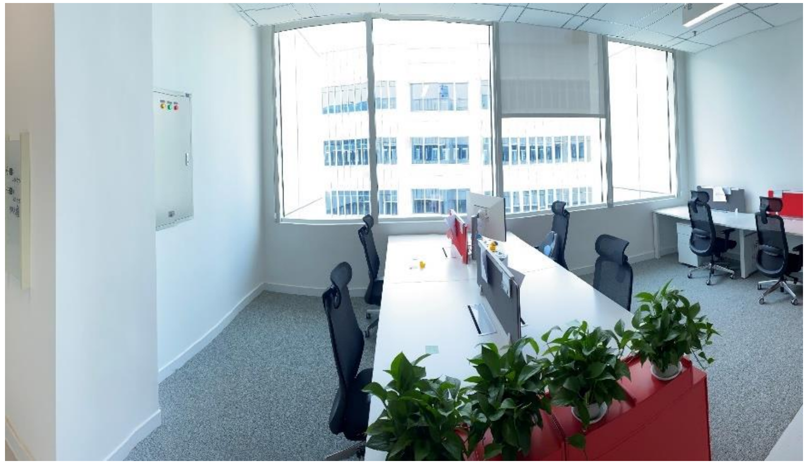
(c) 360 CAM

(d) GoPro Omni

AI et al. "A survey of representation learning, optimization strategies, and applications for omnidirectional vision."IJCV, 2025.

What’s different for 360 cameras?

Perspective Image

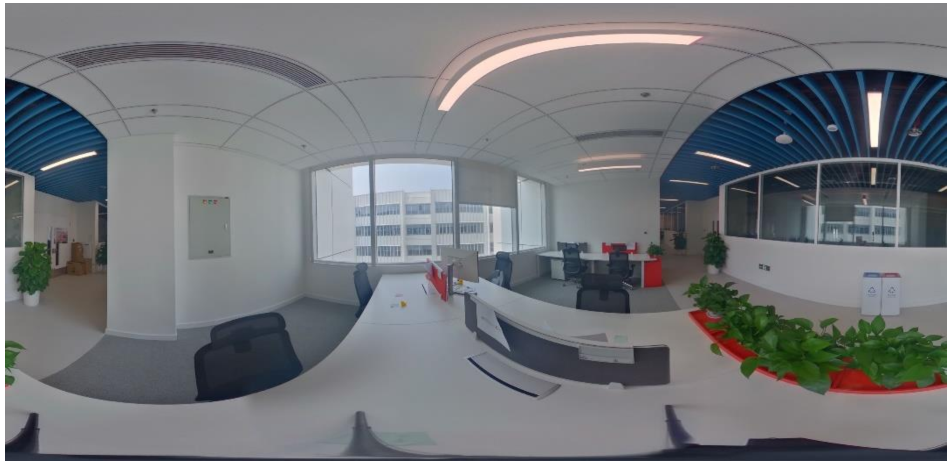


Captured with the iPhone 11

Some drawbacks:

- ✓ **Limited Field-of-view**
- ✓ **Weak to capture the 3D information**

360 Image
Equirectangular projection (ERP)



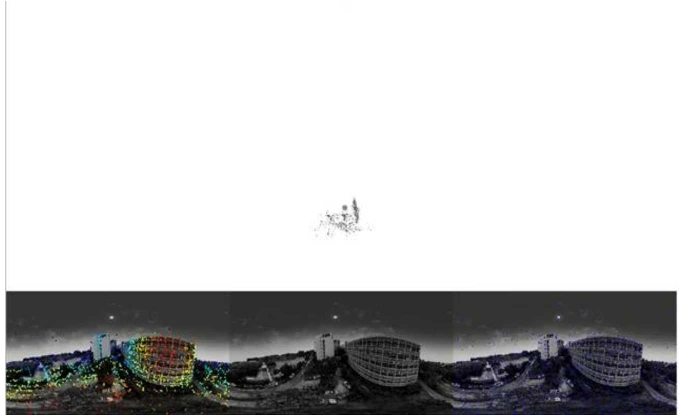
Captured with the THETA 360 camera

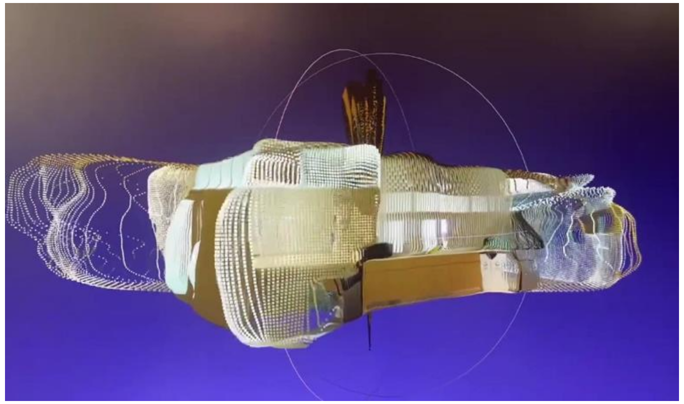
Some advantages:

- ✓ **Large Field-of-view (360x180)**
- ✓ **Strong immersivity and more realism**

AI et al. "A survey of representation learning, optimization strategies, and applications for omnidirectional vision."IJCV, 2025.

Applications of 360 Cameras





VO & 3D Reconstruction

Robotics and self-driving

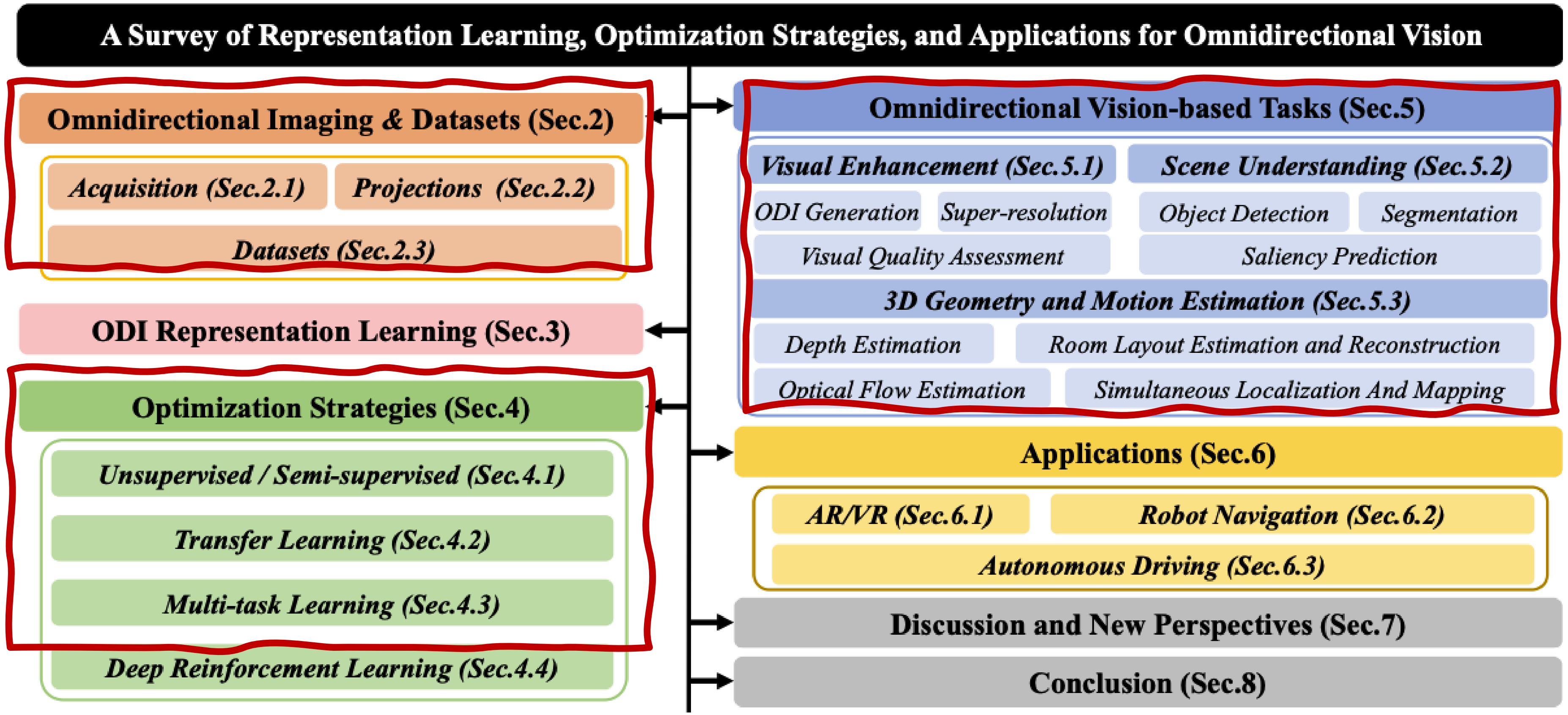
VR & Eye-tracking

Huang et al, 360VO: Visual Odometry Using A Single 360 Camera, RAL, 2022.
AI et al. HRDFuse: Monocular 360°Depth Estimation by Collaboratively Learning Holistic-with-Regional Depth Distributions, CVPR, 2023.
Lee et al. "SpherePHD: Applying CNNs on a Spherical PolyHeDron Representation of 360 Images, CVPR, 2019.

Key Focus of This Talk

A Survey of Representation Learning, Optimization Strategies, and Applications for Omnidirectional Vision

Hao Ai [1] · Zidong Cao [1] · Lin Wang [1,2,✉]



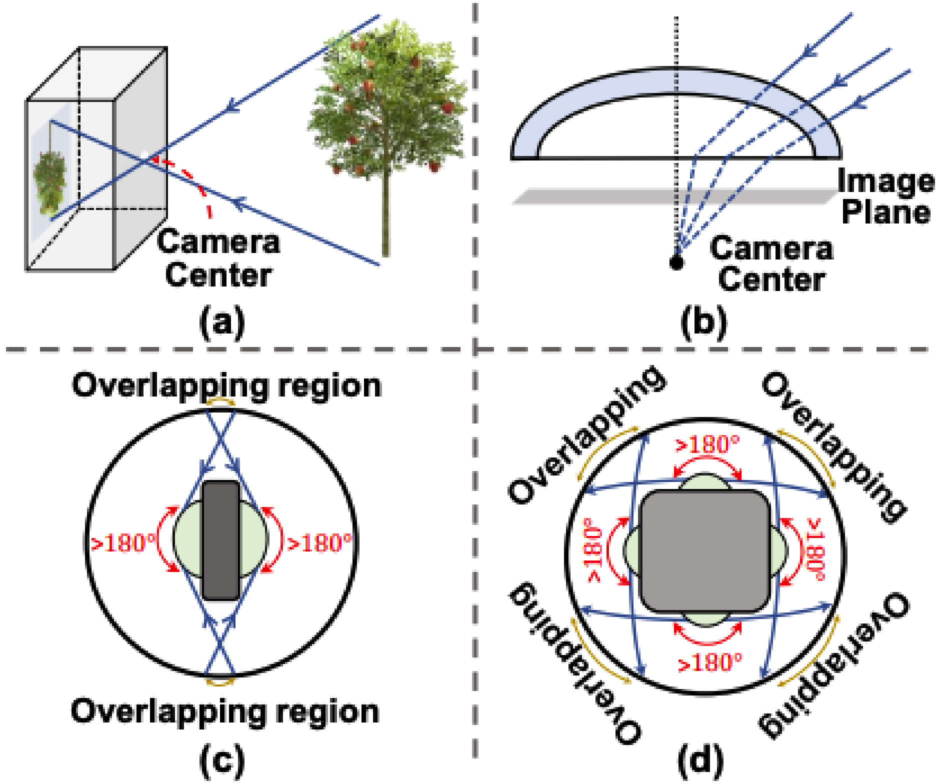
AI et al. "A survey of representation learning, optimization strategies, and applications for omnidirectional vision."IJCV, 2025.

Table of Contents

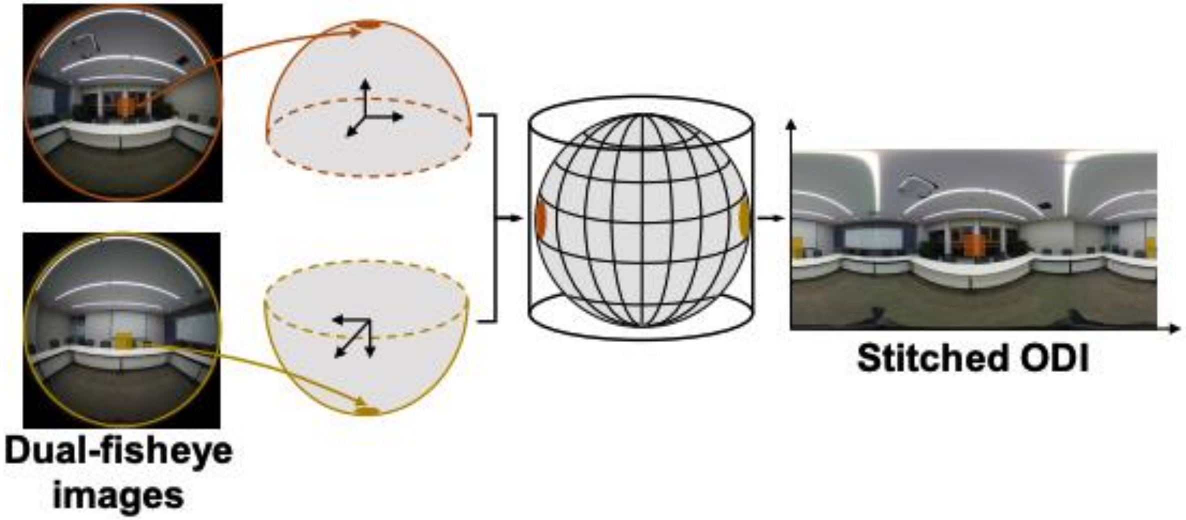
We are here now!

Imaging principles of several cameras

Assigning light from the camera’s surrounding to a specific data structurea



(a) Pinhole camera; (b) Fisheye camera;
(b) 360◦ camera (dual-fisheye);
(c) 360◦ camera (multi-fisheye)



Stitching a pair of dual-fisheye images into an ERP image

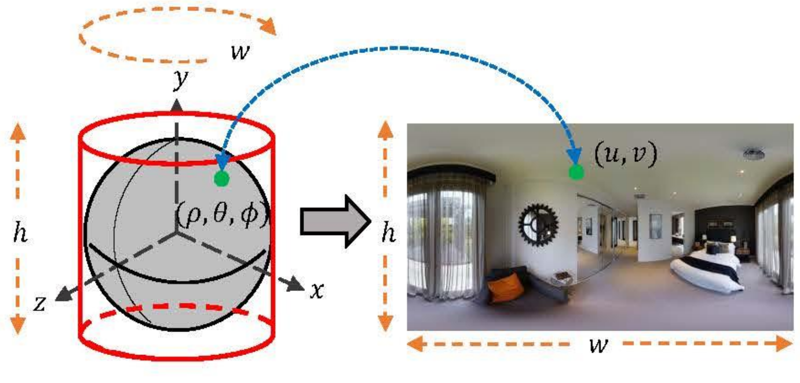
$$\begin{vmatrix}\rho\\ \theta\\ \phi\end{vmatrix} = \begin{vmatrix}(x^2+y^2+z^2)^{1/2}\\ \arctan(x/z)\\ \arccos(y/\rho)\end{vmatrix}, \begin{vmatrix}x\\ y\\ z\end{vmatrix} = \begin{vmatrix}\rho\sin(\theta)\sin(\phi)\\ \rho\cos(\phi)\\ \rho\cos(\theta)\sin(\phi)\end{vmatrix}.$$

Spherical Projection

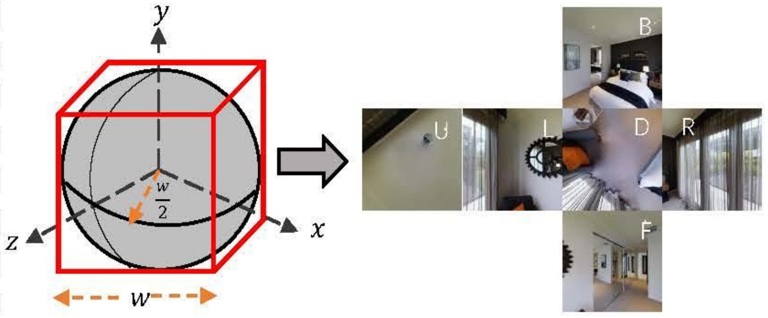
https://www.flir.com/discover/cores-components/can-thermal-imaging-see-through-walls/
AI et al. "A survey of representation learning, optimization strategies, and applications for omnidirectional vision."IJCV, 2025.

Representation of 360 images

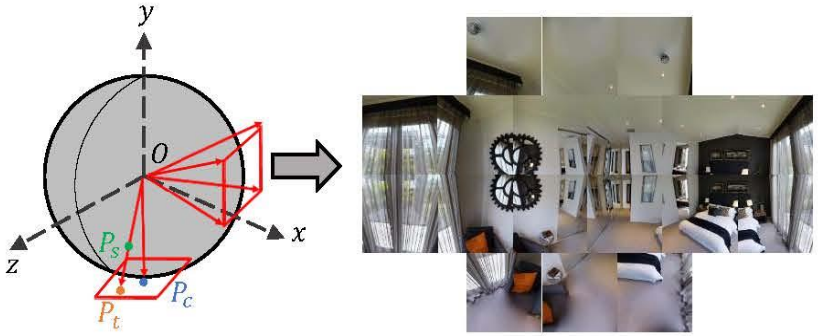
Due to spherical imaging, 360° Image owns multiple projection formats



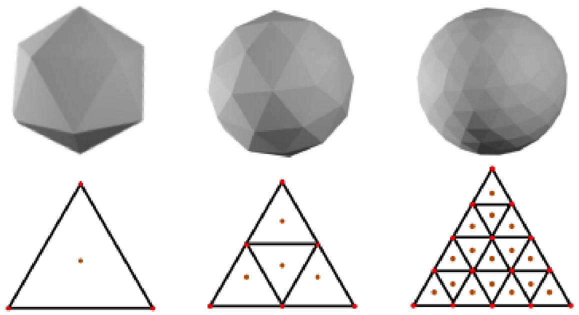
Equirectangular Projection (ERP)



Cubemap Projection (CP)

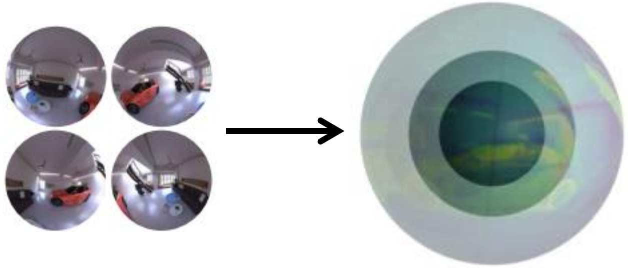


Tangent Projection (TP)



level=0 level=1 level=2

Lcosahedron (ICOSAP)



Cube MSI (IEEE RAL, 2025)

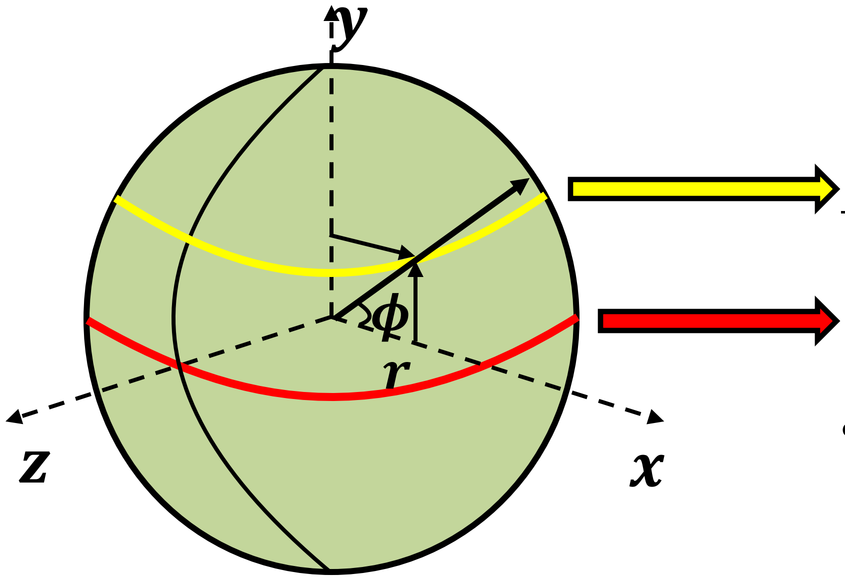
Vertical/horizontal Slicing

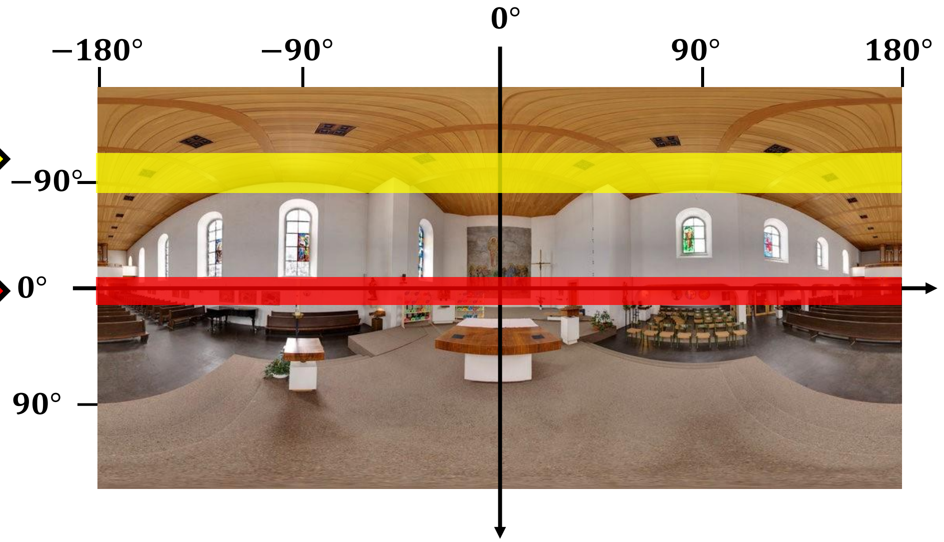
AI et al. "A survey of representation learning, optimization strategies, and applications for omnidirectional vision."IJCV, 2025.

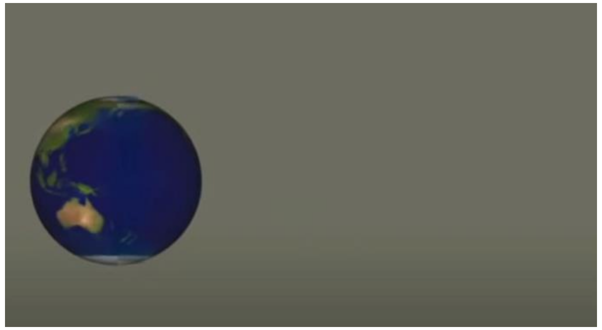
Representation of 360 images: ERP

ERP is the most common representation

- ERP images contains severe spherical distortions, especially at two poles.
- Normal 2D-based convolution filters can not handle distortion problem.



