JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Both Style and Distortion Matter: Dual-Path Unsupervised Domain Adaptation for Panoramic Semantic Segmentation

Xu Zheng[2], Jinjing Zhu[1], Yexin Liu[1], Zidong Cao[1], Chong Fu[2,4], Lin Wang[1,3]

[1] VLISLab, AI Thrust, HKUST(GZ) [2] Northeastern University, [3] Dept. of CSE, HKUST,

[4]Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, NEU, China

VLIS LAB

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

東北大學 Northeastern University

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Quick Preview

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU) · Northeastern University · VLIS LAB · THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Both Style and Distortion Matter: Dual-Path Unsupervised Domain Adaptation for Panoramic Semantic Segmentation

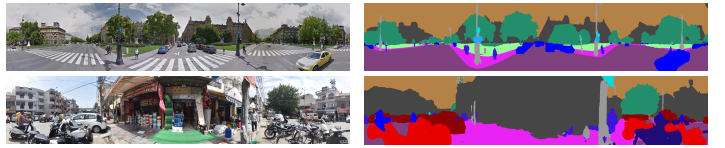
Xu Zheng[2], Jinjing Zhu[1], Yexin Liu[1], Zidong Cao[1], Chong Fu[2,4], Lin Wang[1,3]

[1] VLISLab, AI Thrust, HKUST(GZ) [2] Northeastern University, [3] Dept. of CSE, HKUST, [4]KLICMI, NEU

GitHub Page: https://https://vlis2022.github.io/cvpr23/DPPASS/

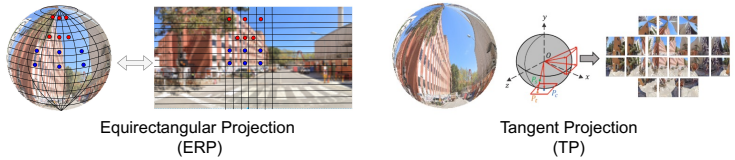
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Why we need Panoramic Semantic Segmentation (PSS)?



Captured with 360 camera *Segmentation map estimated by our DPPASS*

- **Large field-of-view of 360° x 180°**
- **Strong interaction and realism**



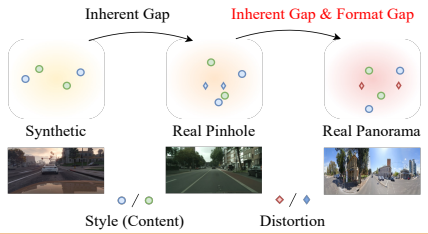
- **Multiple projection types provide more potential of processing**

What makes it difficulty in PSS?

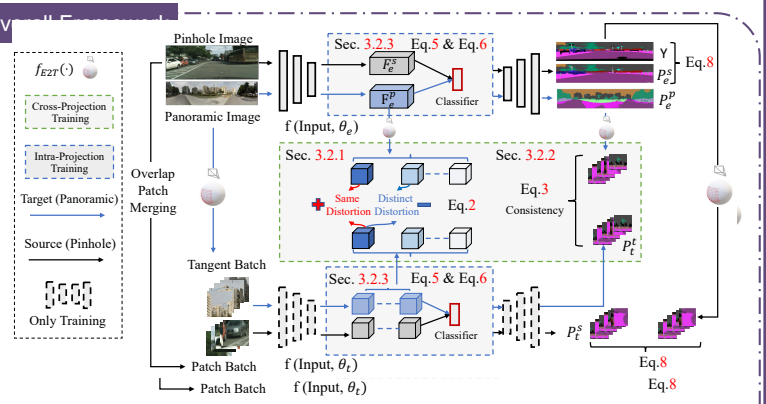
1. Pixel-wise annotated panoramic datasets are scarce;
2. Specifically designed networks are less generalizable to other spherical image data.

What makes it difficulty in UDA for PSS?

1. Inherent Gap: Diverse camera sensors and captured scenes;
2. Format Gap: Distinct image representation formats (ERP and pinhole images).



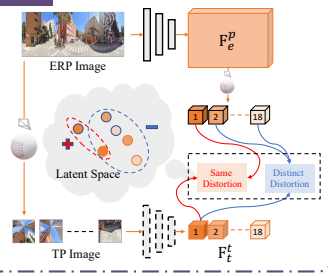
Methodology



1. In the cross-projection training, the prediction consistency training is proposed to fully utilize the distinct characteristics of tangent images to reduce the format gap.;
2. In the feature-level cross-projection training, tangent-wise feature contrastive training (TFCT) is proposed to align the tangent-wise high-level features to reduce the format gap.

Tangent-wise Feature Contrastive Training (TFCT)

1. We explore to impose consistency between the ERP and TP paths at the feature map level to diminish the format gap ;
2. The feature (cube), only the ones in the same color having the same projection locations are the positive example pairs, and vice versa.

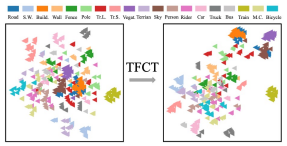


Results

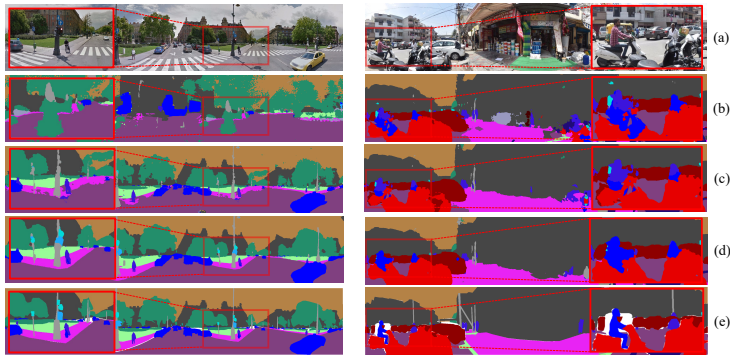
DPPASS *vs.* SOTA methods

Method	mIoU	road	sidewalk	building	wall	fense	pole	traffic Light	traffic Sign	vegetation	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	59.60	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

Ablation Study



Losses				mIoU	△
$\mathcal{L}_s$	$\mathcal{L}_g$	$\mathcal{L}_{pc}$	$\mathcal{L}_{fc}$		
✓				42.40	-
✓	✓			50.12	+7.72
✓		✓		49.53	+7.13
✓			✓	47.30	+4.90
✓	✓	✓	✓	55.30	+12.9



[1] Zhang et al. OmniDepth: Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. CVPR 2-22

Background

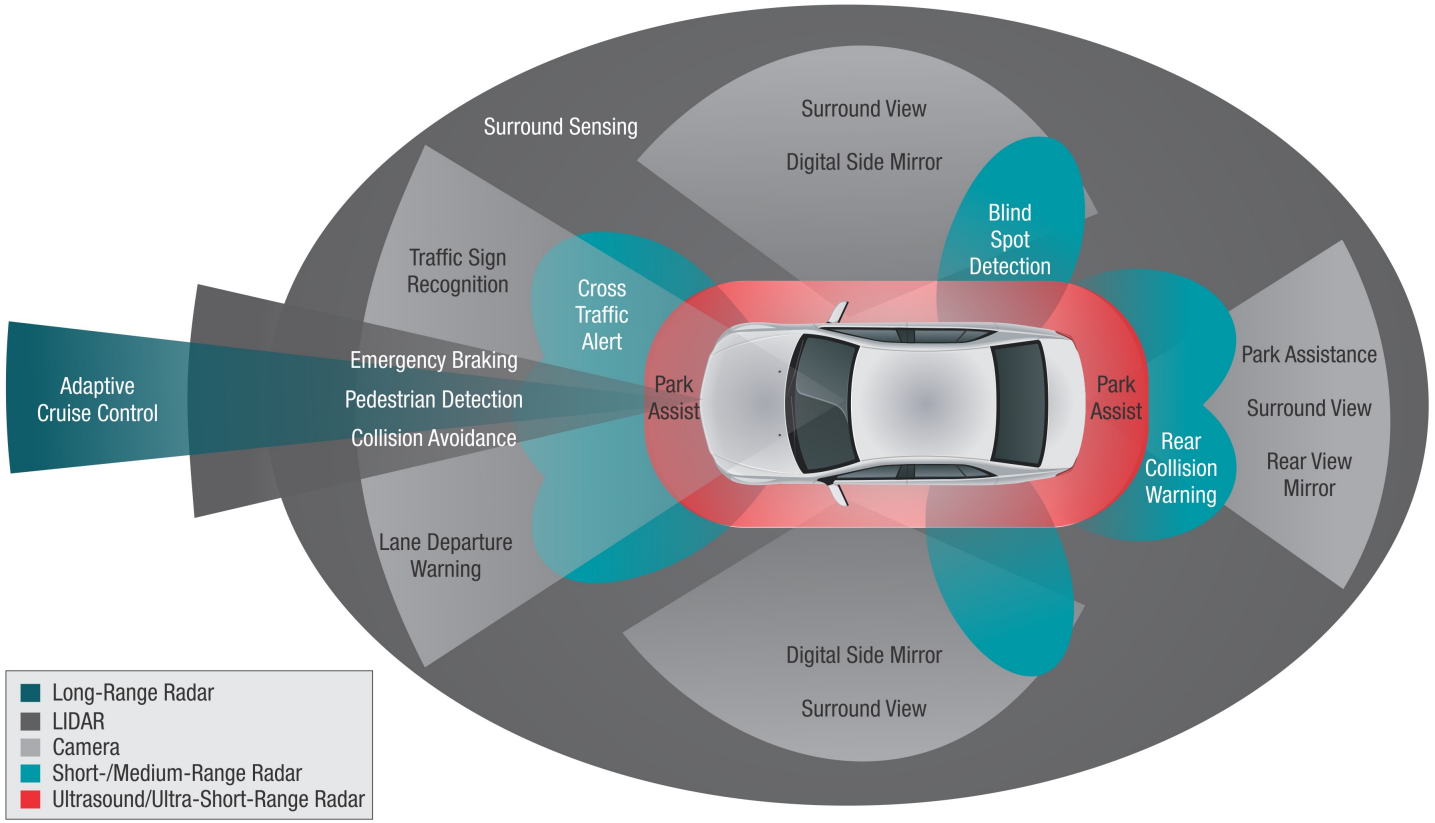
Autonomous driving [1]