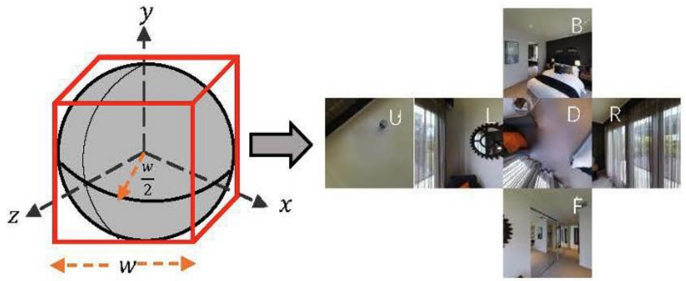




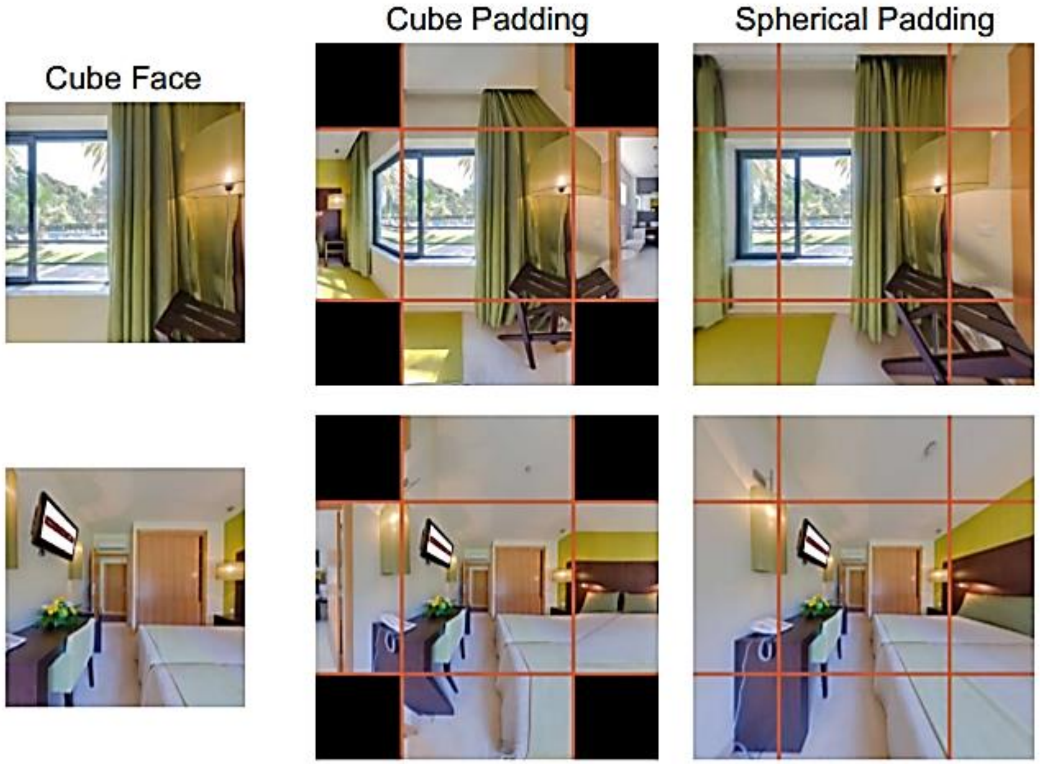
AI et al. "A survey of representation learning, optimization strategies, and applications for omnidirectional vision." IJCV, 2025.

Representation of 360 images: Cubemap Projection

CP padding is needed



Spherical padding is important!



Cube padding

- Cube padding directly pads the feature of the connected faces.
- The values of four corners are undefined.

Spherical padding

- The padding area is calculated with spherical projection.
- Both the missing corner and inconsistency at the boundary can be addressed.

Wang et al. BiFuse: Monocular 360◦ Depth Estimation via Bi-Projection Fusion, CVPR, 2020.

Representation of 360 images: Tangent Projection

A set of local planar image grids tangent to the subdivided icosahedron

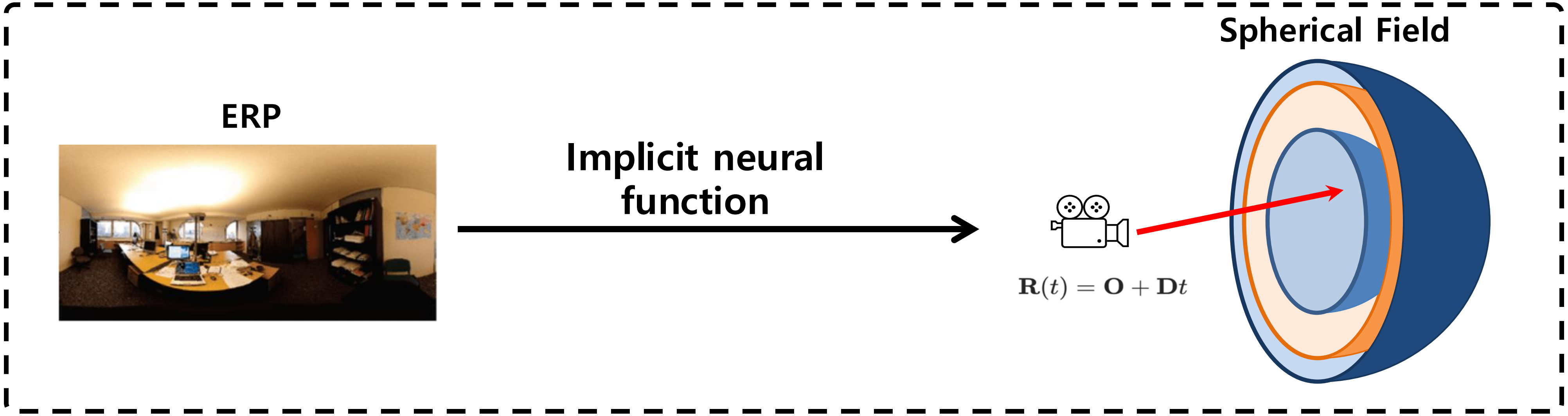
ERP image

TP patches (N=10)

TP patches (N=18)

Representation of 360 images

Multi-spherical image (MSI) representation



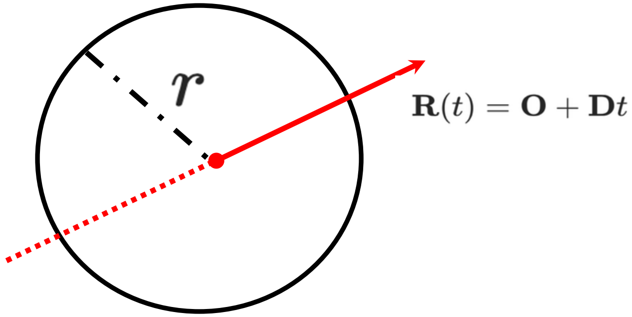
Ray-sphere sampling from the constructed spherical field:

$$a = 1, b =< \mathbf{O}, \mathbf{D} >, c =< \mathbf{O}, \mathbf{O} > -r^2$$

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

Ray

$$\mathbf{R}(x) = \mathbf{O} + x\mathbf{D}$$

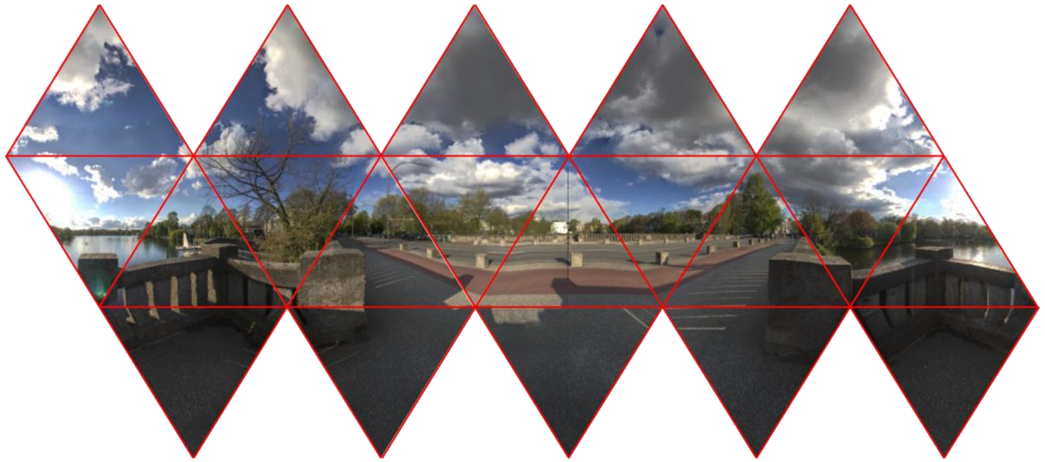


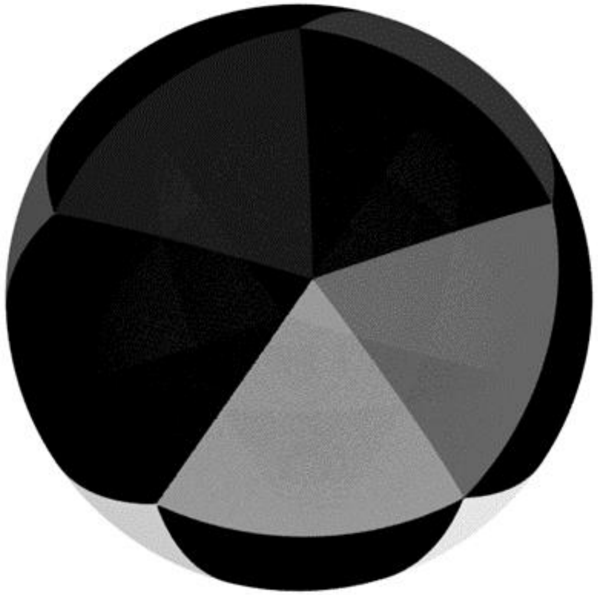
Change et al, CUBE360: Learning Cubic Field Representation for Monocular 360 Depth Estimation, RAL, 2025.

Representation of 360 images

Icosahedron-based 360° Image Representation

- Regular **20-sided** polyhedron-based representation.
- A set of **ray vectors** equal to the number of pixels.





Advantage: Much less irregularity

Lee et al. "SpherePHD: Applying CNNs on a Spherical PolyHeDron Representation of 360 Images, CVPR, 2019.

Takeaways

Which representation is good?

No one, not even one!

Project fusion might be a good solution!

(HRDFuse, CVPR2023; Elite360D, CVPR 2024; CUBE360, IEEE RAL 2025)

Table of Contents

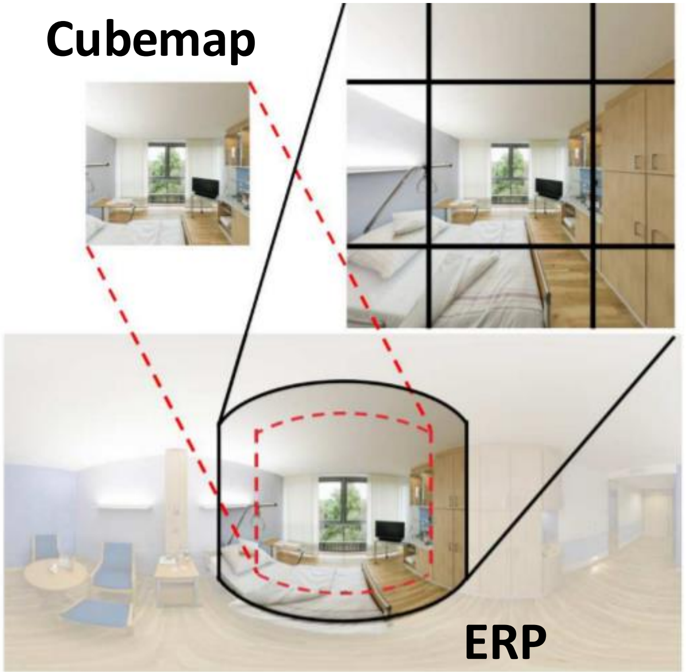
We are here now!

Project fusion is a crucial to learn holistic-to-local semantic & geometric info from 360 data!

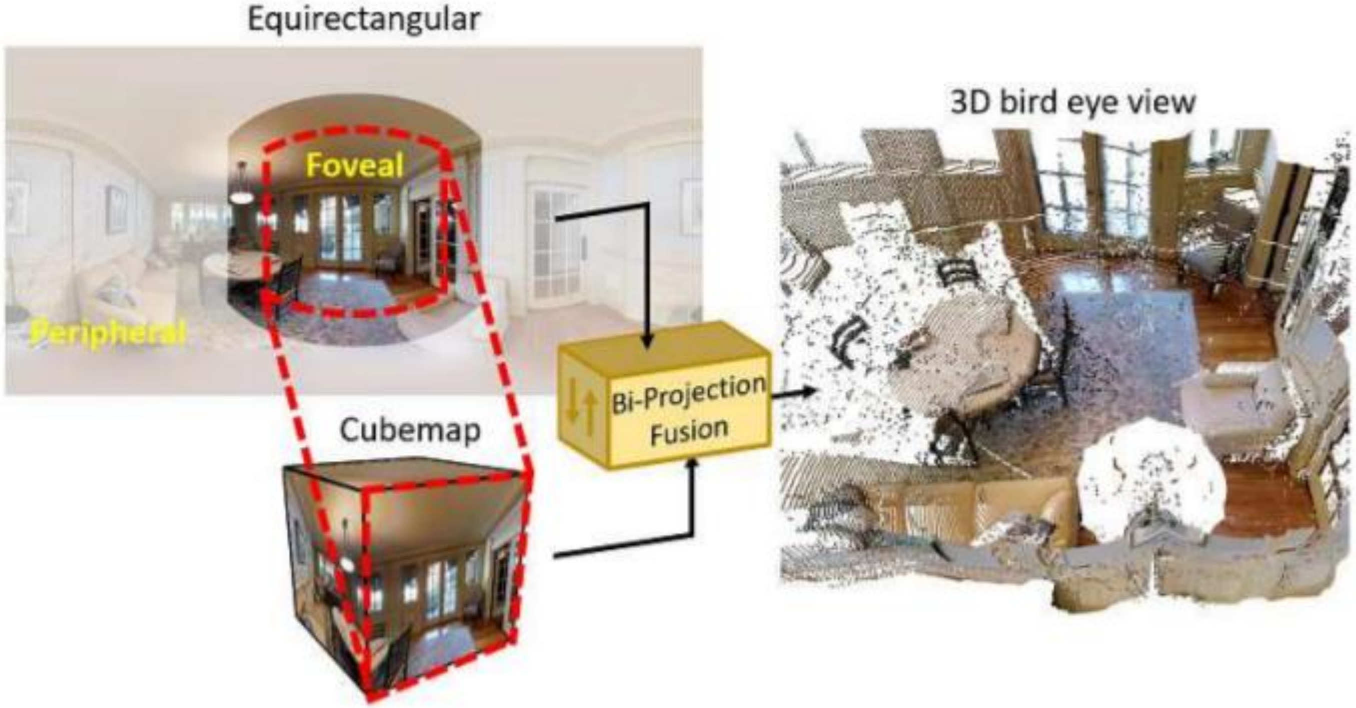
ERP + CP Fusion for Depth Estimation

Is it possible to fuse both representations of 360 images?

- Mimicking both **peripheral and foveal vision** of the human eye



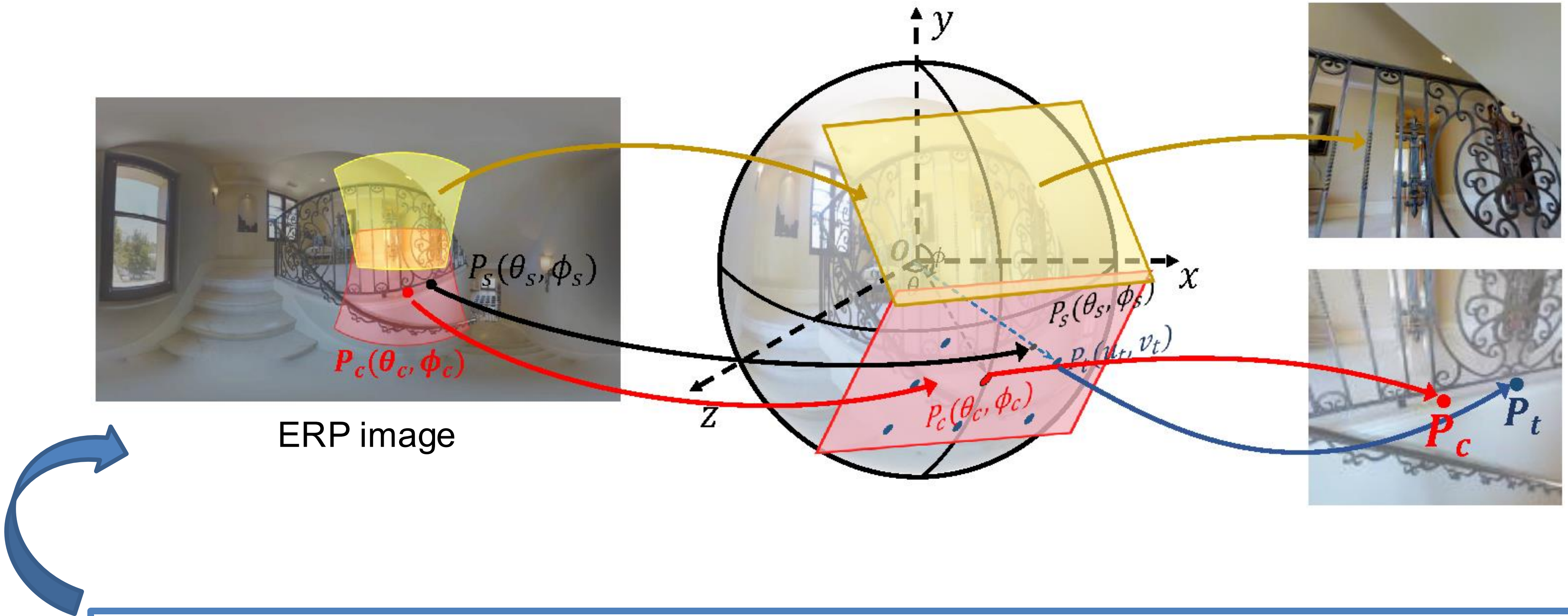
ERP has the **largest** FoV compared to each face on the cubemap projection



- ERP to CP transform (E2C)
- CP to ERP transform (C2E)

Wang et al. BiFuse: Monocular 360◦ Depth Estimation via Bi-Projection Fusion, CVPR, 2020.

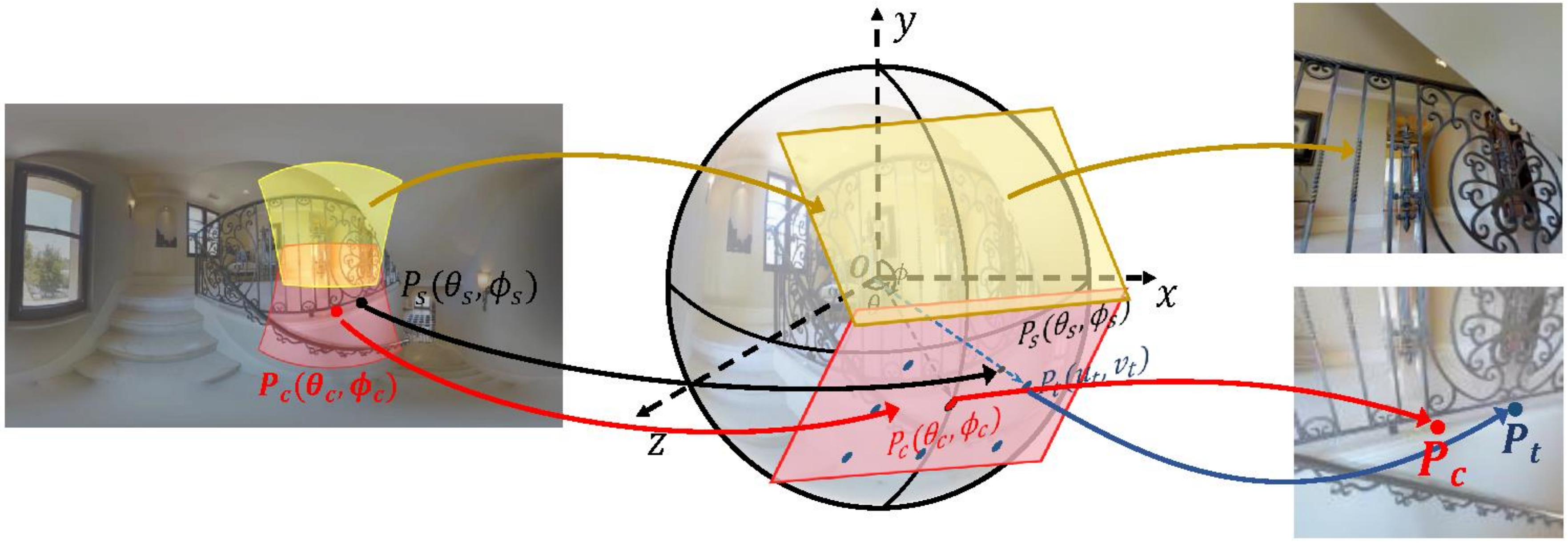
ERP+ TP Fusion for Depth Estimation



ERP can provide holistic contextual information, but it is distorted.

AI et al. HRDFuse: Monocular 360°Depth Estimation by Collaboratively Learning Holistic-with-Regional Depth Distributions, CVPR, 2023.

ERP+ TP Fusion for Depth Estimation

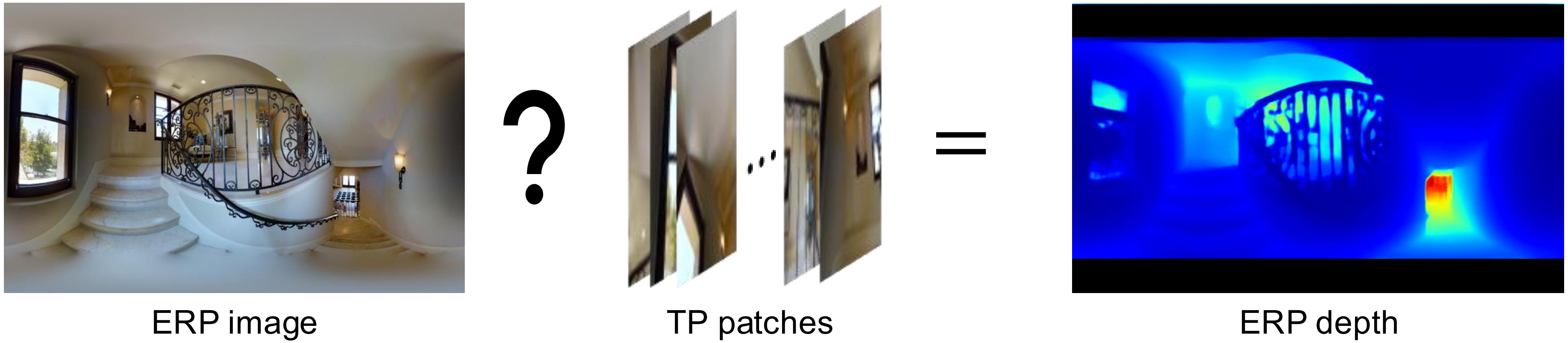


TP patches

TP patches are less distorted but exist unavoidable overlapping areas between two neighboring TP patches.

AI et al. HRDFuse: Monocular 360°Depth Estimation by Collaboratively Learning Holistic-with-Regional Depth Distributions, CVPR, 2023.

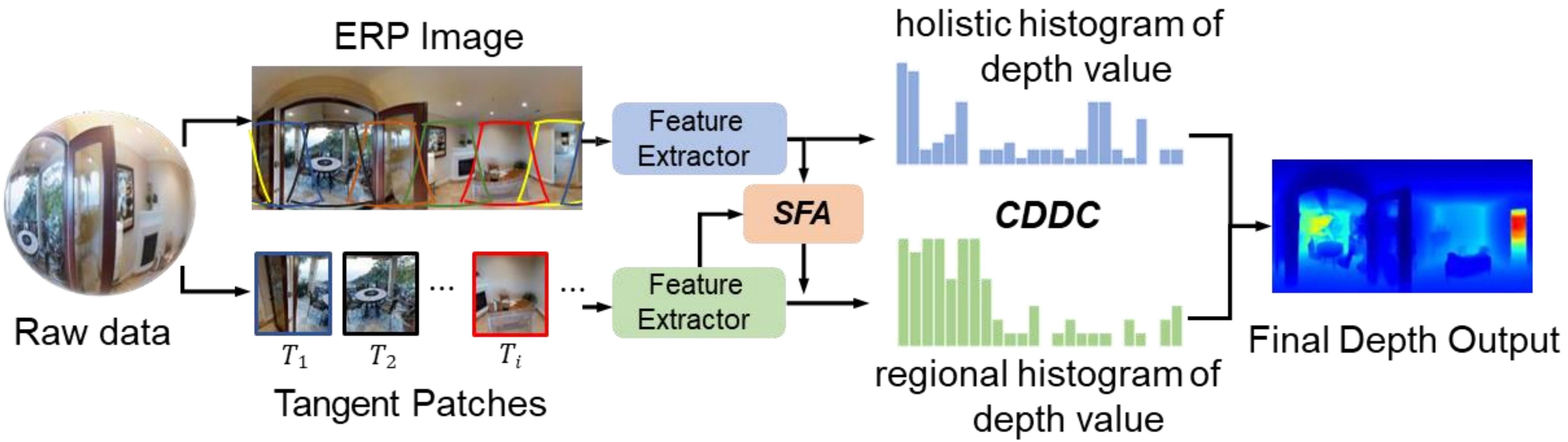
ERP+ TP Fusion for Depth Estimation



How to better employ the holistic contextual Information in the distorted ERP images and regional structural Information in the less-distorted TP patches?