Scene Understanding

Scene understanding allows the vehicle to detect and track objects, estimate their distance and speed, and predict their behavior to make informed decisions.

[1] https://www.mi.com.

Background



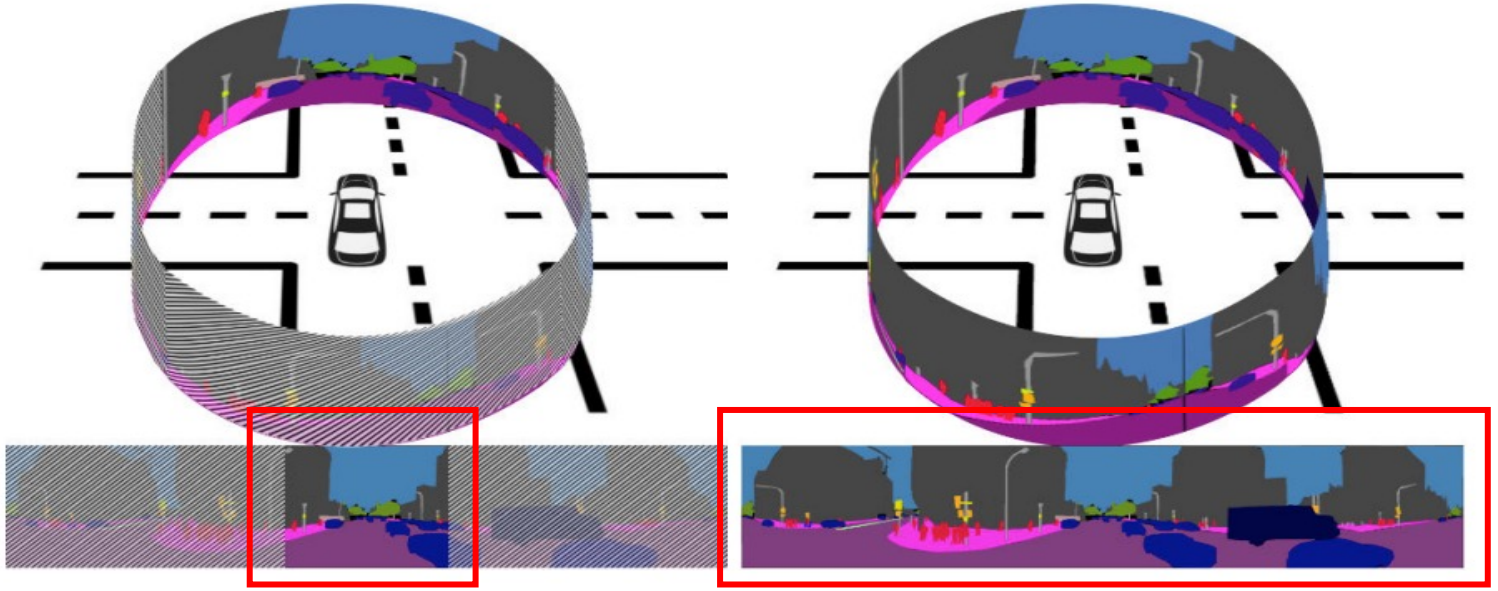
- High cost
- Modality Fusion
- …

Multi–sensors syetem for omnidirectional perception [2]