AI et al. HRDFuse: Monocular 360°Depth Estimation by Collaboratively Learning Holistic-with-Regional Depth Distributions, CVPR, 2023.

ERP+ TP Fusion for Depth Estimation



- **SFA** learns the similarities between the TP features and ERP features .
- **CDDC** learns the holistic and regional depth distribution histograms.

AI et al. HRDFuse: Monocular 360°Depth Estimation by Collaboratively Learning Holistic-with-Regional Depth Distributions, CVPR, 2023.

ERP+ TP Fusion for Depth Estimation

Datasets	Method	Patch size/FoV	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE(log) ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Stanford2D3D	FCRN [28]	–/–	0.1837	-	0.5774	-	0.7230	0.9207	0.9731
	BiFuse with fusion [41]	–/–	0.1209	-	0.4142	-	0.8660	0.9580	0.9860
	UniFuse with fusion [23]	–/–	0.1114	-	0.3691	-	0.8711	0.9664	0.9882
	OmniFusion (2-iter) [30]	256 × 256 / 80°	0.0950	0.0491	0.3474	0.1599	0.8988	0.9769	0.9924
	PanoFormer* [35]	–/–	0.1131	0.0723	0.3557	0.2454	0.8808	0.9623	0.9855
	HRDFuse,Ours	128 × 128 / 80°	0.0984	0.0530	0.3452	0.1465	0.8941	0.9778	0.9923
	HRDFuse,Ours	256 × 256 / 80°	**0.0935**	0.0508	**0.3106**	**0.1422**	**0.9140**	**0.9798**	**0.9927**
Matterport3D	FCRN [28]	–/–	0.2409	-	0.6704	-	0.7703	0.9714	0.9617
	BiFuse with fusion [41]	–/–	0.2048	-	0.6259	-	0.8452	0.9319	0.9632
	UniFuse with fusion [23]	–/–	0.1063	-	0.4941	-	0.8897	0.9623	0.9831
	OmniFusion (2-iter) * [30]	256 × 256 / 80°	0.1007	0.0969	0.4435	0.1664	0.9143	0.9666	0.9844
	PanoFormer* [35]	–/–	0.0904	0.0764	0.4470	0.1650	0.8816	0.9661	0.9878
	HRDFuse,Ours	128 × 128 / 80°	0.0967	0.0936	**0.4433**	**0.1642**	**0.9162**	**0.9669**	0.9844
	HRDFuse,Ours	256 × 256 / 80°	0.0981	0.0945	0.4466	0.1656	0.9147	0.9666	0.9842
3D60	FCRN [28]	–/–	0.0699	0.2833	-	-	0.9532	0.9905	0.9966
	Mapped Convolution [15]	–/–	0.0965	0.0371	0.2966	0.1413	0.9068	0.9854	0.9967
	BiFuse with fusion [41]	–/–	0.0615	-	0.2440	-	0.9699	0.9927	0.9969
	UniFuse with fusion [23]	–/–	0.0466	-	0.1968	-	0.9835	0.9965	0.9987
	ODE-CNN [10]	–/–	0.0467	0.0124	0.1728	0.0793	0.9814	0.9967	0.9989
	OmniFusion (2-iter) [30]	128 × 128 / 80°	0.0430	0.0114	0.1808	0.0735	0.9859	0.9969	0.9989
	HRDFuse,Ours	128 × 128 / 80°	0.0363	0.0103	0.1565	0.0594	0.9888	**0.9974**	0.9990
	HRDFuse,Ours	256 × 256 / 80°	**0.0358**	**0.0100**	**0.1555**	**0.0592**	**0.9894**	0.9973	**0.9990**

Best results compared with TP-based transformer approach!

Problems

1. Merging plenty of patches is **non-trivial process.**

2. The **heavy computation memory and cost** from cross-projection fusion.

3. **Special encoders need to be designed to address distortion issues.**

AI et al. HRDFuse: Monocular 360°Depth Estimation by Collaboratively Learning Holistic-with-Regional Depth Distributions, CVPR, 2023.

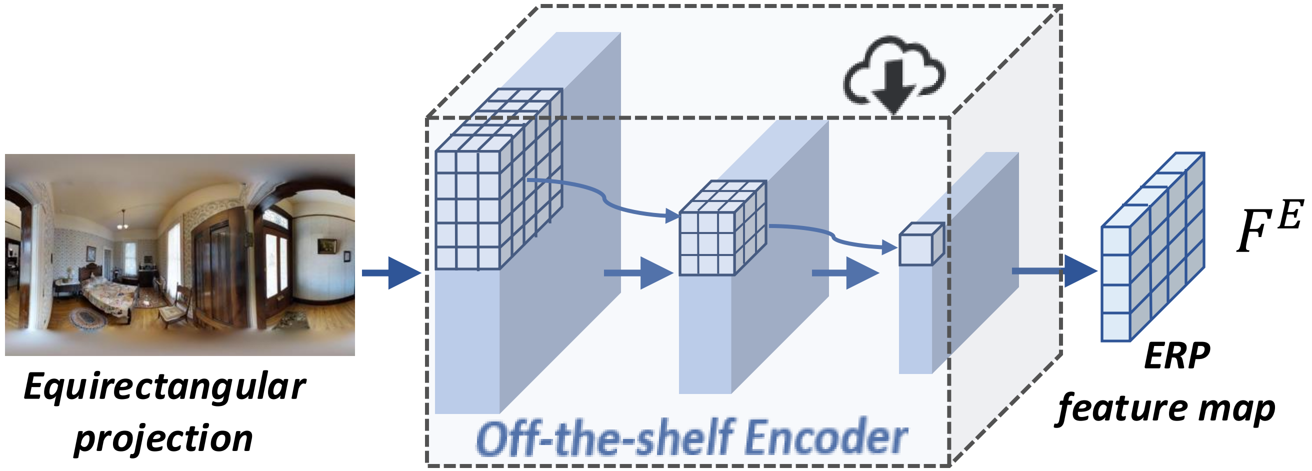
Is it possible to use the off-the-shelf 2D encoders and learn computationally-cheap models?

(Elite360D, CVPR 2024)

ERP+ ICOSAP Fusion for Depth Estimation

Take the best of ERP and ICOSAP by learning a representation from a local-with-global perspective

ERP Branch



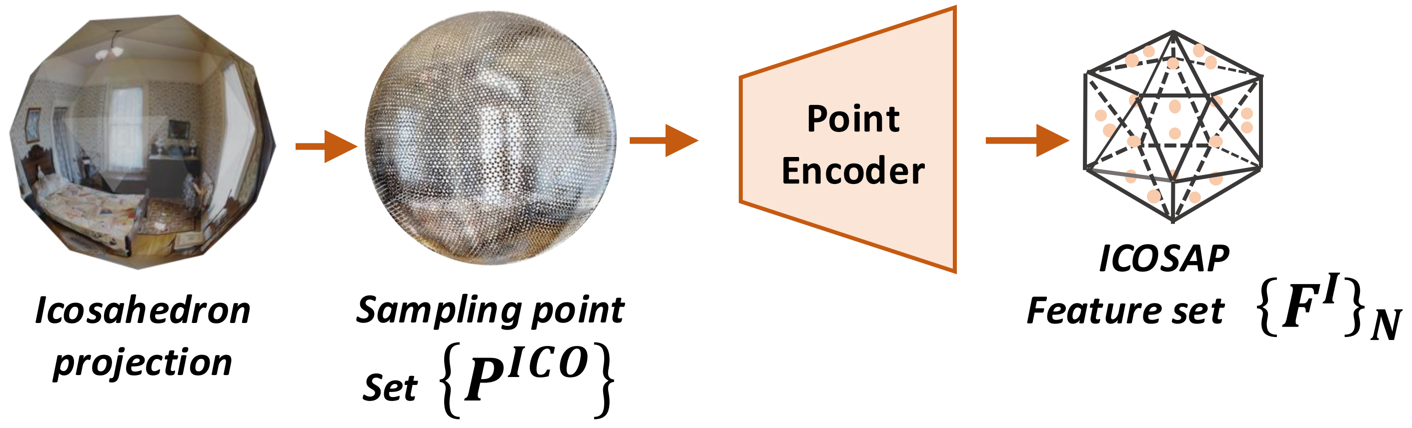
Support a wide range of 2D pretrained models as encoder backbones.

AI et al. Elite360D: Towards Efficient 360 Depth Estimation via Semantic- and Distance-Aware Bi-Projection Fusion, CVPR, 2024.

ERP+ ICOSAP Fusion for Depth Estimation

Take the best of ERP and ICOSAP by learning a representation from a local-with-global perspective

ICOSAP Branch

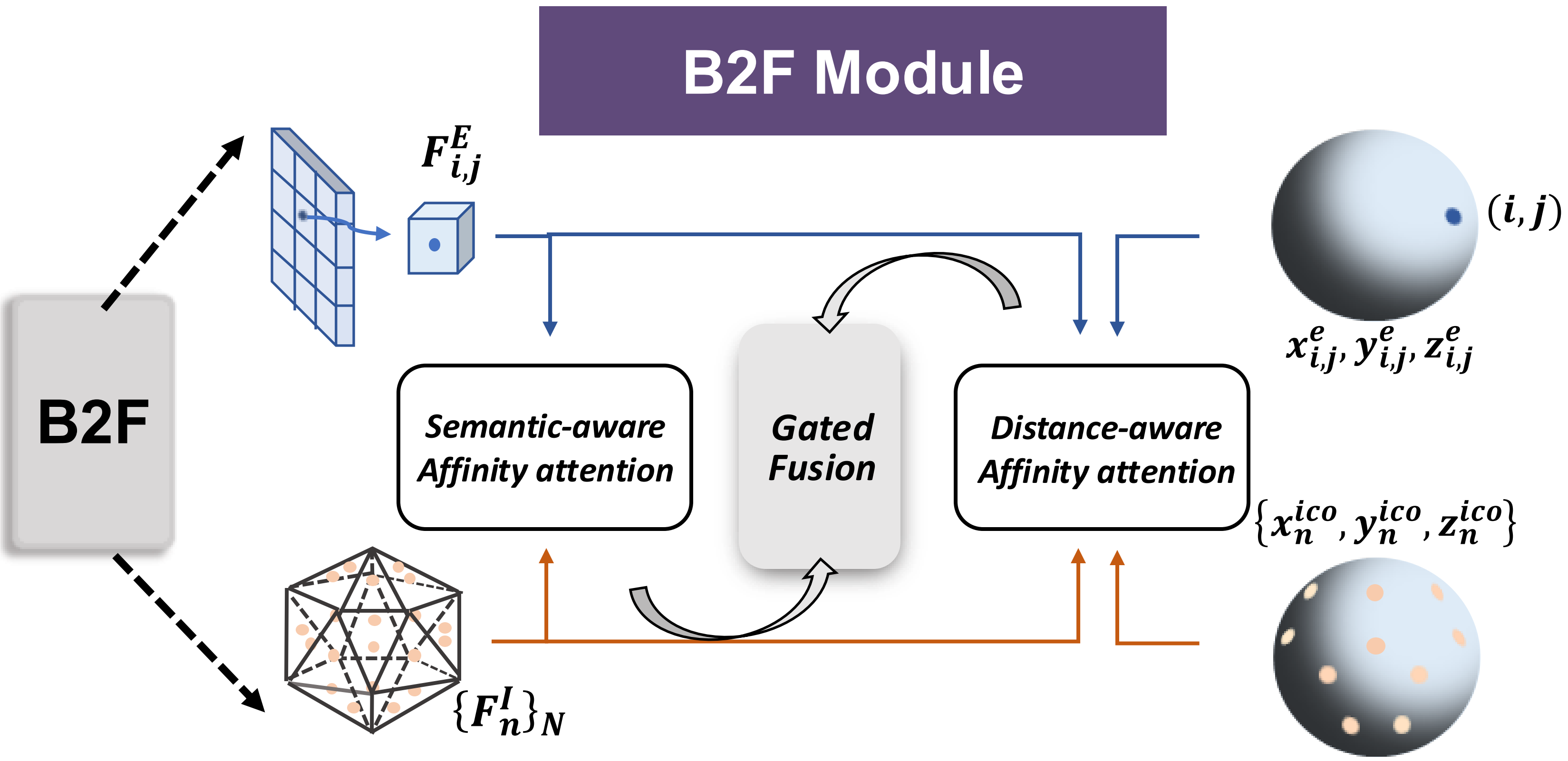


Represent ICOSAP sphere as the *point set*, which contain the spatial information and global perception.

AI et al. Elite360D: Towards Efficient 360 Depth Estimation via Semantic- and Distance-Aware Bi-Projection Fusion, CVPR, 2024.

ERP+ ICOSAP Fusion for Depth Estimation

B2F: Bi-Projection Bi-attention Fusion Module

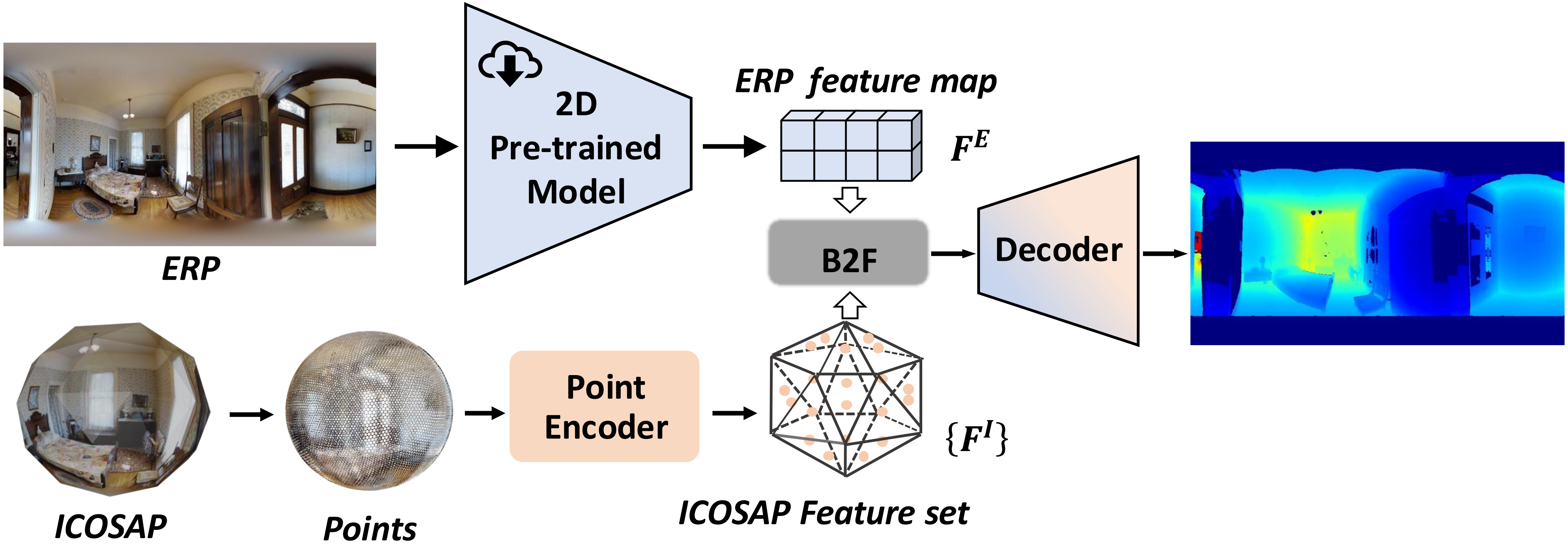


Capture the ***semantic- and distance-aware dependencies*** between each ERP pixel feature and entire ICOSAP feature set

AI et al. Elite360D: Towards Efficient 360 Depth Estimation via Semantic- and Distance-Aware Bi-Projection Fusion, CVPR, 2024.

ERP+ ICOSAP Fusion for Depth Estimation

Take the best of ERP and ICOSAP by learning a representation from a local-with-global perspective



AI et al. Elite360D: Towards Efficient 360 Depth Estimation via Semantic- and Distance-Aware Bi-Projection Fusion, CVPR, 2024.

ERP+ ICOSAP Fusion for Depth Estimation

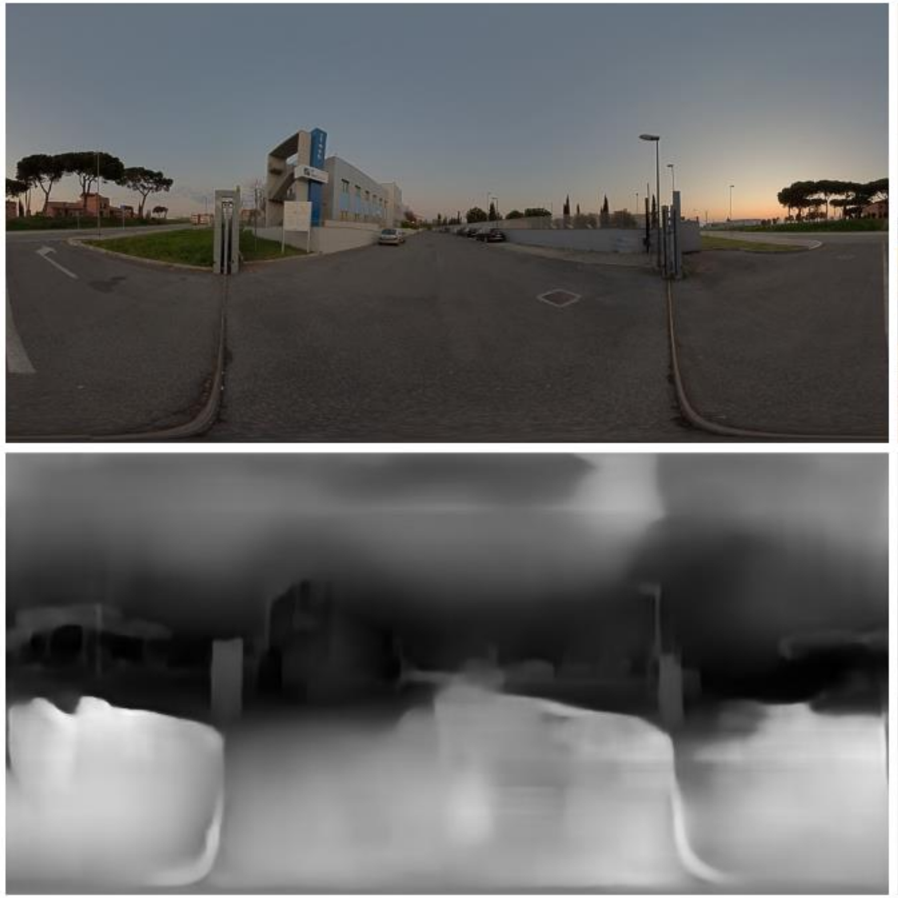
Smallest model size with on par performance

Datasets	Backbone	Method	Pub'Year	#Params (M)	#FLOPs (G)	Abs Rel ↓	Sq Rel ↓	RMSE ↓	$\delta_1(\%)\uparrow$	$\delta_2(\%)\uparrow$	$\delta_3(\%)\uparrow$
M3D	Transformer	EGFormer	ICCV'23	15.39	66.21	0.1473	0.1517	0.6025	81.58	93.90	97.35
		PanoFormer	ECCV'22	20.38	81.09	0.1051	0.0966	0.4929	89.08	96.23	98.31
	ResNet-18	BiFuse	CVPR'20	35.80	165.66	0.1360	0.1202	0.5488	83.27	95.12	98.10
		UniFuse	RAL'21	30.26	62.60	0.1191	0.1030	0.5158	86.04	95.84	98.30
		OmniFusion	CVPR'22	32.35	98.68	0.1209	0.1090	0.5055	86.58	95.81	98.36
		HRDFuse†	CVPR'23	26.09	50.59	0.1414	0.1241	0.5507	81.48	94.89	98.20
		Ours	-	15.43	45.91	0.1272	0.1070	0.5270	85.28	95.28	98.49
	ResNet-34	BiFuse	CVPR'20	56.01	199.58	0.1126	0.0992	0.5027	88.00	96.13	98.47
		BiFuse++	TPAMI'22	52.49	87.48	0.1123	0.0915	0.4853	88.12	96.56	98.69
		UniFuse	RAL'21	50.48	96.52	0.1144	0.0936	0.4835	87.85	96.59	98.73
		OmniFusion	CVPR'22	42.46	142.29	0.1161	0.1007	0.4931	87.72	96.15	98.44
		HRDFuse†	CVPR'23	46.31	80.87	0.1172	0.0971	0.5025	86.74	96.17	98.49
		Ours	-	25.54	65.29	0.1115	0.0914	0.4875	88.15	96.46	98.74
	ResNet-50*	BiFuse	CVPR'20	253.08	775.24	0.1179	0.0981	0.4970	86.74	96.27	98.66
		UniFuse	RAL'21	131.30	222.30	0.1185	0.0984	0.5024	86.66	96.18	98.50
		Ours	-	42.99	170.11	0.1112	0.0980	0.4870	86.70	96.01	98.61

AI et al. Elite360D: Towards Efficient 360 Depth Estimation via Semantic- and Distance-Aware Bi-Projection Fusion, CVPR, 2024.

Data-specific models are difficult to be generalized to unseen scenes (outdoor and indoor)

Can we get 360 foundation model?

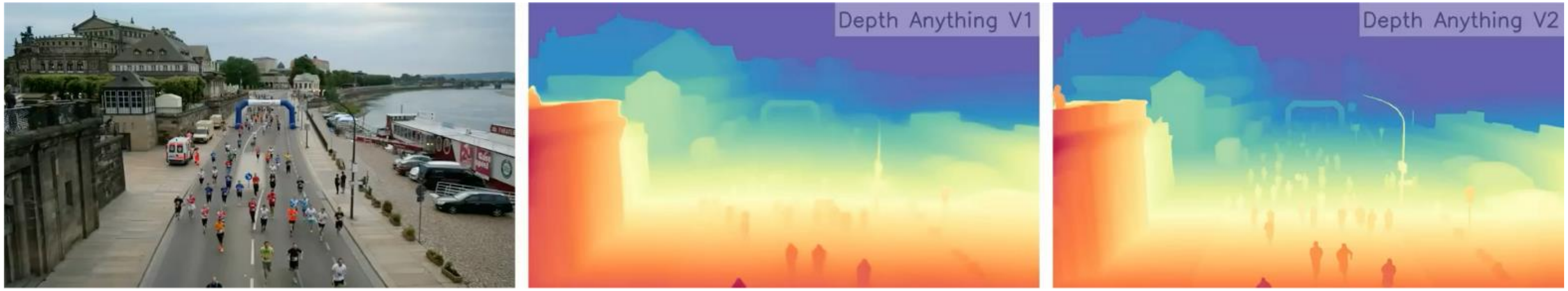


Unseen outdoor scenes

Vision Foundation Models

Depth Anything v1 & v2

- Utilize large amount of labeled and unlabeled data for training.



Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

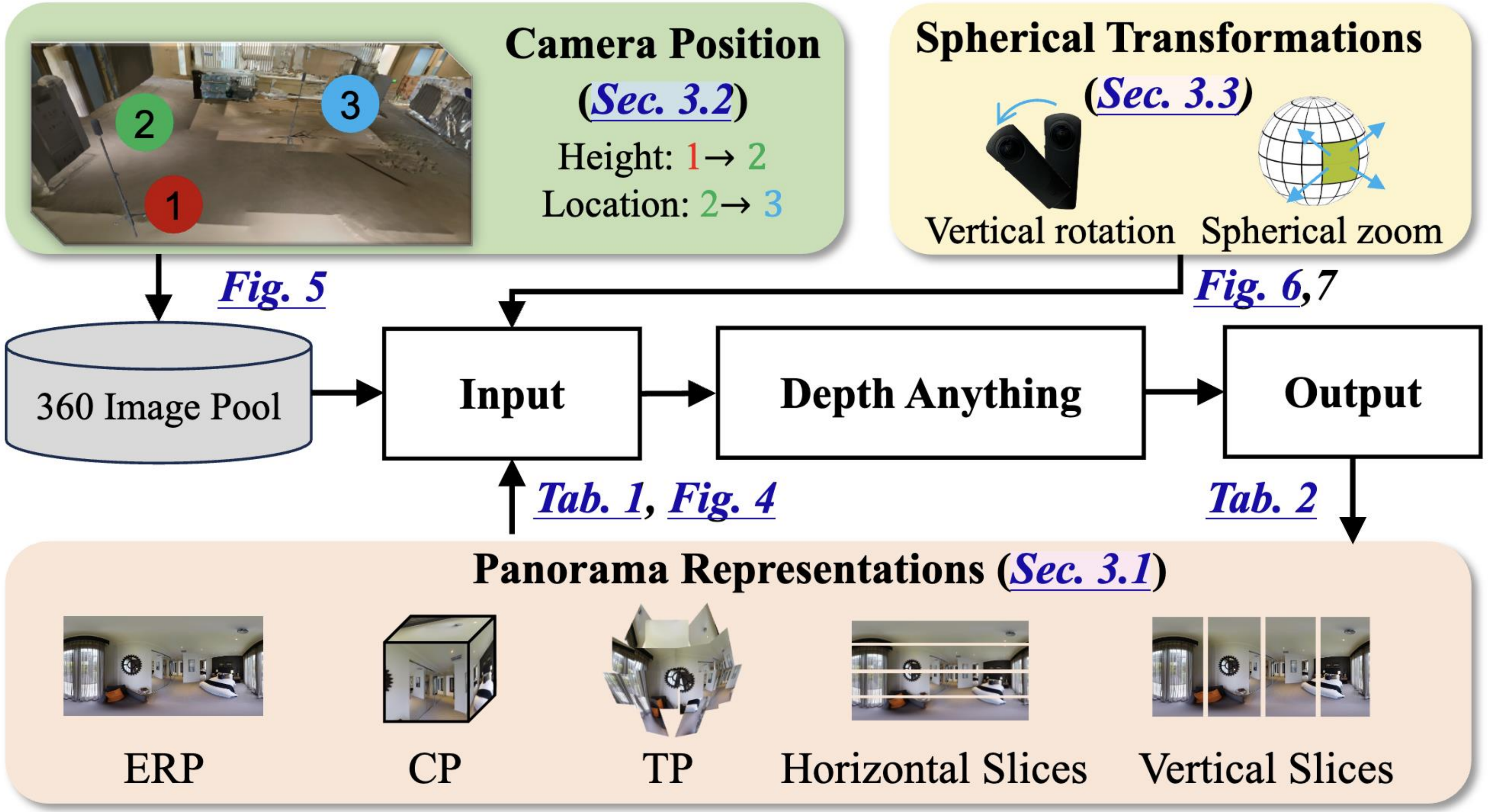
Do depth foundation models generalize well to 360 data across diverse scenes? (PanDA, CVPR 2025)

Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

Leaderboard of Performance

Check the spherical properties of Depth Anything for panoramas

- Panoramic representations; Camera positions; and Spherical transformations.

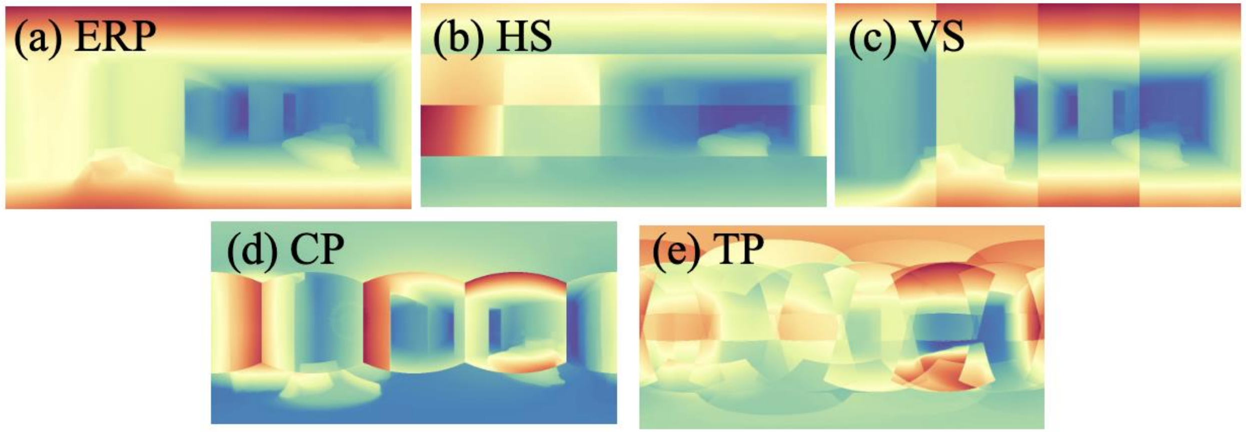


Finding 1

Check which panoramic representation is better for Depth Anything

- When the output space is ERP, ERP performs the best.

Method	Backbone	ERP	CP	TP	HS	VS
DAM v1 [49]	ViT-S	**0.1687**	0.2144	0.2289	0.2104	0.1873
	ViT-B	**0.1629**	0.2238	0.2251	0.2073	0.1889
	ViT-L	**0.1614**	0.2165	0.2046	0.2043	0.1858
DAM v2 [50]	ViT-S	**0.1692**	0.2205	0.2317	0.2186	0.1962
	ViT-B	**0.1662**	0.2249	0.2460	0.2149	0.2006
	ViT-L	**0.1654**	0.2238	0.2363	0.2101	0.1984



Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

Finding 2

Check which panoramic representation is better for Depth Anything

- When the output space is other projections, **ERP performs mostly the best**.
- The **prior knowledge** in Depth Anything can **address spherical distortion in some level**.

Inp. → Out.	Equator	Pole	Average
ERP → CP	**0.1129**	**0.1201**	**0.1153**
CP → CP	0.1164	0.1357	0.1228
ERP → TP	**0.1235**	**0.1232**	**0.1234**
TP → TP	0.1416	0.1492	0.1441
ERP → HS	**0.1322**	**0.0965**	**0.1145**
HS → HS	0.1760	0.1251	0.1507
ERP → VS	–	–	0.1438
VS → VS			**0.1355**

Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

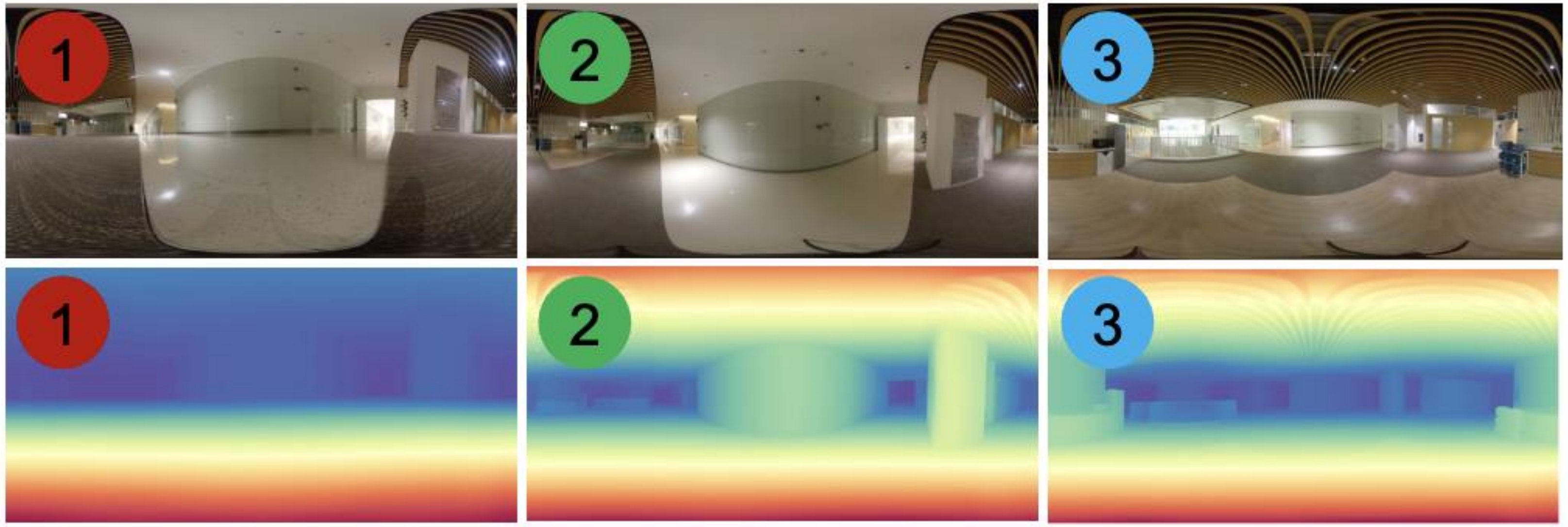
Finding 3

Check the camera position

1. Near the ground, **poor result** due to **large portion of polar regions**;

2. Lift the height, **better result**;

3. Lift the height and move towards the **interested objects**, **best result**.

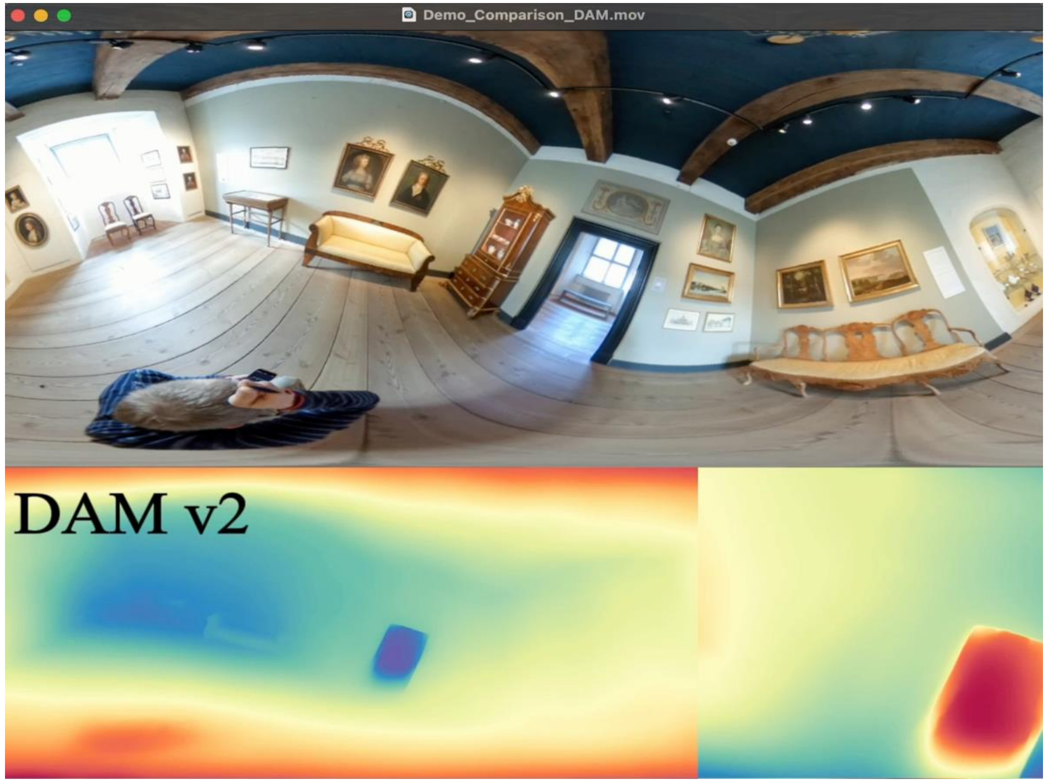


Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

Finding 4

Check the robustness of Depth Anything for panoramas

- Depth Anything performs poorly with spherical transformations.

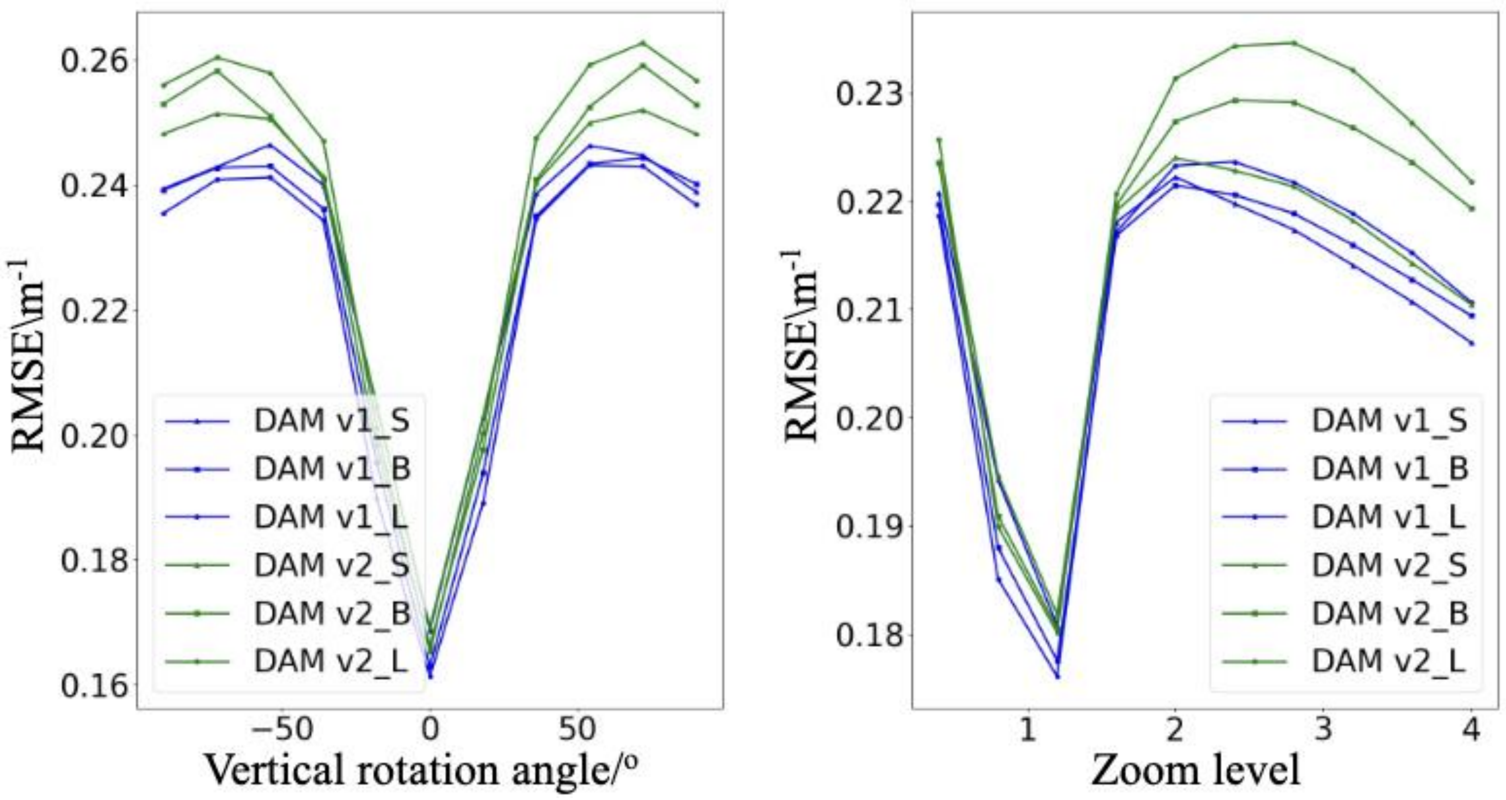


Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

Finding 5

Check the robustness of Depth Anything for panoramas

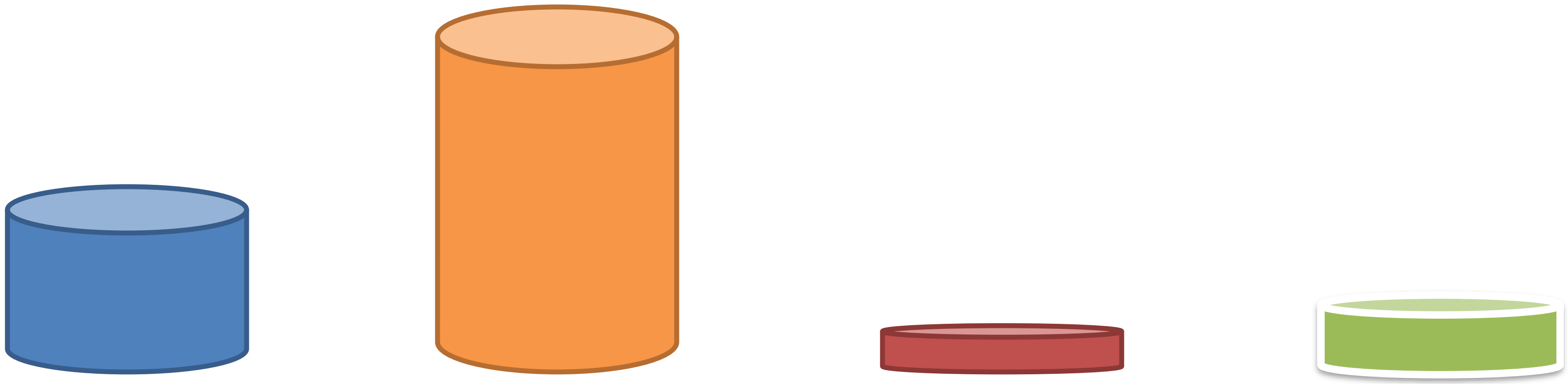
- The performance of Depth Anything **changes rapidly**.



Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

Hurdles for Achieving 360 Foundation Model

360 depth ground truth is difficult to acquire



Labeled Perspective

> 10M

Unlabeled

Perspective

>50M

Labeled 360

< 0.1M

Unlabeled 360

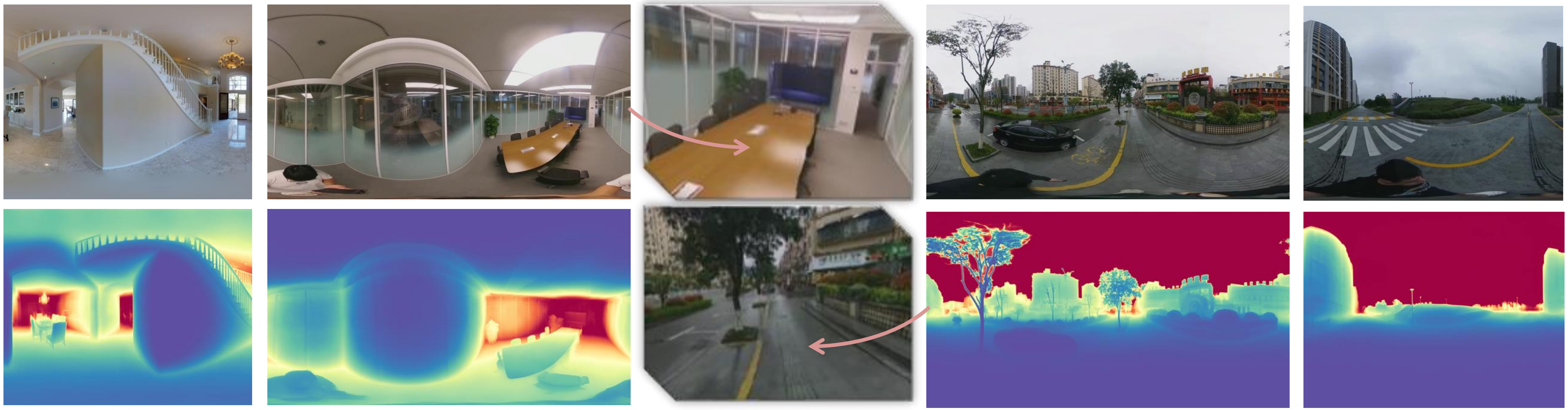
About 0.4M

Complicated annotation on curves, and require stitching

Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

Achieving 360 Depth Anything

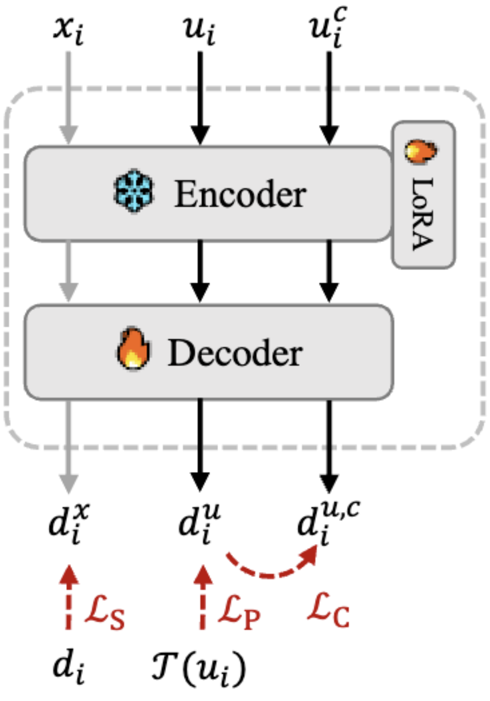
- **Fine-tune** Depth Anything to omnidirectional images with three stages of semi-supervised learning
- **Teacher model training** with synthetic panoramas;
- Collect **100k unlabeled real-world images** and generate pseudo labels from the teacher model;
- **Student model training** using both labeled and unlabeled data.



Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

Achieving 360 Depth Anything

- Semi-supervision pipeline
- Enforce consistency between the original unlabeled panorama and color-augmented ones.

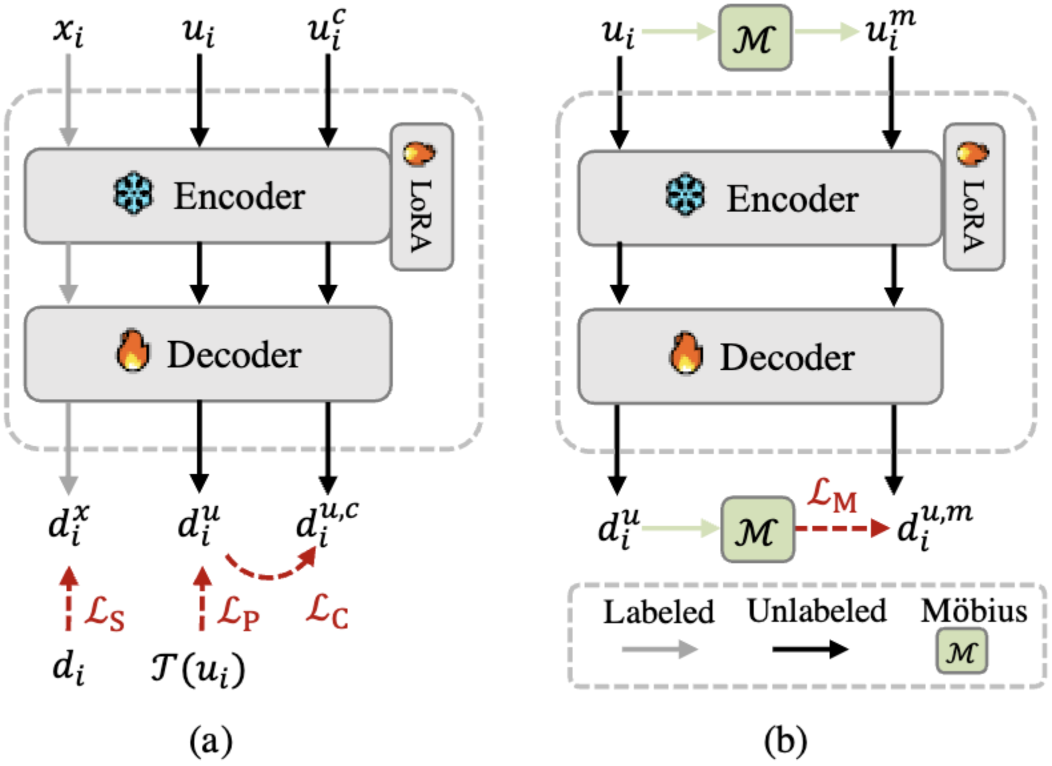


(a)

Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

Achieving 360 Depth Anything

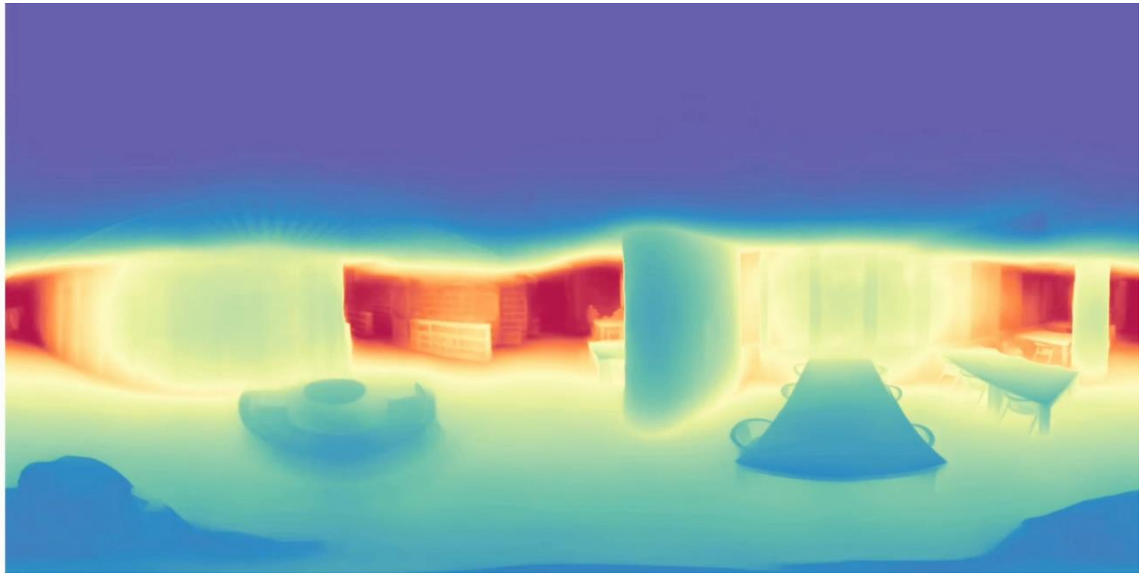
- Semi-supervision pipeline
- Enforce consistency between the original unlabeled panorama and color-augmented ones.
- Enforce consistency between the original unlabeled **panorama and spherical-transformed** ones.



Mobius transformation-based spatial augmentation (MTSA)

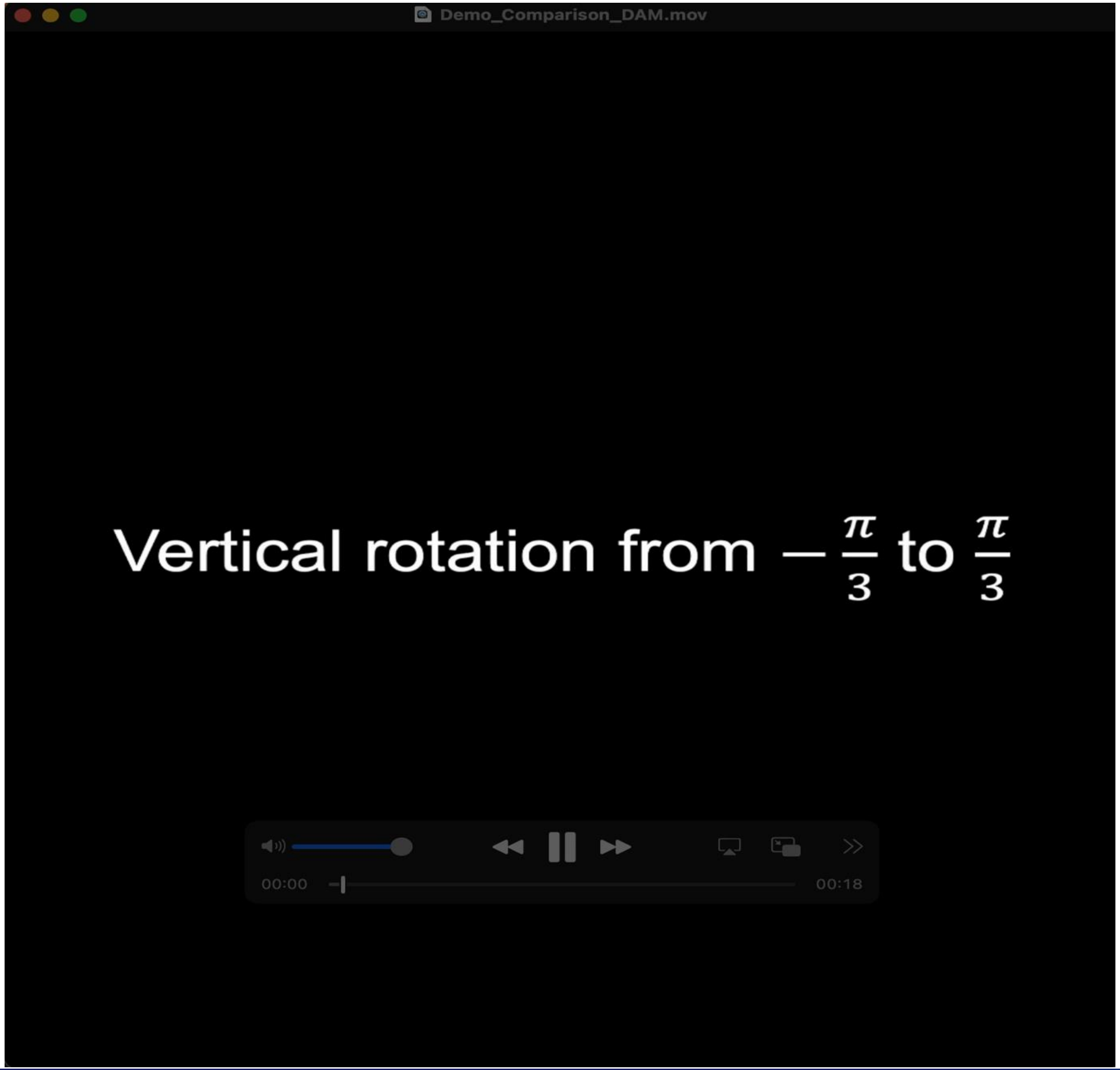
Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

Open World Results



Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

Open-World Results (Random Transformation)



Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

New Approach to 360 Depth Anything

Scale in Data
31% **Improvement**

Scale in Data and Model
45% **Improvement**

Methods	*AbsRel* ↓	*RMSE* ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
BiFuse [40]	0.1209	0.4142	86.60	95.80	98.60
UniFuse [17]	0.1114	0.3691	87.11	96.64	98.82
HoHoNet [38]	0.1014	0.3834	90.54	96.93	98.86
BiFuse++ [41]	–	0.3720	87.83	96.49	98.84
ACDNet [55]	0.0984	0.3410	88.72	97.04	98.95
PanoFormer [35]	0.1131	0.3557	88.08	96.23	98.55
HRDFuse [2]	0.0935	0.3106	91.40	97.98	99.27
S2Net [24]	0.0903	0.3383	91.91	97.82	99.12
Depth Anywhere [42]	0.1180	0.3510	91.00	97.10	98.70
PanDA-S	0.0762	0.2866	95.31	98.60	99.36
PanDA-B	0.0635	0.2682	95.84	98.95	99.51
PanDA-L	**0.0609**	**0.2540**	**96.82**	**99.05**	**99.52**

Cao et al. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation, CVPR, 2025.

Takeaways

ERP+ CP (2021)

ERP+TP (2023)

ERP+ ICOSAP (2024)

360 foundation model (2025)

What's next?

Edge AI Applications

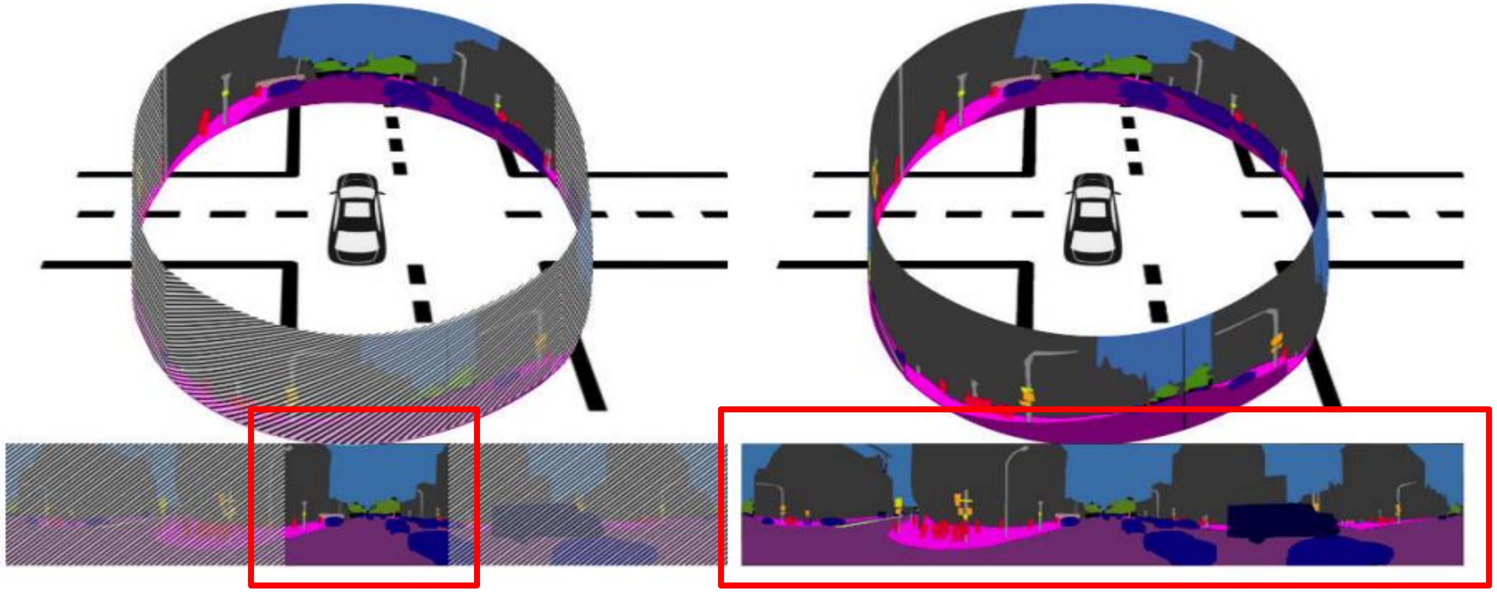
Bi-projection can complement each other but models are not scalable

Table of Contents

We are here now!

Takeaways: Learning strategies are important for efficient and effective 360-based scene understanding!

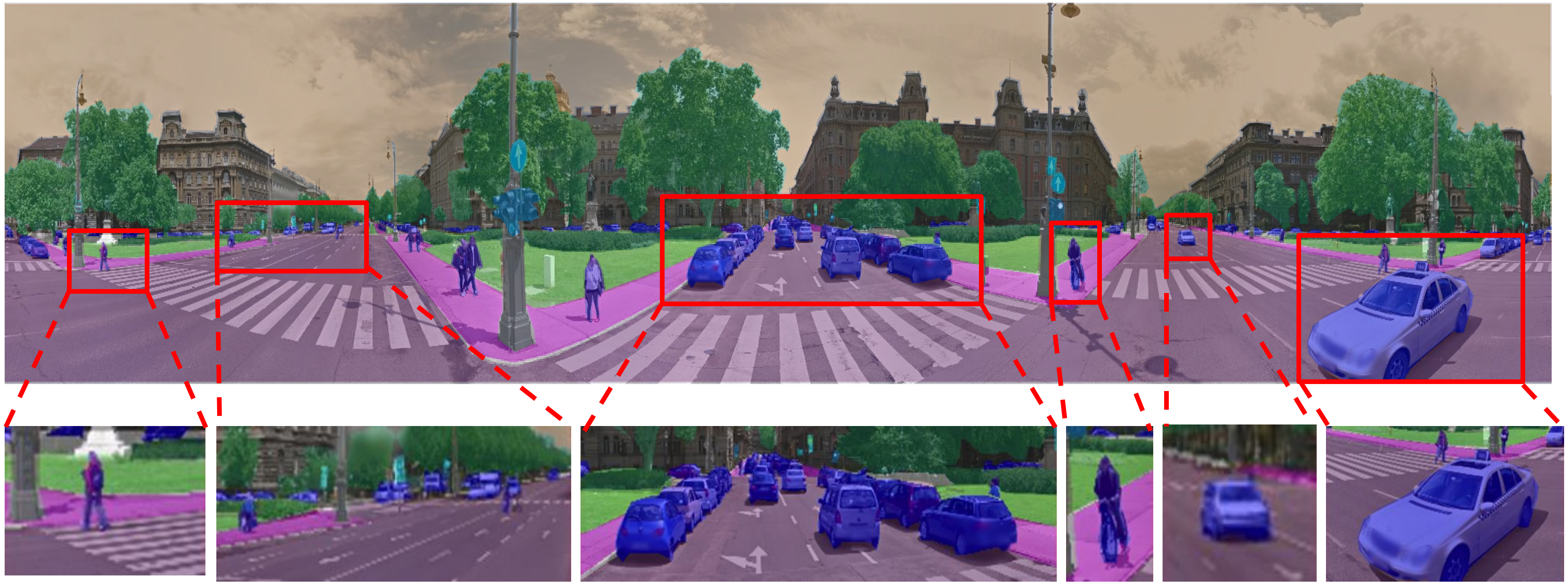
What can we benefit from 360 cameras?



Perspective vs 360 camera

360 cameras'comprehensive view of the vehicle's surroundings, eliminating blind spots and increasing situational awareness.

What can we benefit from 360 cameras?



Lack of labeled data (labeling 360 images is very time-consuming and labor-intensive) !

Zheng et al, Both Style and Distortion Matter: Dual-Path Unsupervised Domain Adaptation for Panoramic Semantic Segmentation, CVPR, 2023.

Transferring knowledge

Pinhole image

Panoramic (ERP) image

Limited FoV

No Distortion

Sufficient Labels

Broader / 360 FoV

Severe Distortion

Scarce Labels

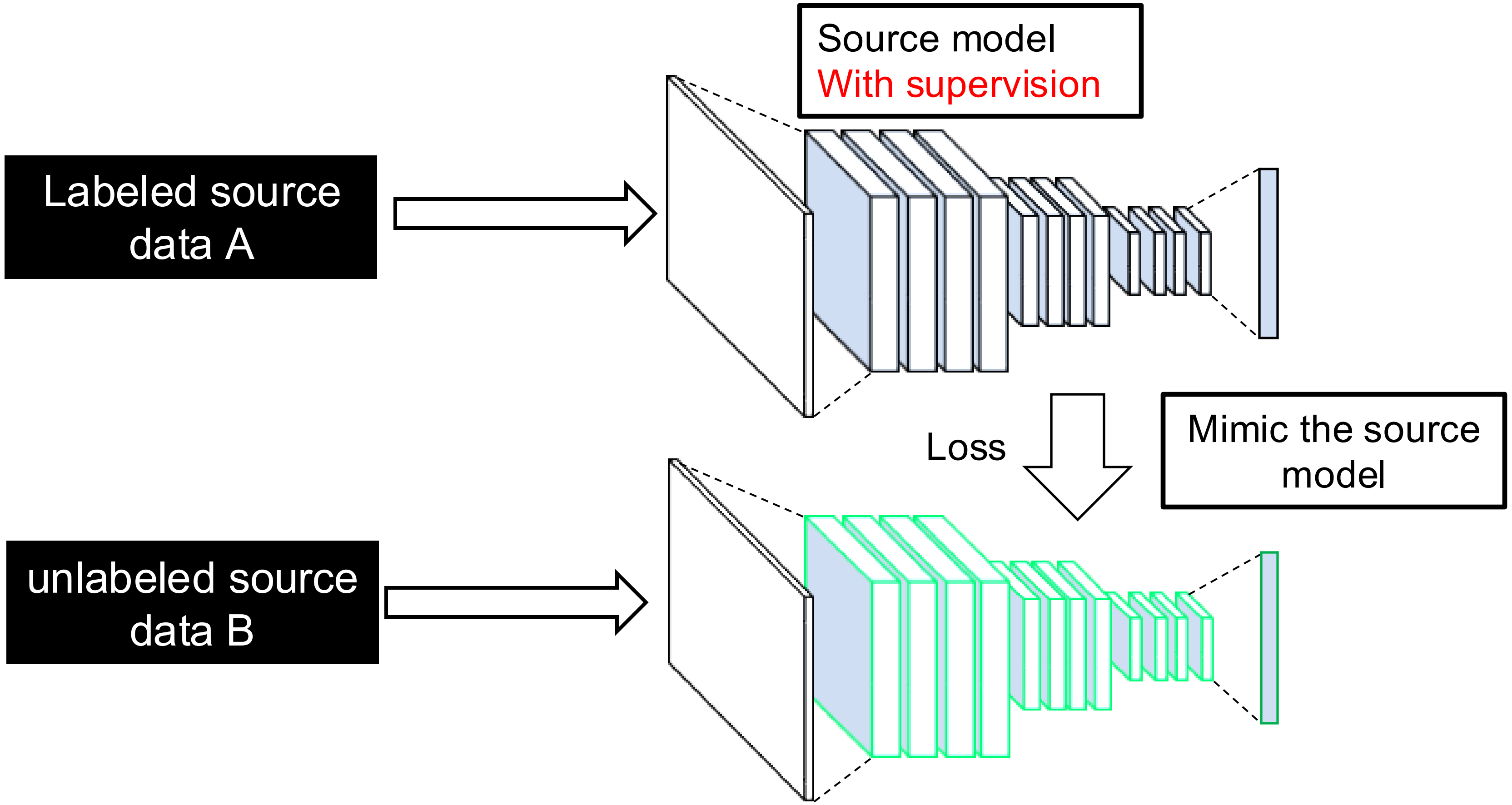
Unsupervised Domain Adaptation (UDA)

Zheng et al, Both Style and Distortion Matter: Dual-Path Unsupervised Domain Adaptation for Panoramic Semantic Segmentation, CVPR, 2023.

Transferring knowledge

How it works?

- Transfer knowledge from a trained model on source data A to a new model for target data B.
- To improve training of the new model (on unlabeled target data B)



Wang, Lin, and Kuk-Jin Yoon. "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks." IEEE TPAMI (2021).

Research Questions

Synthetic

(GTA5)

Real Pinhole

(Cityscapes)

Real Panorama

(DensePASS)

Domain Gaps: Style

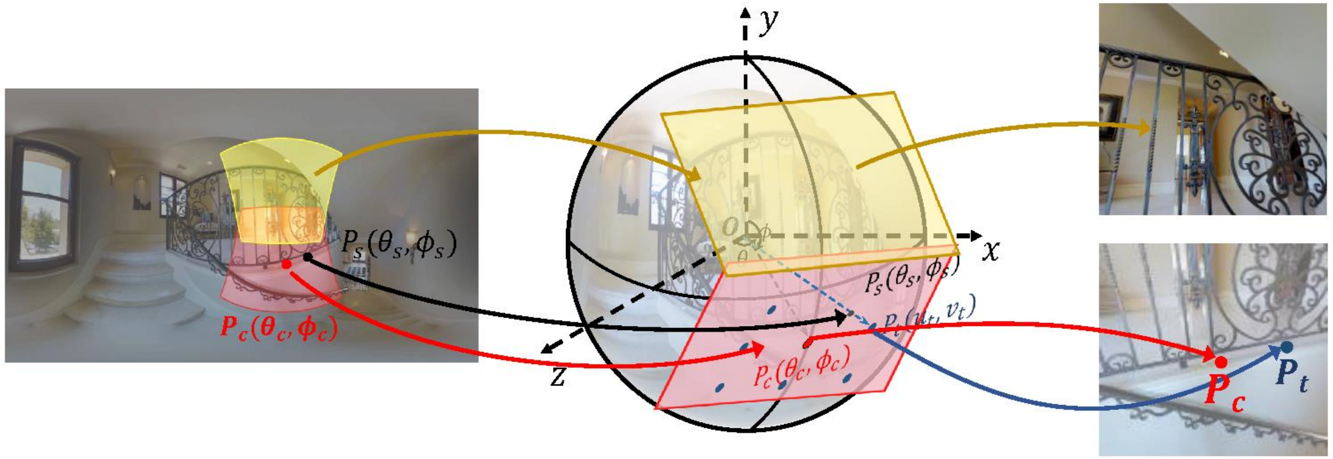
Domain Gaps: Style & Distortion

Domain Gaps
- **Inherent Gaps: Diverse camera sensors and captured scenes**
- **Format Gaps: Distinct image representation formats**

Research question: How to alleviate these domain gaps?

Zheng et al, Both Style and Distortion Matter: Dual-Path Unsupervised Domain Adaptation for Panoramic Semantic Segmentation, CVPR, 2023.

Key Idea



Use tangent projection (TP) along with ERP for alleviating the format gap caused by distortion.

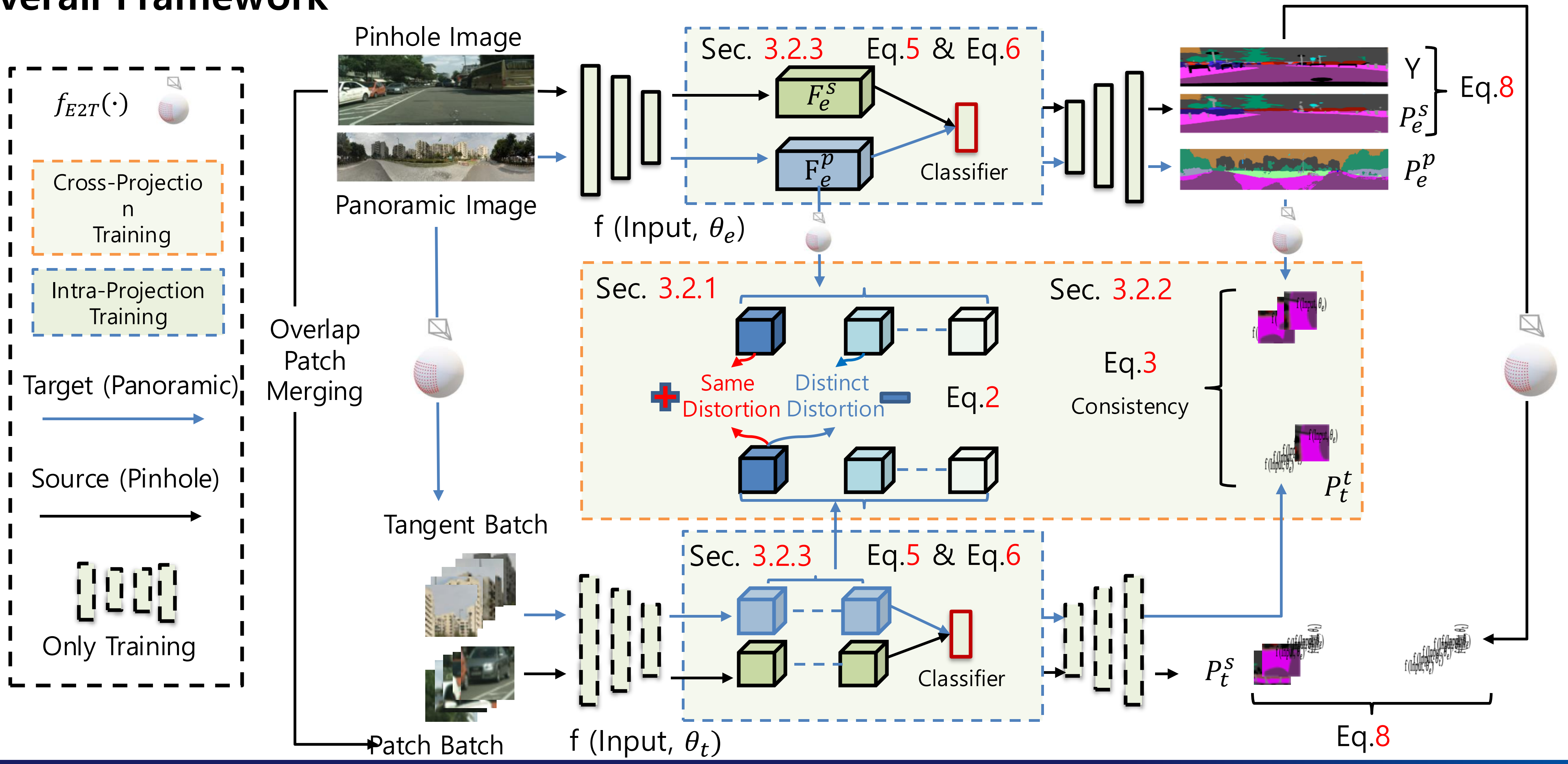
Our Key Idea!