[2] https://www.elecfans.com/techweek/442327.html

Background

Perspective vs. 360° $camera^{[3]}$



Perspective vs. 360° $camera^{[3]}$

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

[3] Zhang J et al. Transfer Beyond the Field of View: Dense Panoramic Semantic Segmentation via Unsupervised Domain Adaptation[J]. IEEE TITS, 2022(7):23.

Background

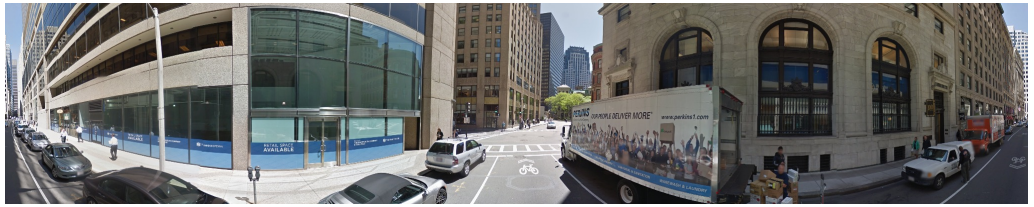
Omnidirectional Scene Perception Abilities[4]

[4] Ma C, et al. DensePASS: Dense Panoramic Semantic Segmentation via Unsupervised Domain Adaptation with Attention-Augmented Context Exchange[J]. 2021.

Background

Pinhole image

Panoramic image



Limited FoV

No Distortion

Sufficient Labels

Broader / 360 FoV

Severe Distortion

Scarce Labels

Unsupervised Domain Adaptation (UDA)

Background

Synthetic

(GTA5)

Real Pinhole

(Cityscapes)

Real Panorama

(DensePASS)

Domain Gaps: Style

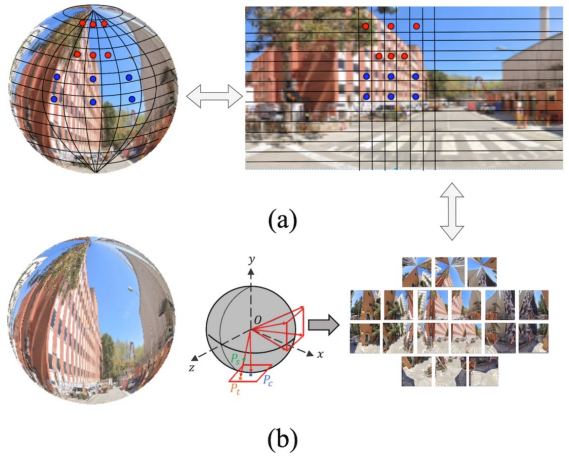
Domain Gaps: Style & Distortion

Domain Gaps

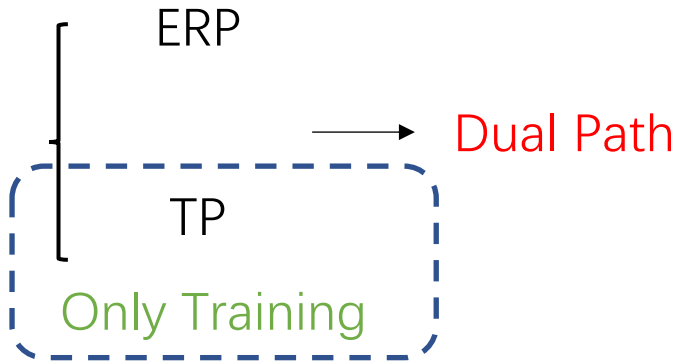
- Inherent Gaps: Diverse camera sensors and captured scenes
- Format Gaps: Distinct image representation formats

How to alleviate these domain gaps?