Dual-Path Framework
- Cross-Projection Training
  - ERP & TP
    - Feature level
      - Tangent-wise feature contrastive training
    - Prediction level
      - Prediction consistency training
- Intra-Projection Training
  - ERP path
  - TP path

Zheng et al, Both Style and Distortion Matter: Dual-Path Unsupervised Domain Adaptation for Panoramic Semantic Segmentation, CVPR, 2023.

Transferring knowledge

Overall Framework



Transferring knowledge

Cross Projection Training:

- Tangent-wise feature contrastive training:

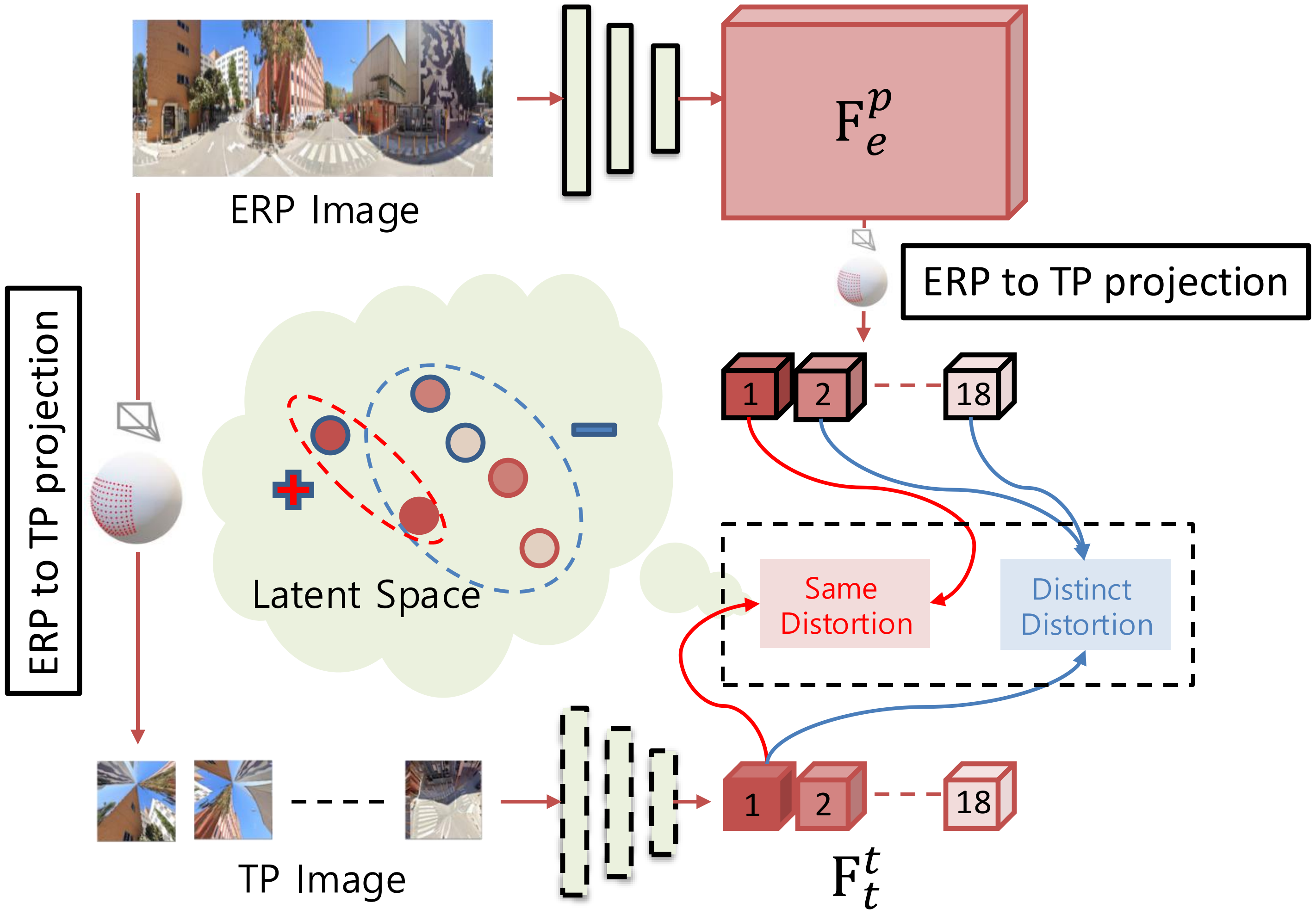
Positive Negative

$$L_{fc} = \frac{1}{F_i} \sum_{f_+ \in F_i} -log \frac{exp(f_+/\tau)}{exp(f_+/\tau) + \sum_f exp(f_-/\tau)}$$

- Prediction consistency training:

$$\mathcal{L}_{pc} = \sum_{i=1}^{18} f_{E2T}(P_{ei}^p) \log \frac{f_{E2T}(P_{ei}^p))}{P_{ti}^t}$$

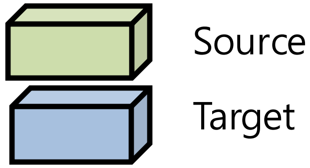
Choose 18 patches for prediction consistency



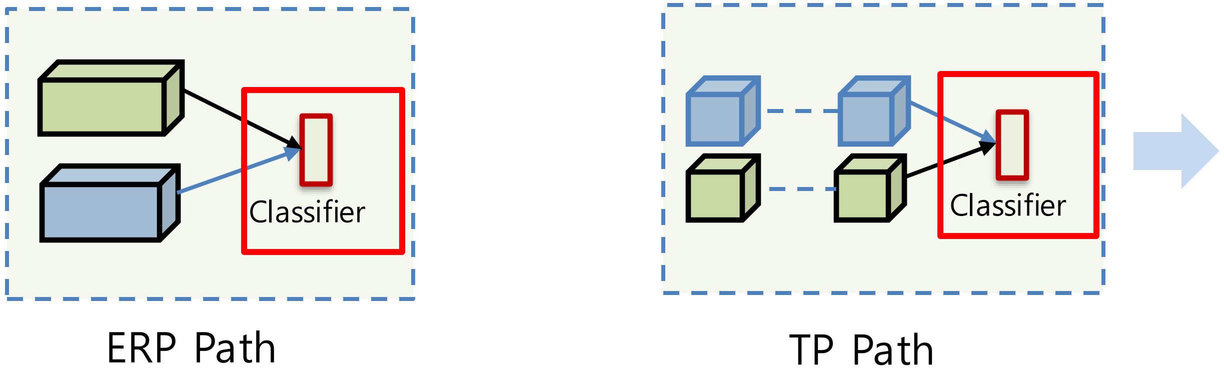
Zheng et al, Both Style and Distortion Matter: Dual-Path Unsupervised Domain Adaptation for Panoramic Semantic Segmentation, CVPR, 2023.

Transferring knowledge

Intra Projection Training:

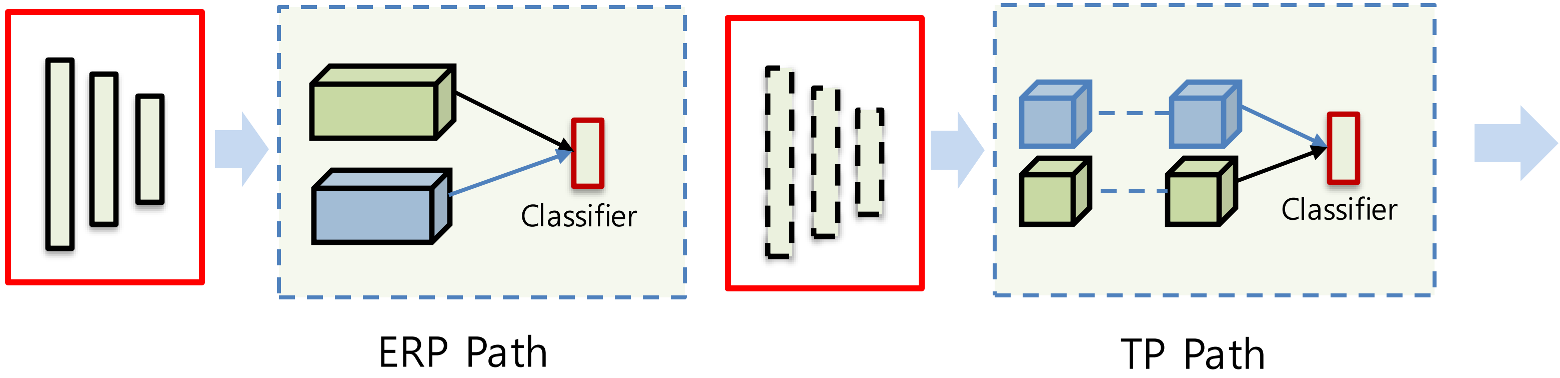


Classifier:



Distinguish features (distortion)

Feature Extractor:



Generate domain-invariant features

Zheng et al, Both Style and Distortion Matter: Dual-Path Unsupervised Domain Adaptation for Panoramic Semantic Segmentation, CVPR, 2023.

Transferring knowledge

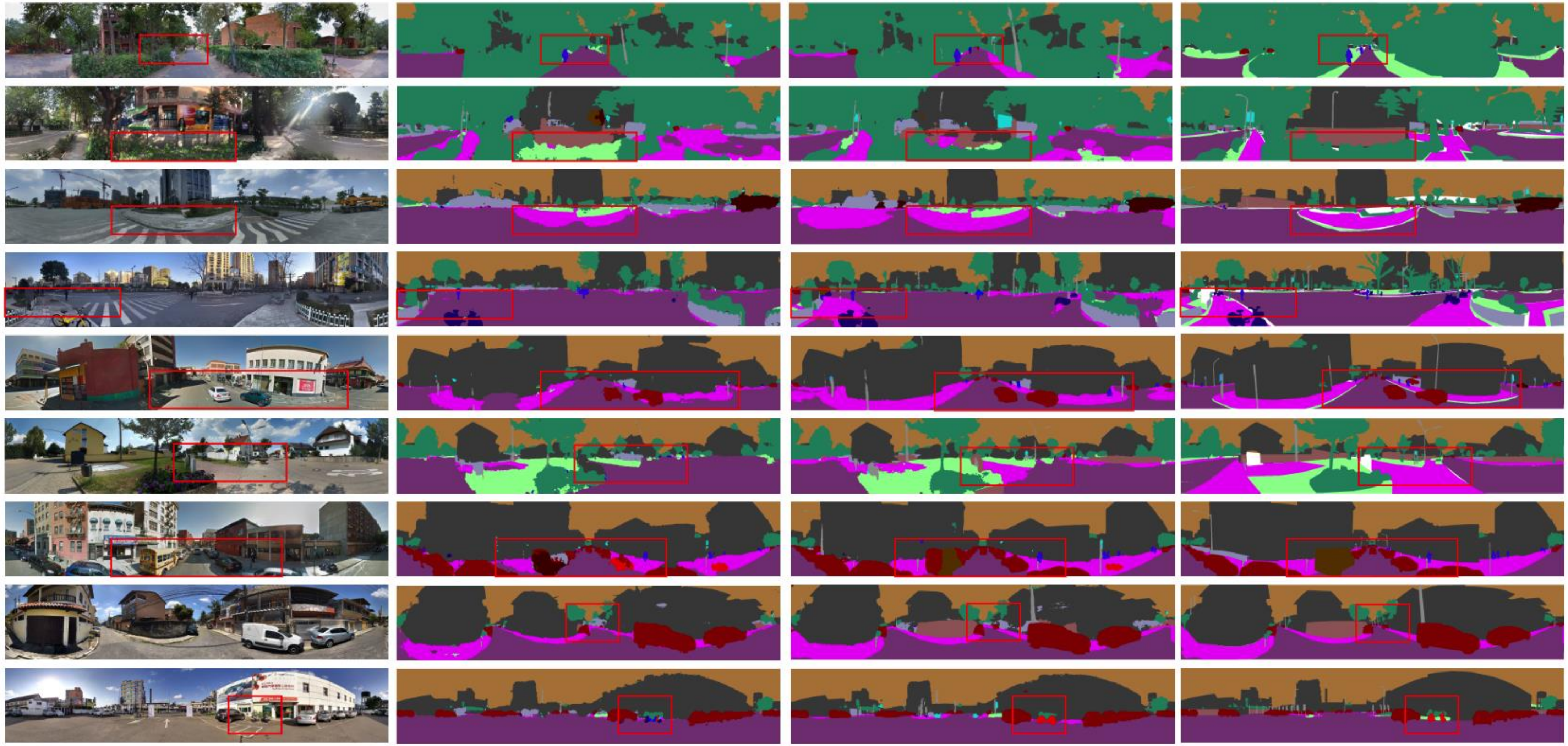
Per-class results of the SoTA panoramic image semantic segmentation methods on DensePASS test set.

Method	mIoU	road	sidewalk	building	wall	fense	pole	traffic Light	traffic Sign	tegetation	terrain	sky	Person	rider	car	truck	bus	train	motorcycle	bicycle
ERFNet	16.65	63.59	18.22	47.01	9.45	12.79	17.00	8.12	6.41	34.24	10.15	18.43	4.96	2.31	46.03	3.19	0.59	0.00	8.30	5.55
PASS(ERFNet)	23.66	67.84	28.75	59.69	19.96	29.41	8.26	4.54	8.07	64.96	13.75	33.50	12.87	3.17	48.26	2.17	0.82	0.29	23.76	19.46
Omni-sup(ECANet)	43.02	81.60	19.46	81.00	32.02	39.47	25.54	3.85	17.38	79.01	39.75	94.60	46.39	12.98	81.96	49.25	28.29	0.00	55.36	29.47
P2PDA(Adversarial)	41.99	70.21	30.24	78.44	26.72	28.44	14.02	11.67	5.79	68.54	38.20	85.97	28.14	0.00	70.36	60.49	38.90	77.80	39.85	24.02
PCS	53.83	78.10	46.24	86.24	30.33	45.78	34.04	22.74	13.00	79.98	33.07	93.44	47.69	22.53	79.20	61.59	67.09	83.26	58.68	39.80
Trans4PASS-T †	53.18	78.13	41.19	85.93	29.88	37.02	32.54	21.59	18.94	78.67	45.20	93.88	48.54	16.91	79.58	65.33	55.76	84.63	59.05	37.61
Trans4PASS-S †	55.22	78.38	41.58	86.48	31.54	45.54	33.92	22.96	18.27	79.40	41.07	93.82	48.85	23.36	81.02	67.31	69.53	86.13	60.85	39.09
DPPASS-T(Ours)	**55.30**	78.74	46.29	87.47	**48.62**	40.47	**35.38**	24.97	17.39	79.23	40.85	93.49	52.09	**29.40**	79.19	58.73	47.24	**86.48**	66.60	38.11
DPPASS-S(Ours)	**56.28**	78.99	**48.14**	**87.63**	42.12	44.85	34.95	**27.38**	**19.21**	78.55	43.08	92.83	**55.99**	29.10	80.95	61.42	55.68	79.70	**70.42**	38.40

Huge boost to key targets (classes) for *autonomous driving purpose.*

Zheng et al, Both Style and Distortion Matter: Dual-Path Unsupervised Domain Adaptation for Panoramic Semantic Segmentation, CVPR, 2023.

Transferring knowledge



Input | Trans4PASS[5] | Ours | GT

Zheng et al, Both Style and Distortion Matter: Dual-Path Unsupervised Domain Adaptation for Panoramic Semantic Segmentation, CVPR, 2023.

Towards 360 Foundation Models for Segmentation!

Knowledge transfer from Segment Anything Model (GoodSAM, CVPR 2024)

Emergence of Segment Anything Model (SAM)

Segment Anything Model (SAM): a new AI model from Meta AI that can "cut out" any object, in any image, with a single click.

- SAM AMG is a **prompt-free mode of SAM** that automatically generates multi-level masks for all visible objects in an image — **no manual prompts or training required**.

Prompt it with interactive points and boxes.

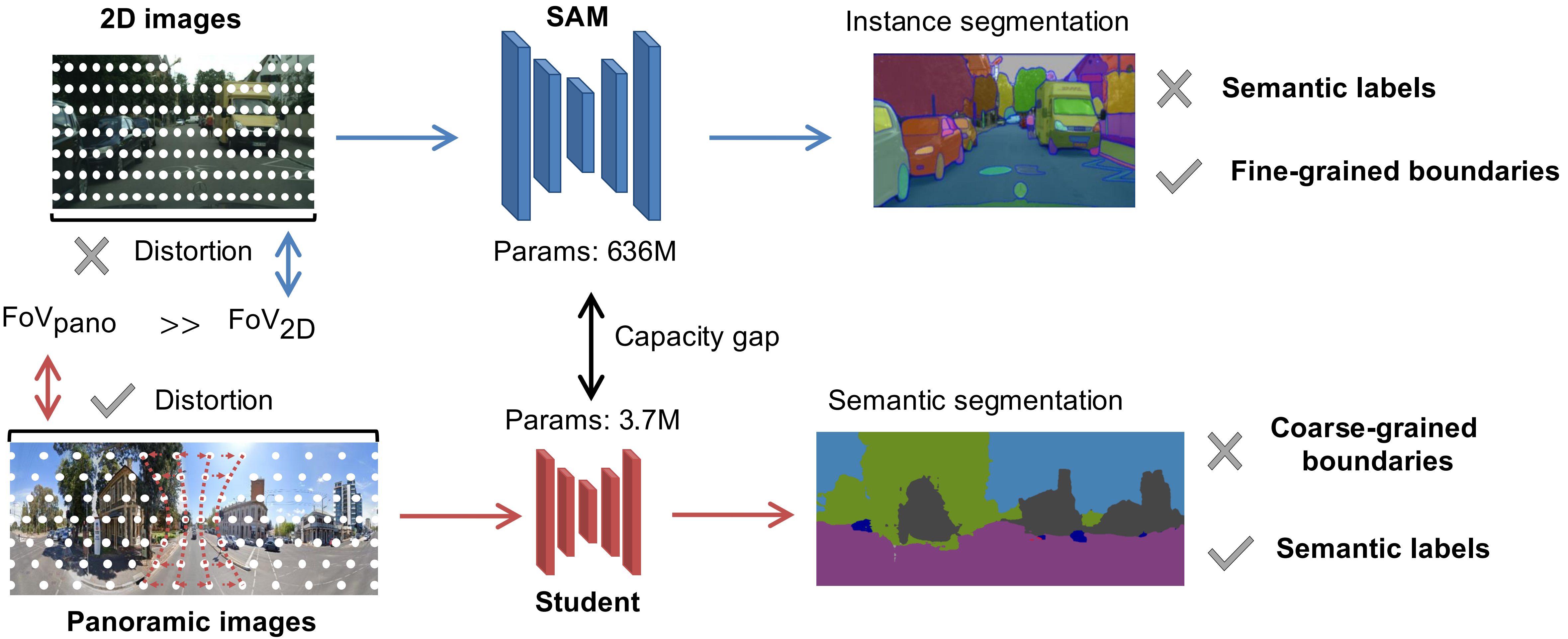
Generate multiple valid masks for ambiguous prompts.

Automatically segment everything (AMG mode) in an image.

All demos are from SAM official website.

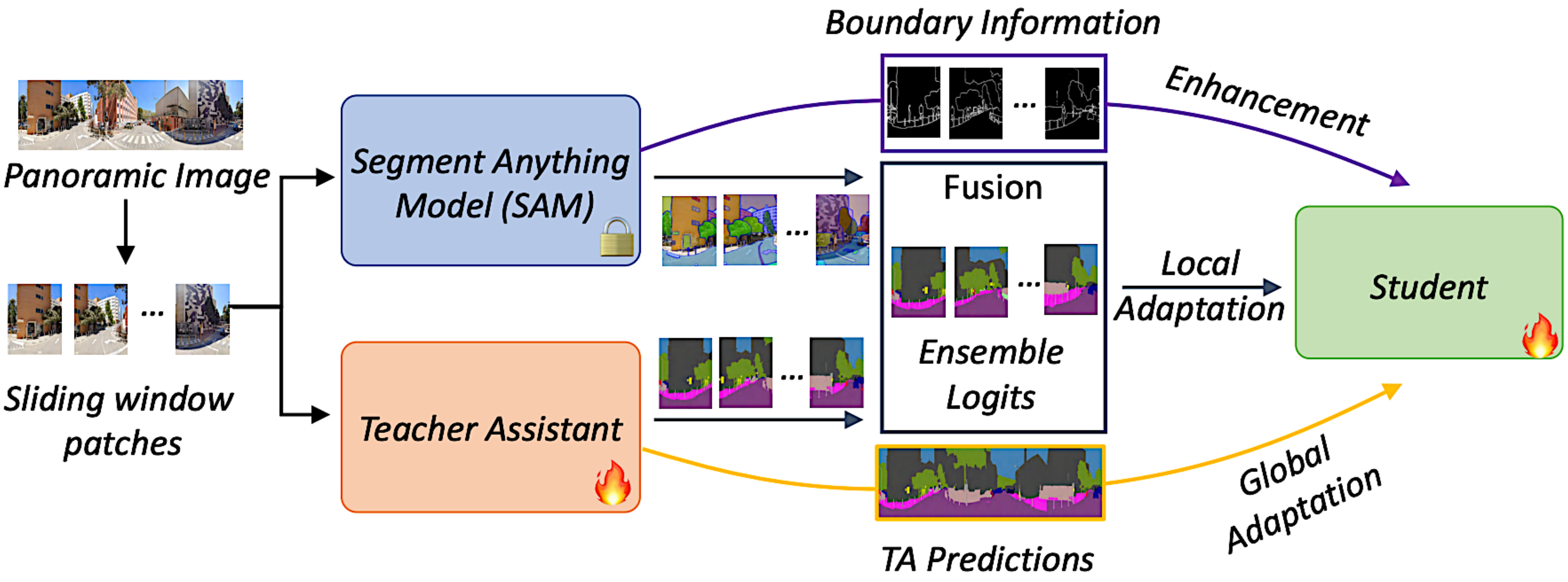
Emergence of Segment Anything Model (SAM)

How to transfer the instance segmentation knowledge from SAM to learn a more compact p anoramic semantic segmentation model (i.e., student) without requiring any labeled data?



Zhang et al, GoodSAM: Bridging Domain and Capacity Gaps via Segment Anything Model for Distortion-aware Panoramic Semantic Segmentation, CVPR, 2025.

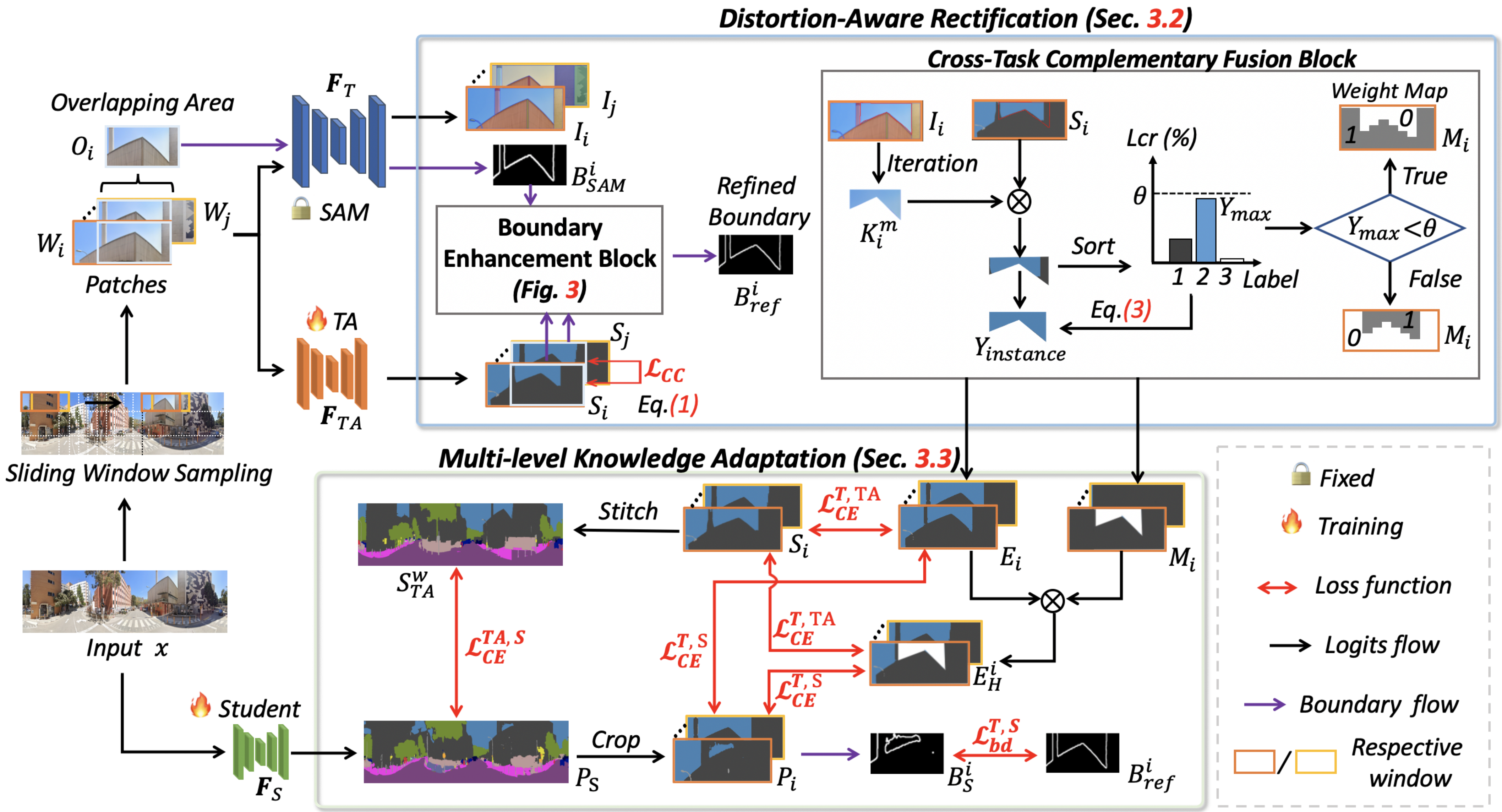
Our Key Idea



Leveraging **instance masks and boundary information** provided by SAM, coupled with **segmentation logits from the teacher assistant**, to obtain reliable ensemble logits for knowledge adaptation to our student.

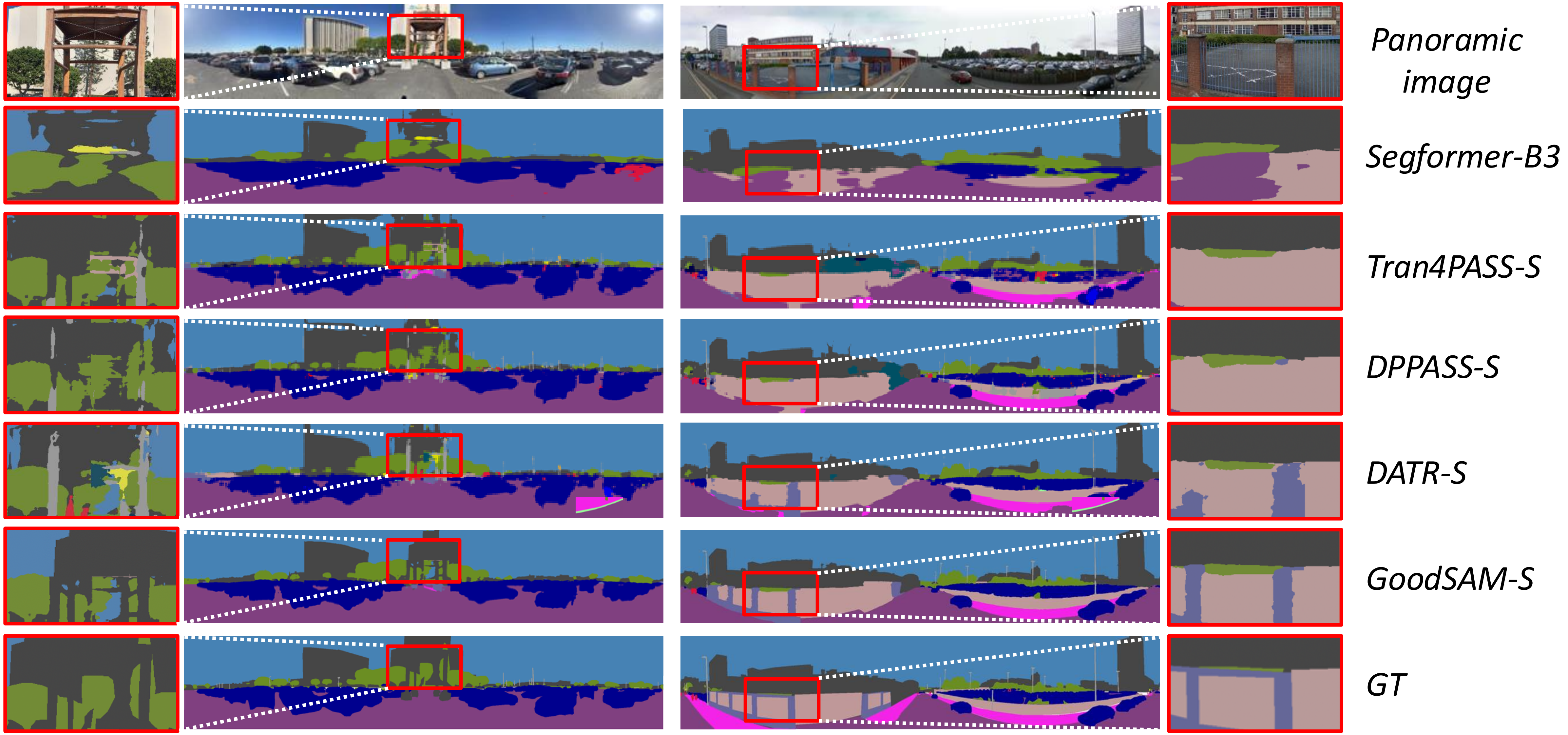
Zhang et al, GoodSAM: Bridging Domain and Capacity Gaps via Segment Anything Model for Distortion-aware Panoramic Semantic Segmentation, CVPR, 2025.

Two Technical Contributions



Zhang et al, GoodSAM: Bridging Domain and Capacity Gaps via Segment Anything Model for Distortion-aware Panoramic Semantic Segmentation, CVPR, 2025.

Visual Comparison



Zhang et al, GoodSAM: Bridging Domain and Capacity Gaps via Segment Anything Model for Distortion-aware Panoramic Semantic Segmentation, CVPR, 2025.

Quantitative results

Method	P. (M)	mIoU	Road	S.W.	Build.	Wall	Fence	Pole	Tr.L.	Tr.S.	Veget.	Terr.	Sky	Person	Rider	Car	Truck	Bus	Train	M.C.	B.C.
ERFNet [16]	-	16.65	63.59	18.22	47.01	9.45	12.79	17.00	8.12	6.41	34.24	10.15	18.43	4.96	2.31	46.03	3.19	0.59	0.00	8.30	5.55
PASS(ERFNet) [28]	-	23.66	67.84	28.75	59.69	19.96	29.41	8.26	4.54	8.07	64.96	13.75	33.50	12.87	3.17	48.26	2.17	0.82	0.29	23.76	19.46
Omni-sup(ECANet) [30]	-	43.02	81.60	19.46	81.00	32.02	39.47	25.54	3.85	17.38	79.01	39.75	94.60	46.39	12.98	81.96	49.25	28.29	0.00	55.36	29.47
P2PDA(Adversarial) [36]	-	41.99	70.21	30.24	78.44	26.72	28.44	14.02	11.67	5.79	68.54	38.20	85.97	28.14	0.00	70.36	60.49	38.90	77.80	39.85	24.02
PCS [33]	25.56	53.83	78.10	46.24	86.24	30.33	45.78	34.04	22.74	13.00	79.98	33.07	93.44	47.69	22.53	79.20	61.59	67.09	83.26	58.68	39.80
Trans4PASS-T [37]	13.95	53.18	78.13	41.19	85.93	29.88	37.02	32.54	21.59	18.94	78.67	**45.20**	93.88	48.54	16.91	79.58	65.33	55.76	84.63	59.05	37.61
Trans4PASS-S [37]	24.98	55.22	78.38	41.58	86.48	31.54	45.54	33.92	22.96	18.27	79.40	41.07	93.82	48.85	23.36	81.02	67.31	**69.53**	86.13	60.85	39.09
DPPASS-T [42]	14.0	55.30	78.74	46.29	87.47	**48.62**	40.47	35.38	24.97	17.39	79.23	40.85	93.49	52.09	29.40	79.19	58.73	47.24	86.48	66.60	38.11
DPPASS-S [42]	25.4	56.28	78.99	48.14	87.63	42.12	44.85	34.95	27.38	19.21	78.55	43.08	92.83	**55.99**	29.10	80.95	61.42	55.68	79.70	70.42	38.40
DATR-M [41]	4.64	52.90	78.71	48.43	86.92	34.92	43.90	33.43	22.39	17.15	78.55	28.38	93.72	52.08	13.24	77.92	56.73	59.53	93.98	51.52	34.06
DATR-T [41]	14.72	54.60	79.43	49.70	87.39	37.91	44.85	35.06	25.16	19.33	78.73	25.75	93.60	53.52	20.20	78.07	60.43	55.82	91.11	67.03	34.32
DATR-S [41]	25.76	56.81	80.63	51.77	87.80	44.94	43.73	**37.23**	25.66	**21.00**	78.61	26.68	93.77	54.62	**29.50**	80.03	67.35	63.75	87.67	67.57	37.10
GoodSAM-M(ours)	3.7	55.93	79.57	51.04	86.24	43.42	44.86	30.92	26.60	20.62	77.79	25.43	92.99	53.77	25.84	82.01	70.94	62.29	91.93	58.24	38.25
GoodSAM-T(ours)	14.0	58.21	80.06	**53.29**	89.75	44.91	46.98	31.13	27.81	19.83	79.58	25.72	93.81	55.44	26.99	84.54	73.07	68.41	93.99	67.36	43.39
GoodSAM-S(ours)	25.4	**60.56**	**80.98**	52.96	**93.22**	48.17	**51.28**	33.51	**28.09**	20.15	**81.64**	30.97	**95.21**	55.13	29.01	**87.89**	**75.28**	69.37	**94.98**	**73.28**	**49.64**

- Significant boost compared with existing methods.
- Generalization is not good in outdoor scenes.

Zhang et al, GoodSAM: Bridging Domain and Capacity Gaps via Segment Anything Model for Distortion-aware Panoramic Semantic Segmentation, CVPR, 2025.

Towards 360 Foundation Models for Segmentation!

Data and Geneeralization are always a challenge!
(NeurIPS 2025, ICCV 2025)

Data is always a challenge

The scale and diversity of existing 360 video datasets remain limited due to:

1. **Long time to annotate because of large field of view.**
2. **Distortion and object deformation, as shown in (a).**
3. **Long time to annotate objects crossing border, as shown in (b).**

(a)

(b)

Zhang et al, Leader360V: The Large-scale, Real-world 360 Video Dataset for Multi-task Learning in Diverse Environment, NeuriPS, 2025.

Our Dataset (Leader360V)- NeurIPS 2025

Large scale (10K+)

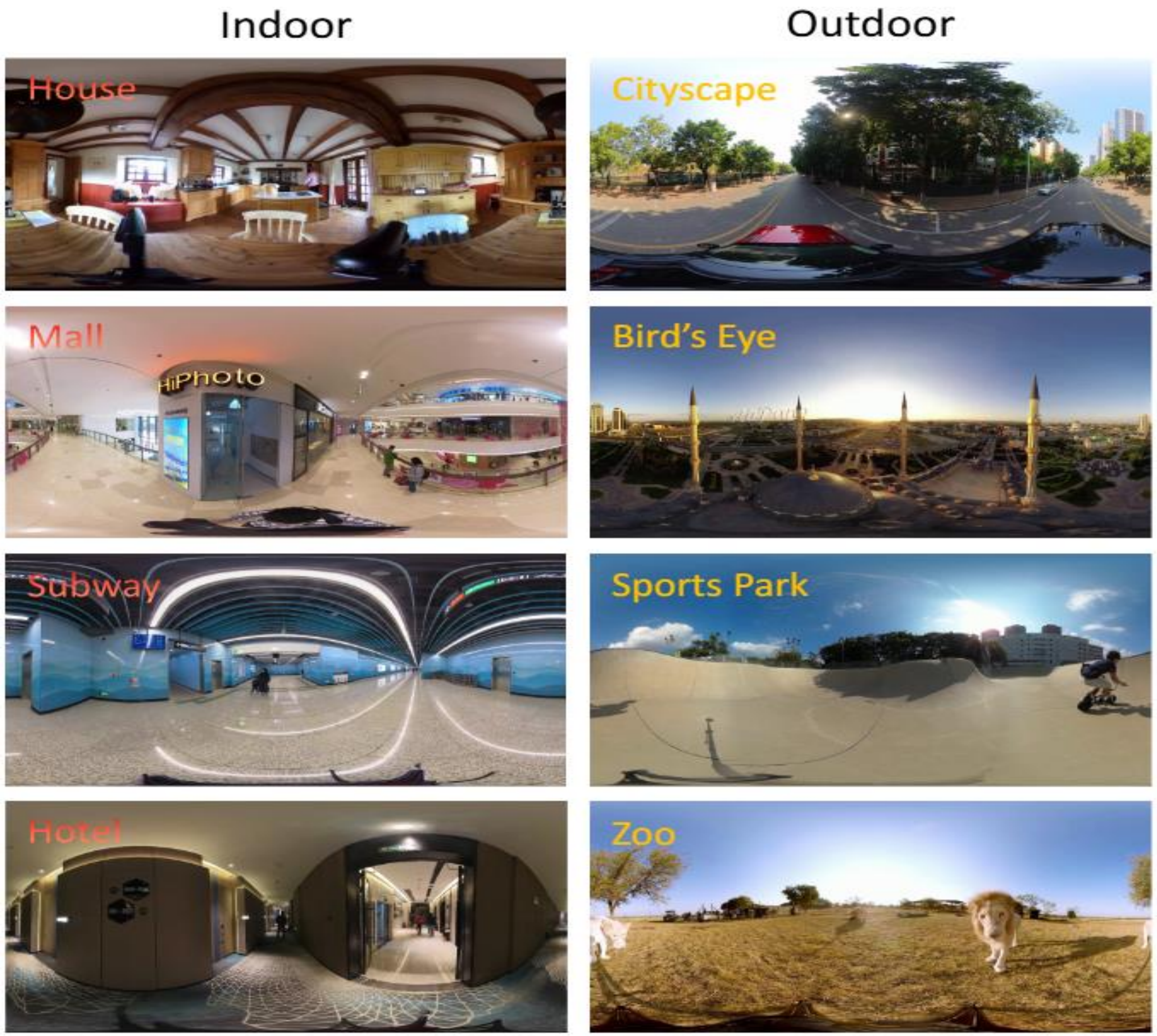
Zhang et al, Leader360V: The Large-scale, Real-world 360 Video Dataset for Multi-task Learning in Diverse Environment, NeurIPS, 2025.

Our Dataset (Leader360V)

Real-world data

High scene diversity

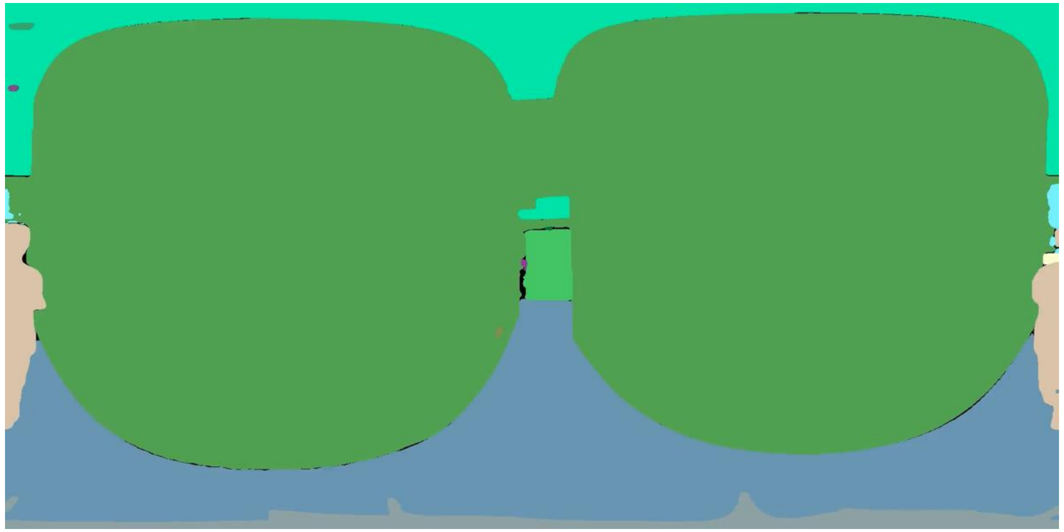
Covering 198 object types



Zhang et al, Leader360V: The Large-scale, Real-world 360 Video Dataset for Multi-task Learning in Diverse Environment, NeurIPS, 2025.

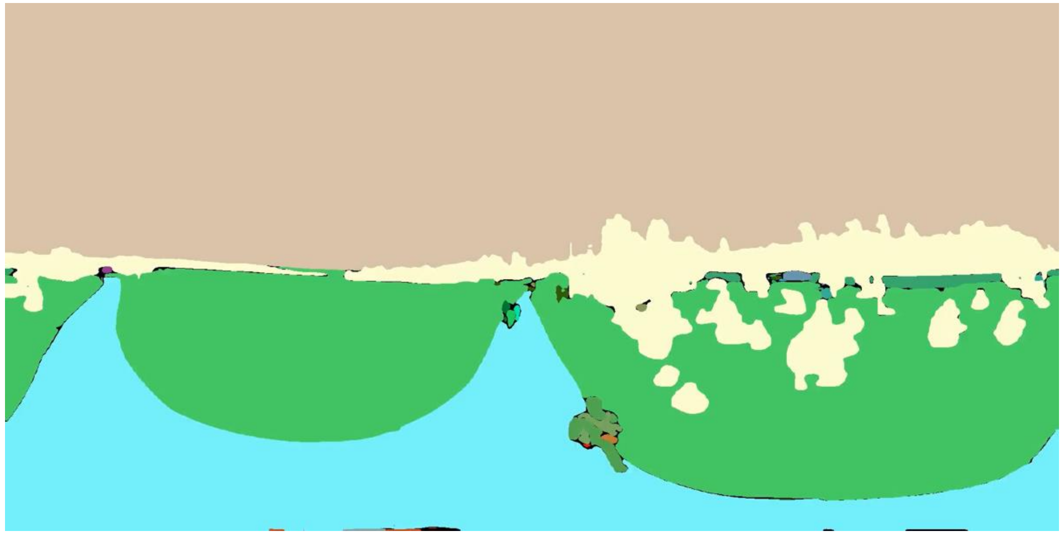
Video Samples

Video Samples



Raw video

Annotation



Zhang et al, Leader360V: The Large-scale, Real-world 360 Video Dataset for Multi-task Learning in Diverse Environment, NeuriPS, 2025.

ICCV OCT 19-23, 2025 HONOLULU HAWAII

E-SAM: Training-Free Segment Every Entity Model

Weiming Zhang[1], Dingwen Xiao[1], Lei Chen[1,2], Lin Wang[3]




Thu 13:30- 17:00

Poster #7238 ☺

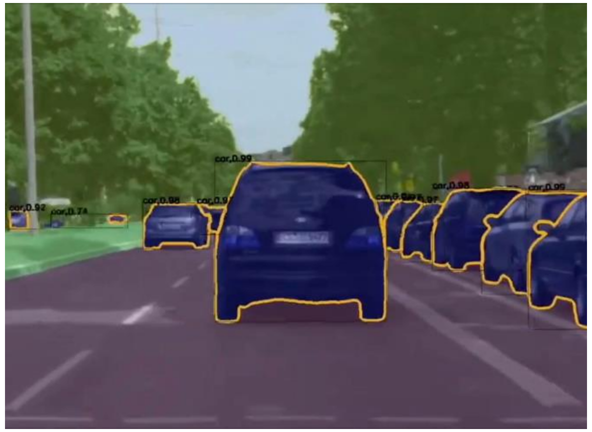
Home Page

Why Entity Segmentation Matters?

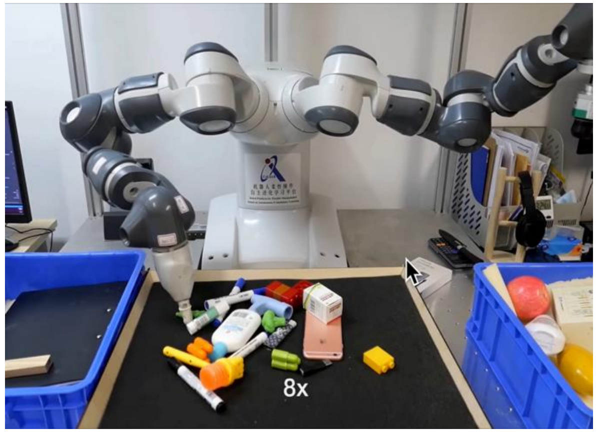
- ❑ The class-agnostic nature makes ES well-suited to various open-world applications:

Video Tracking

Image Inpainting



Autonomous Driving



Robotics Perception

Entity Segmentation provides rich mask proposals for various open-world applications.

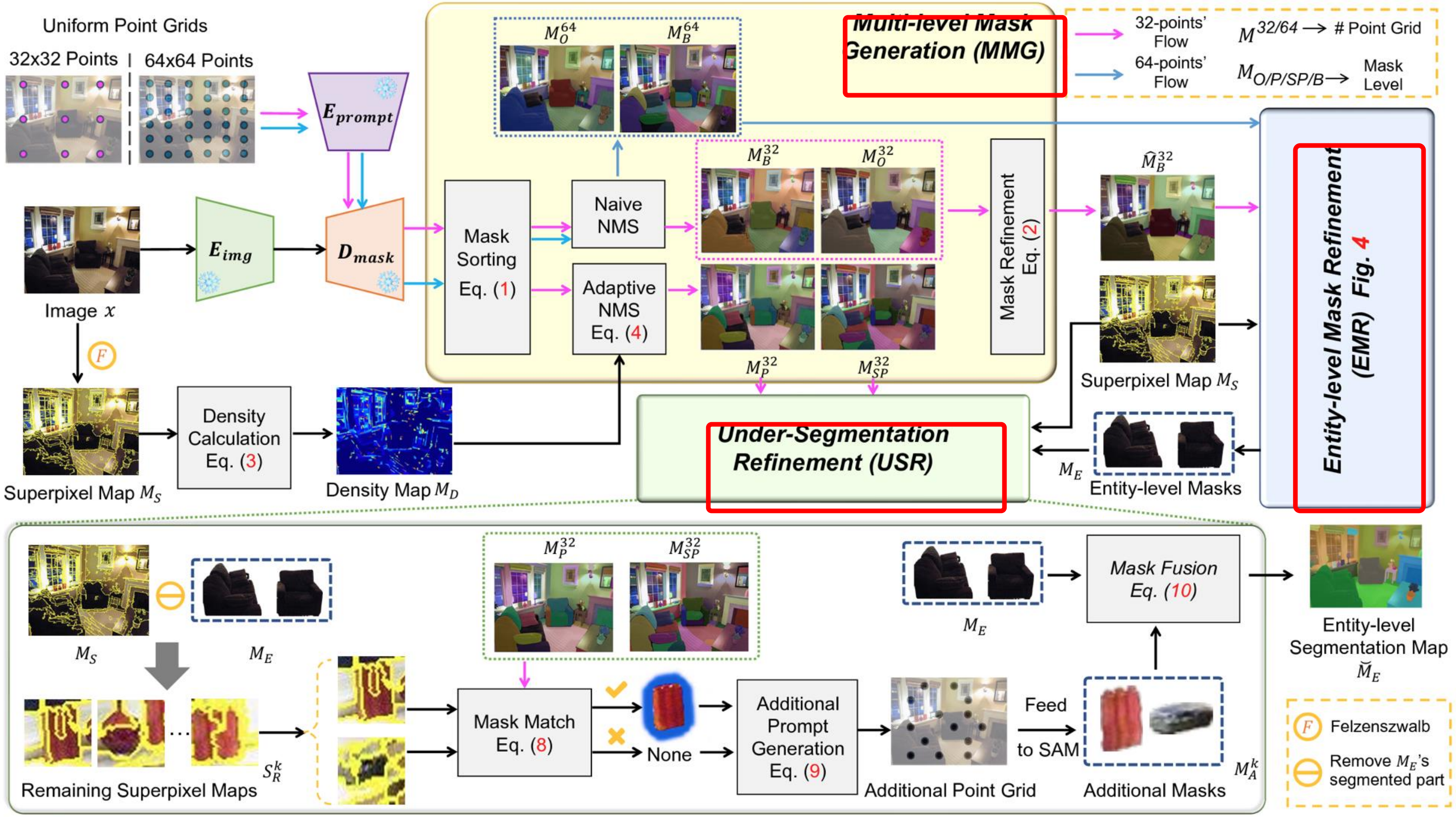
The left video demo is adapted from Track-Anything: Segment and Track Anything in Videos [Gao et al., 2023], available at https://github.com/gaomingqi/Track-Anything.
The middle video fragment is taken from the demo video available at YouTube (https://www.youtube.com/watch?v=cC6lR7ScecU).
The right video segment is adapted from the official Track-Anything demo available at https://www.youtube.com/watch?v=GIXs6TAaPM8.