Methodology



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

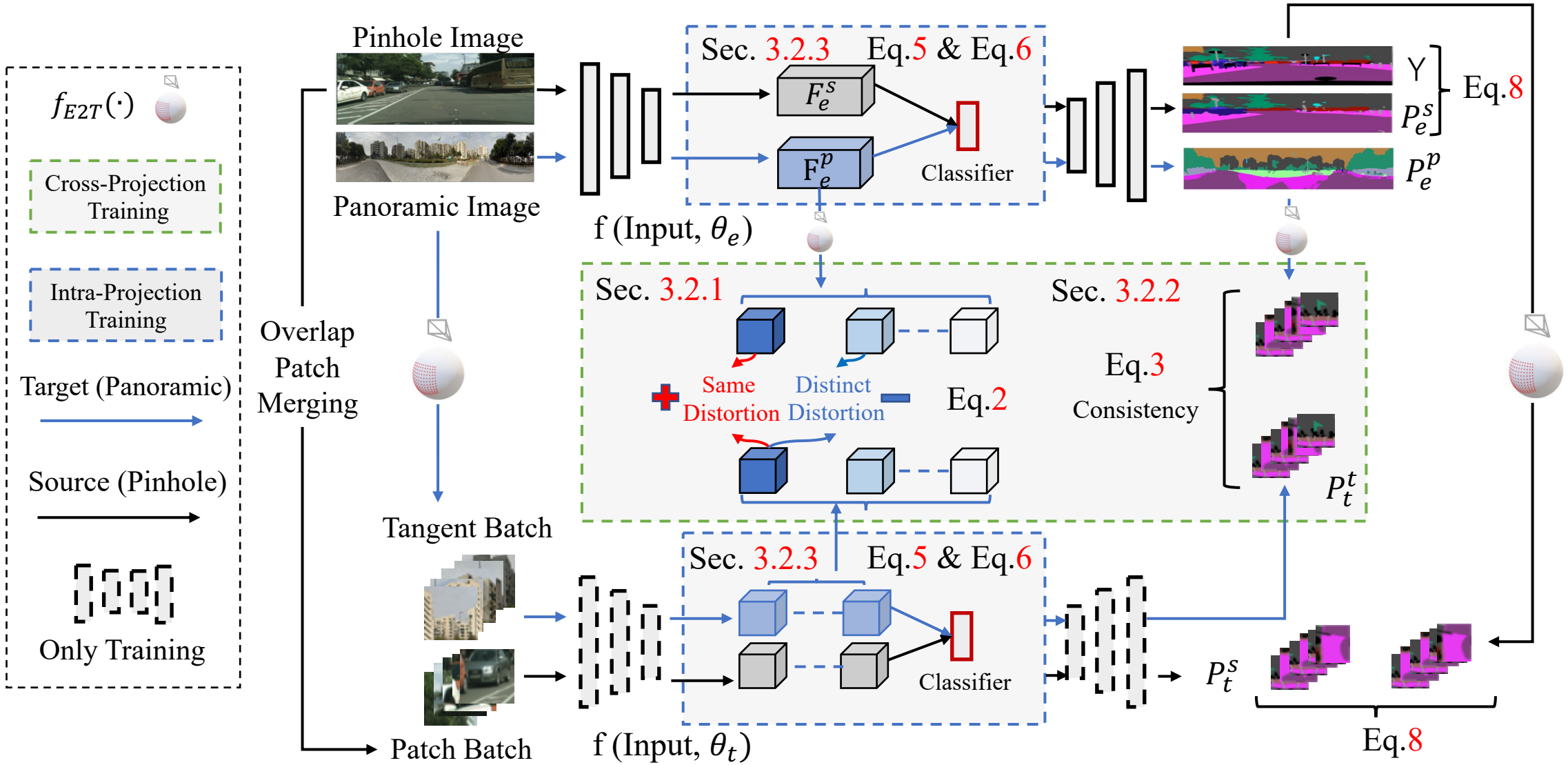


- Dual-Path Framework
  - Cross-Projection Training — ERP & TP
    - Feature level: Tangent-wise feature contrastive training
    - Prediction level: Prediction consistency training
  - Intra-Projection Training
    - ERP path
    - TP path