Key Question & Idea

How can we efficiently and effectively achieve ES of all entities in an image?

Key Idea

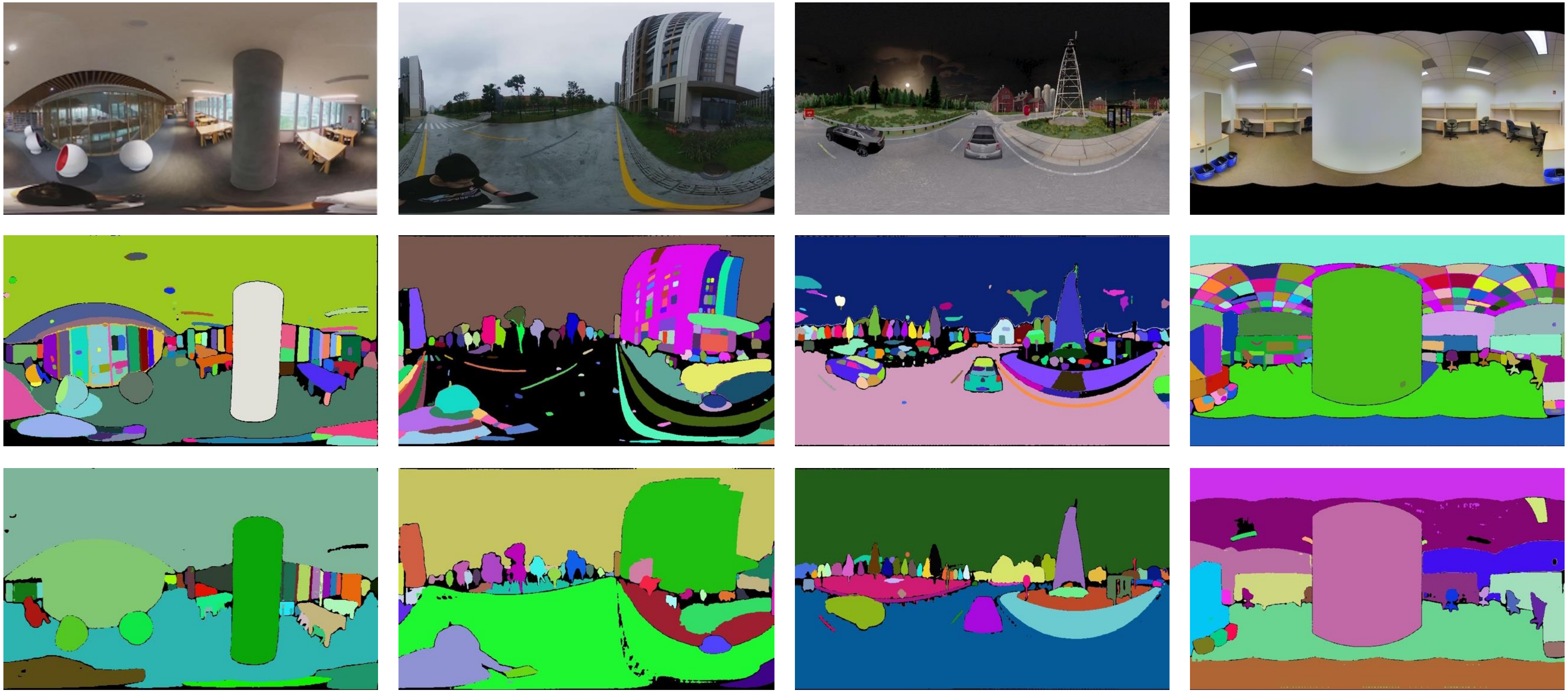
Hierarchical self-refining masking mechanism that transforms SAM's coarse, multi-granular masks into clean, entity-level segmentations — **all in a training-free manner**.



Zhang et al, E-SAM: Training-Free Segment Every Entity Model, ICCV, 2025.

Visual Results from Leader360V Dataset

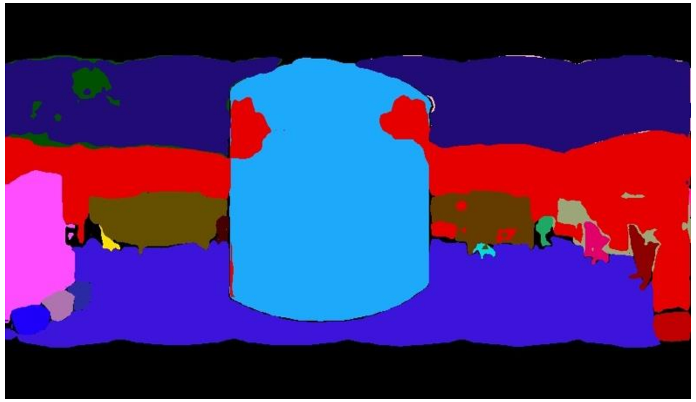
Image
SAM AMG
E-SAM (Ours)



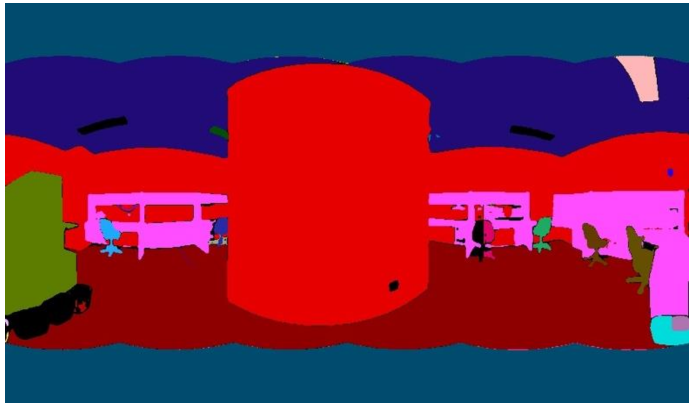
More Open-World Examples:

Image

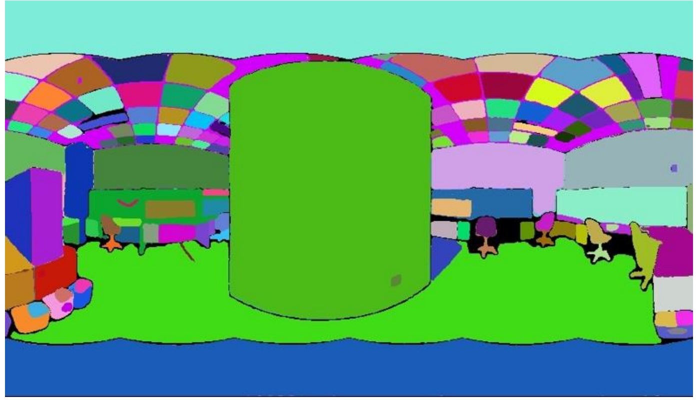
Entity Framework



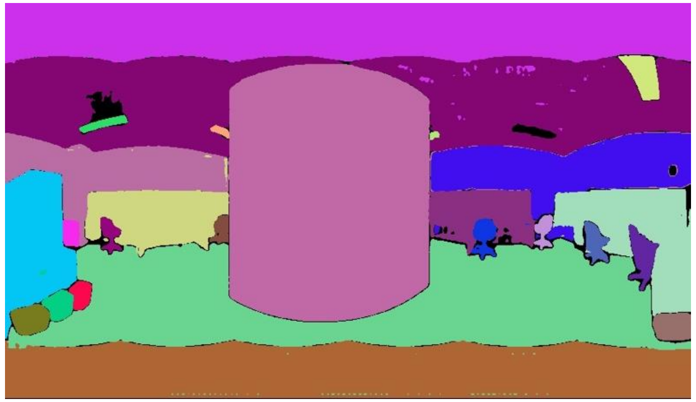
CropFormer



SAM



E-SAM



Zhang et al, E-SAM: Training-Free Segment Every Entity Model, ICCV, 2025.

Takeaways and Some of My Thoughts

Takeaways

1

360 cameras more advantageous in their application potential.

- 360 data are with multi-projection properties with larger field-of-view (FoV).
- By principle, one 360 camera can cover the whole scene.

2

Projection fusion is a way to address distortion and learn complete visual info

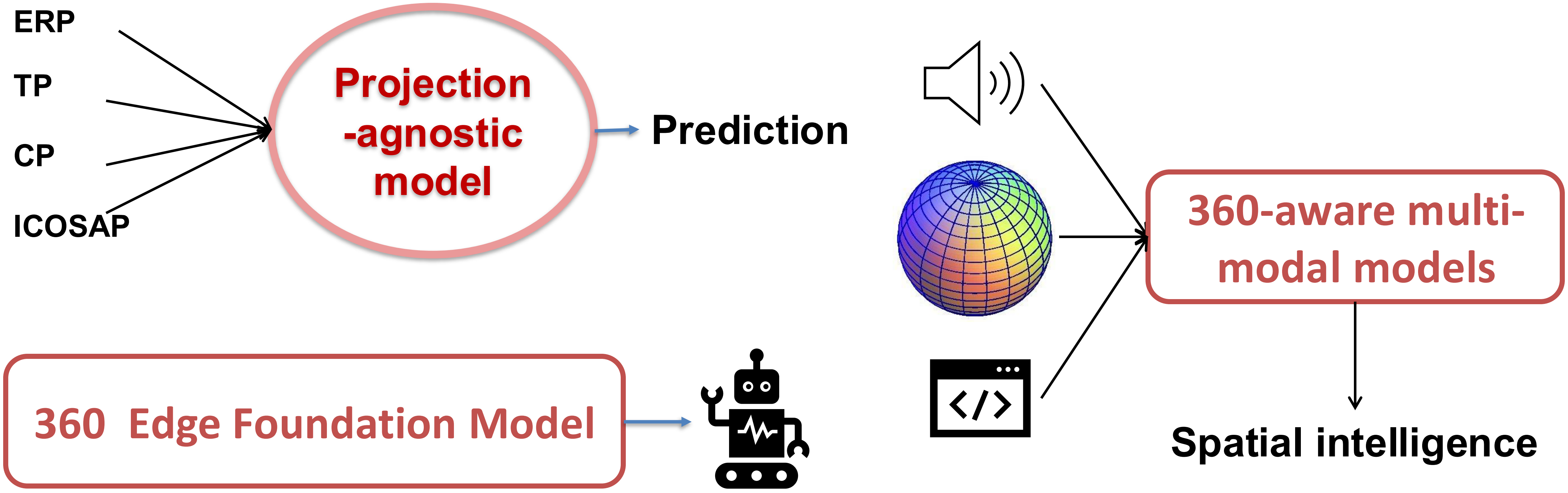
- Uni-projection is trapped by distortion, disconnection, and overlap issues.
- Fusion has to deal with geometric and semantic mismatch issues.

3

Handling data shortage is crucial; transfer learning is important.

- Domain adaptation or knowledge distillation are beneficial to overcome data issues.
- Learning scalable 360 foundation models needs real-world data.

Future Directions



linwang@ntu.edu.sg
www.linkedin/in/addison-lin-wang-62542b222/

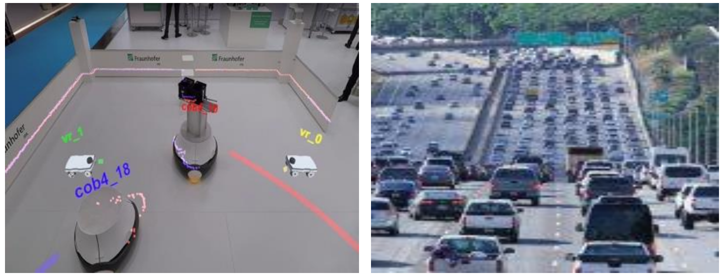
Thank you!
Q/A

Distortion issues are still

Challenges for Embodied Intelligent Systems

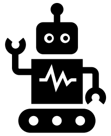
Adverse Visual Conditions

Multi-source Heterogenous Data



Dynamically Varying Env.

External Challenges



Internal Challenges

Adverse Visual Conditions

Dynamic range/ Noise

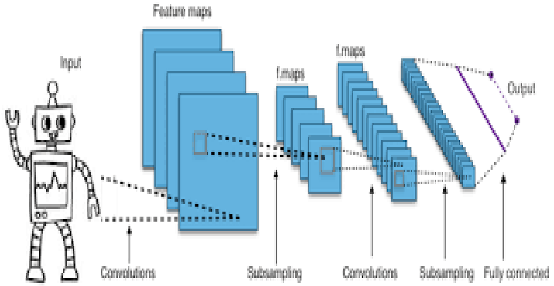
Unideal Sensor Match

Complex Open-world Env.

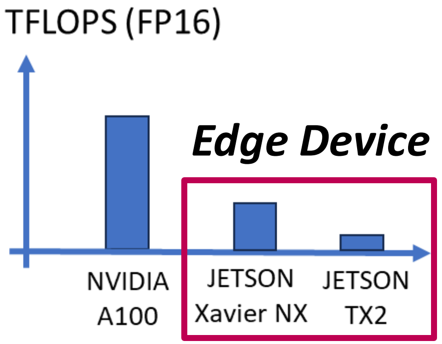
Normal Condi tion

Adverse Co ndition

Cumbersome Model Size



Limited Computation

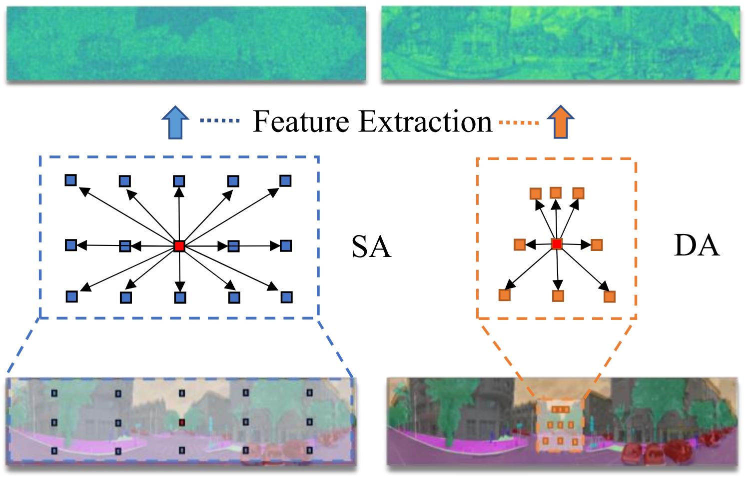


Distortion Matters

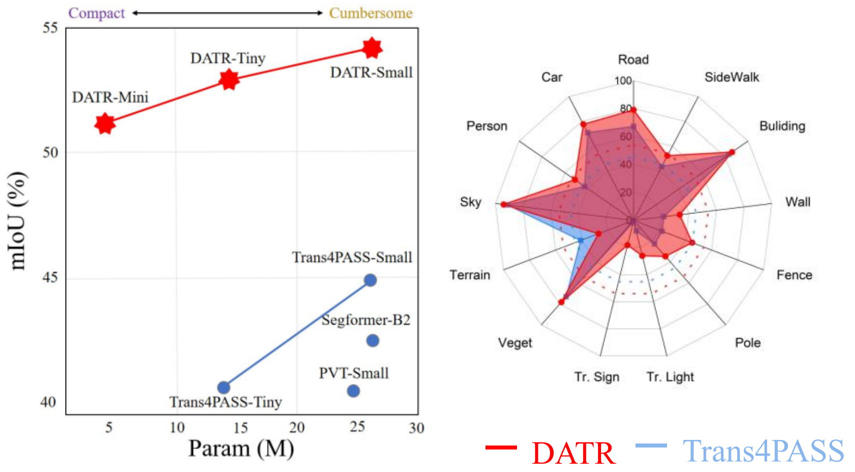
Efficient backbone and algorithm for UDA in Panoramic Semantic Segmentation

We find that the pixels' neighboring region of ERP indeed introduce less distortion

Distortion-aware Attention (DA) & Class-wise Feature Aggregation (CFA)



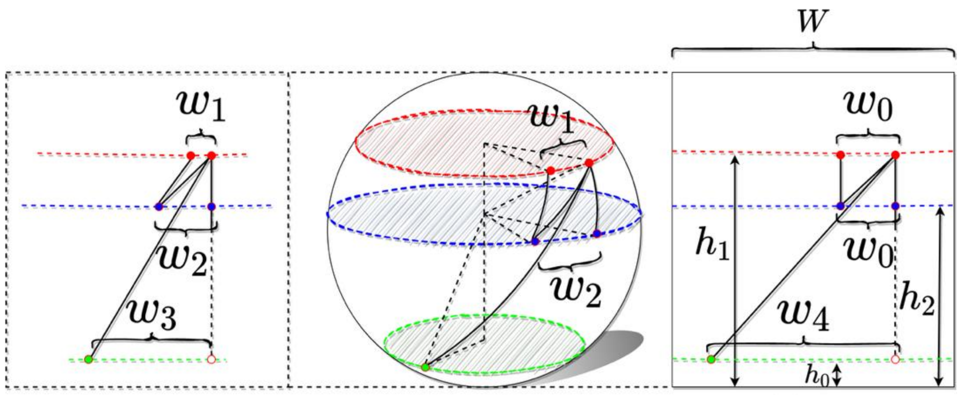
(a)



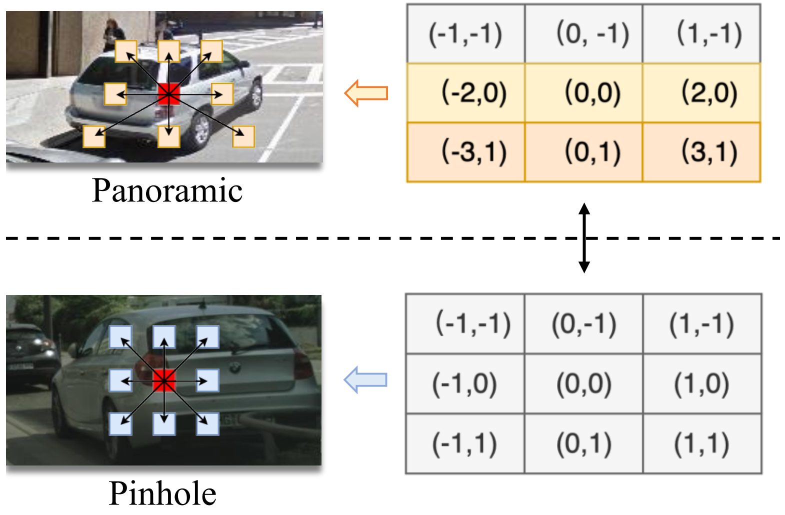
(b)

It is challenging to address distortion issues

Distortion-aware Attention: Why we focus on the neighboring region?



(a) DLP (b) Sphere (c) ERP

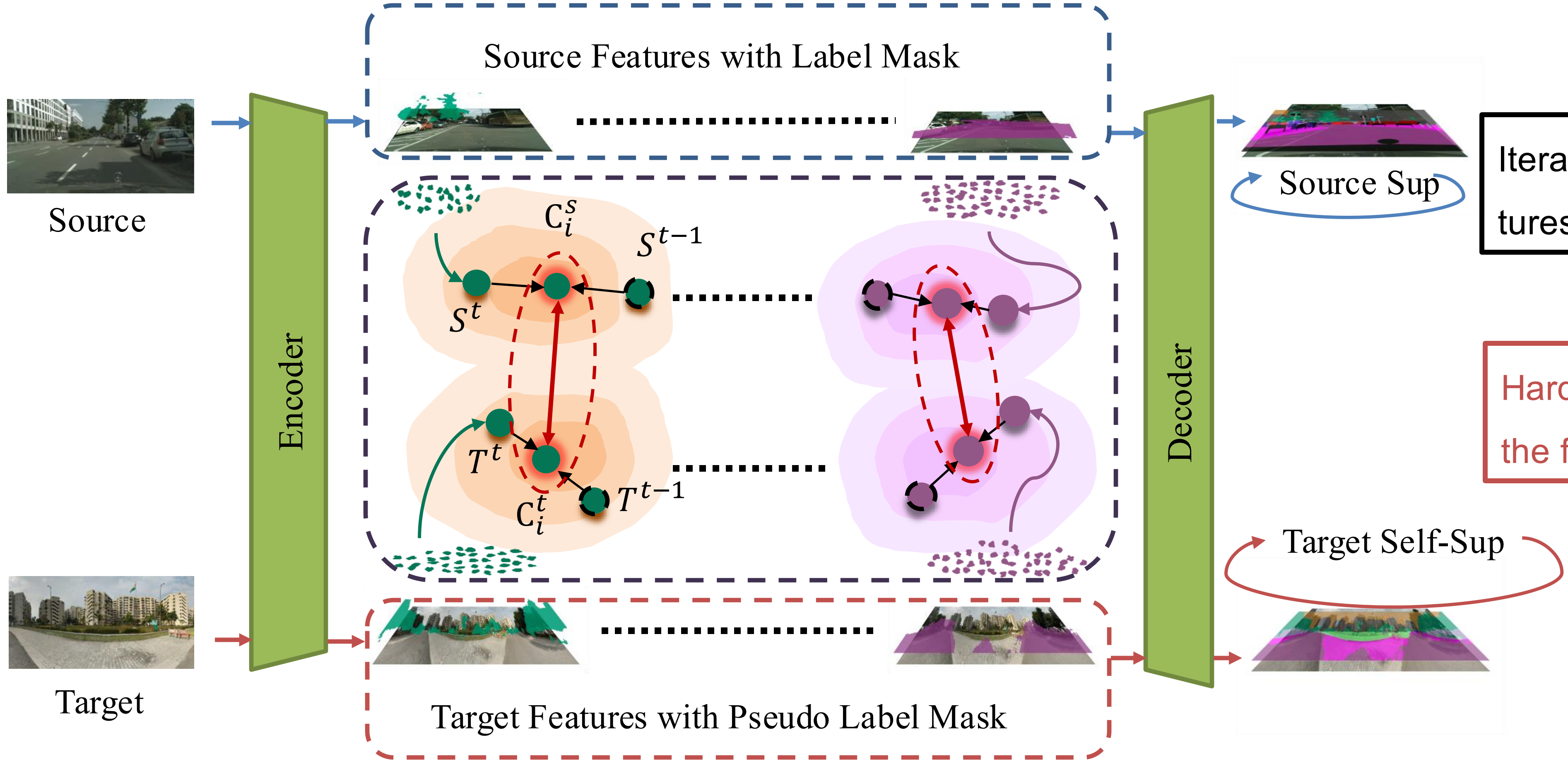


It is easier to capture the positional distribution among the pixels by reducing the receptive field.

This Relative Positional Encoding captures the distribution of different neighboring pixels.

What is challenging?

Class-wise Feature Aggregation: Why class-wise feature aggregation?



Iteratively aggregate class-wise fea tures and updates feature centers.

Hard pseudo labels are softened in the feature space.