Overview

Methodology

Overall Framework



Methodology

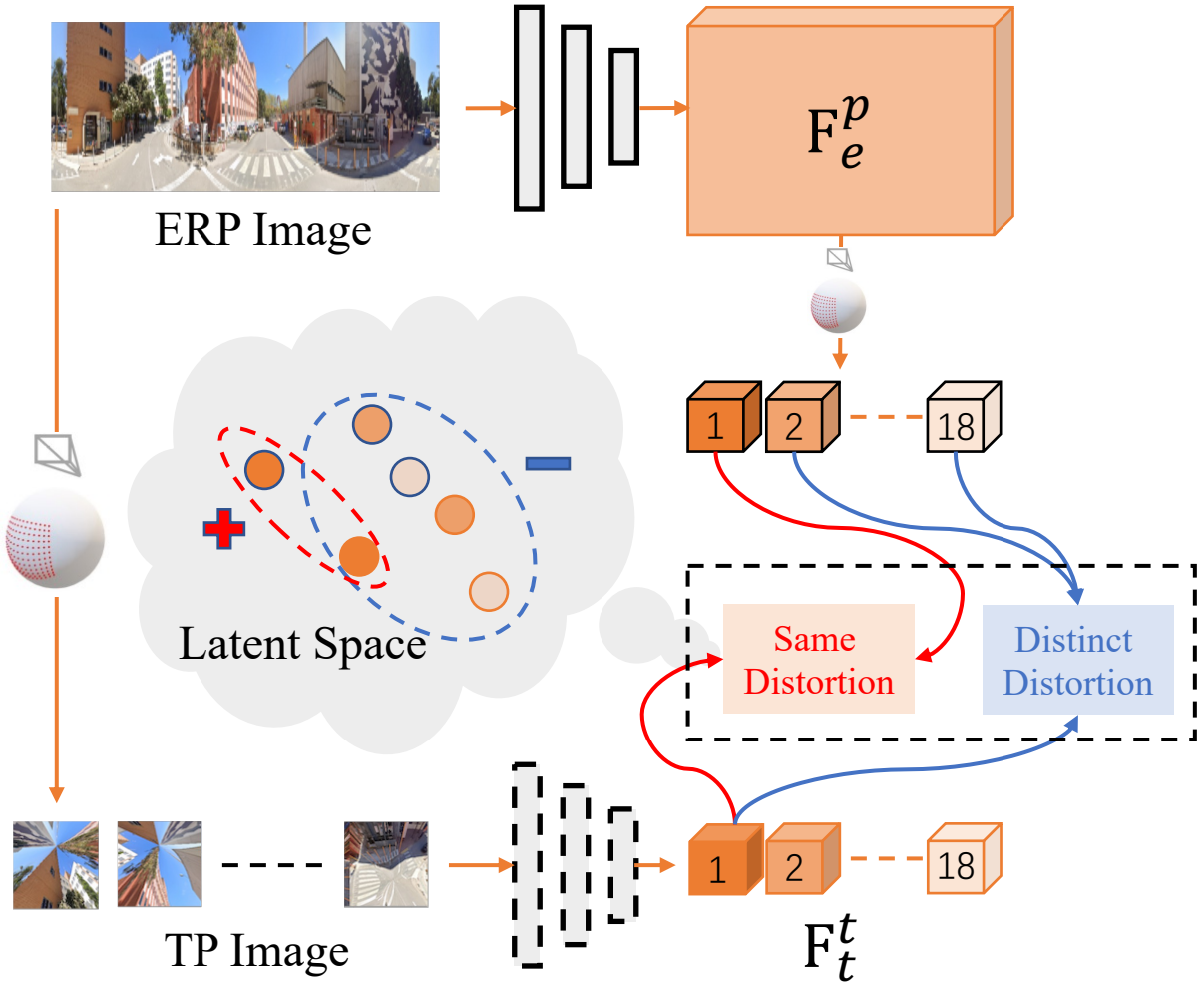
Cross Projection Training:

Tangent-wise feature contrastive training:

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

Prediction consistency training:

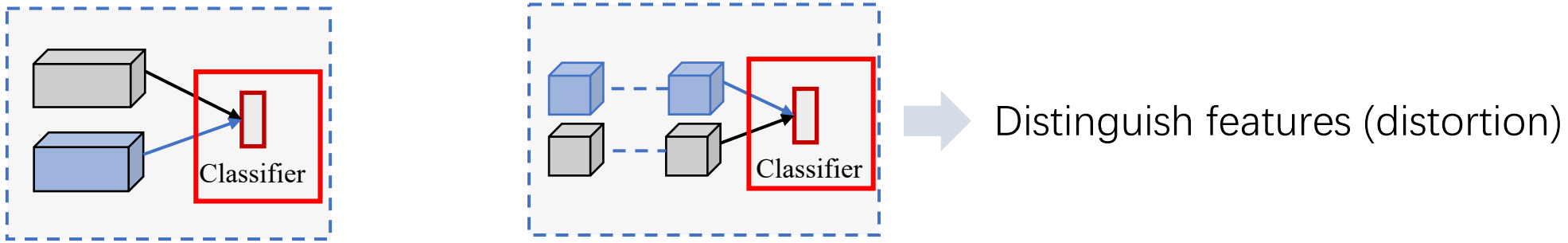
$$\mathcal{L}_{pc} = \sum_{i=1}^{18} f_{E2T}(P_{ei}^p) \log \frac{f_{E2T}(P_{ei}^p))}{P_{ti}^t}$$



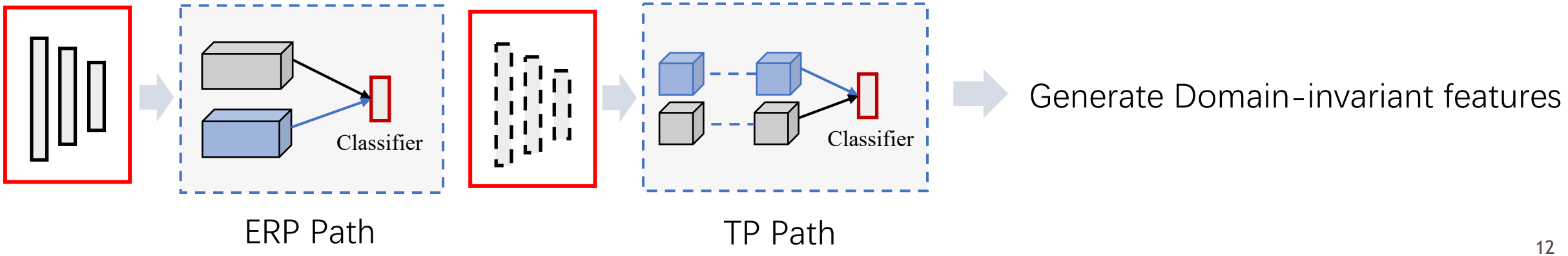
Methodology

Intra Projection Training:

Classifier:



Feature Extractor:



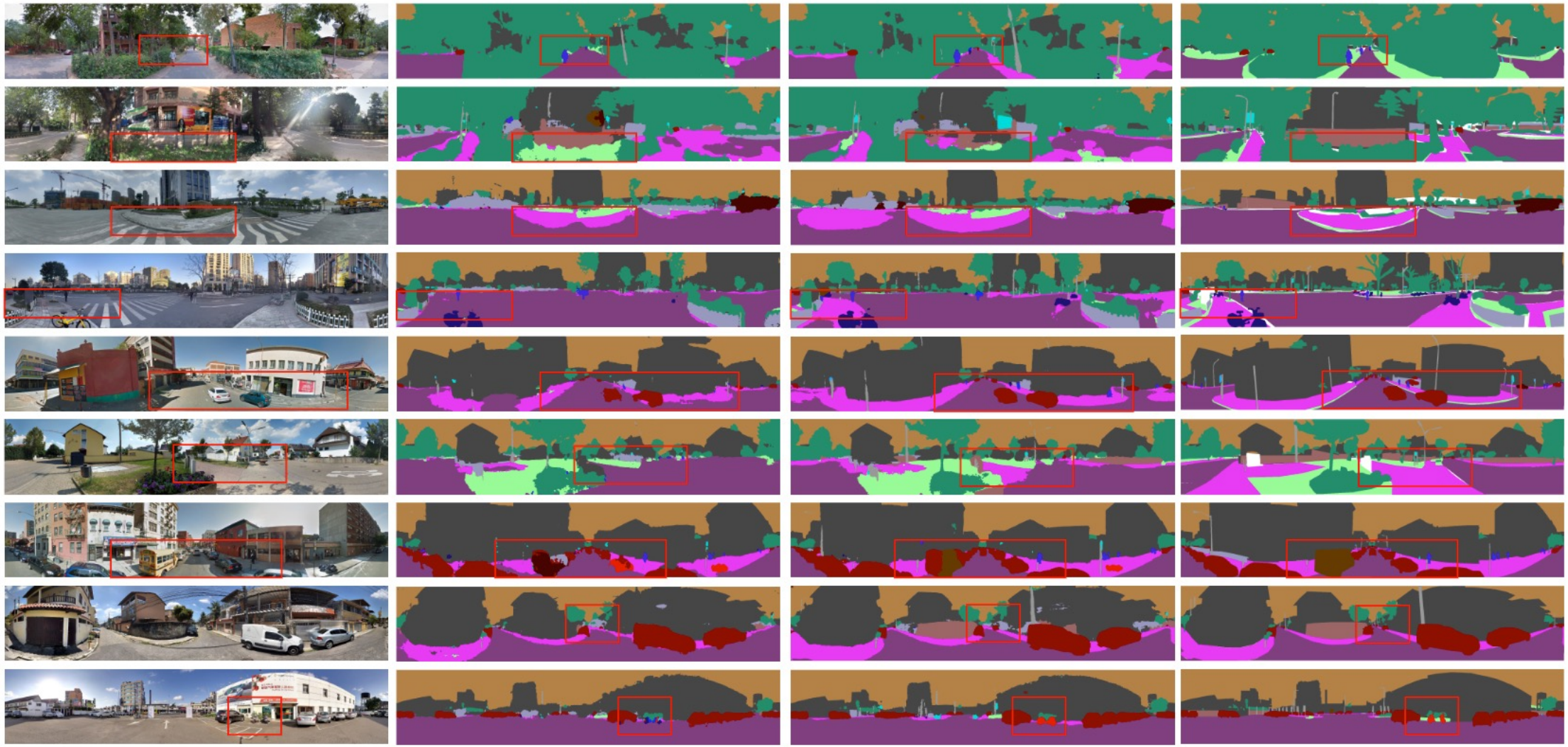
Experiments

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 for autonomous driving

Experiments

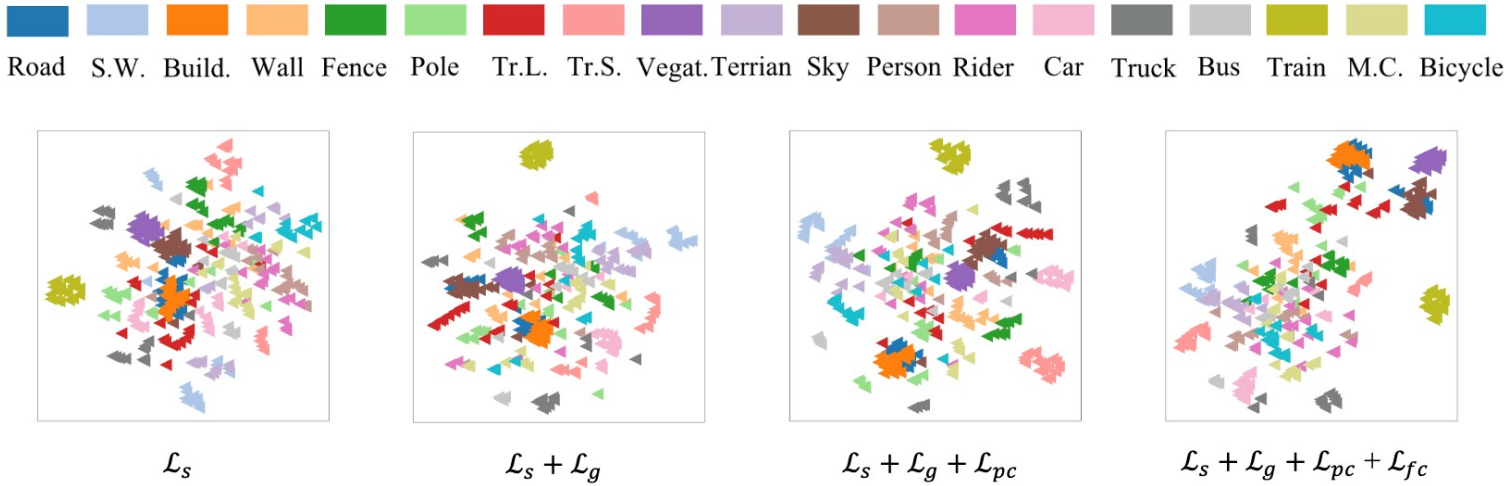


Input | Trans4PASS[5] | Ours | GT

[5] Zhang J, Yang K, Ma C, et al. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation, CVPR. 2022.

Ablation

Loss Combination:



TSNE visualization with different loss combinations.

Tangent Projection Size:

	Tangent Projection				
Size	96 × 96	144 × 144	224 × 224	384 × 384	512 × 512
mIoU	49.98	52.22	**55.30**	55.17	52.56

Not the bigger / smaller the better

Dual Projection:

Dual Projection (49.53%) vs. Single Projection (45.22%)