Generalizable Knowledge Distillation from Vision Foundation Models for Semantic Segmentation

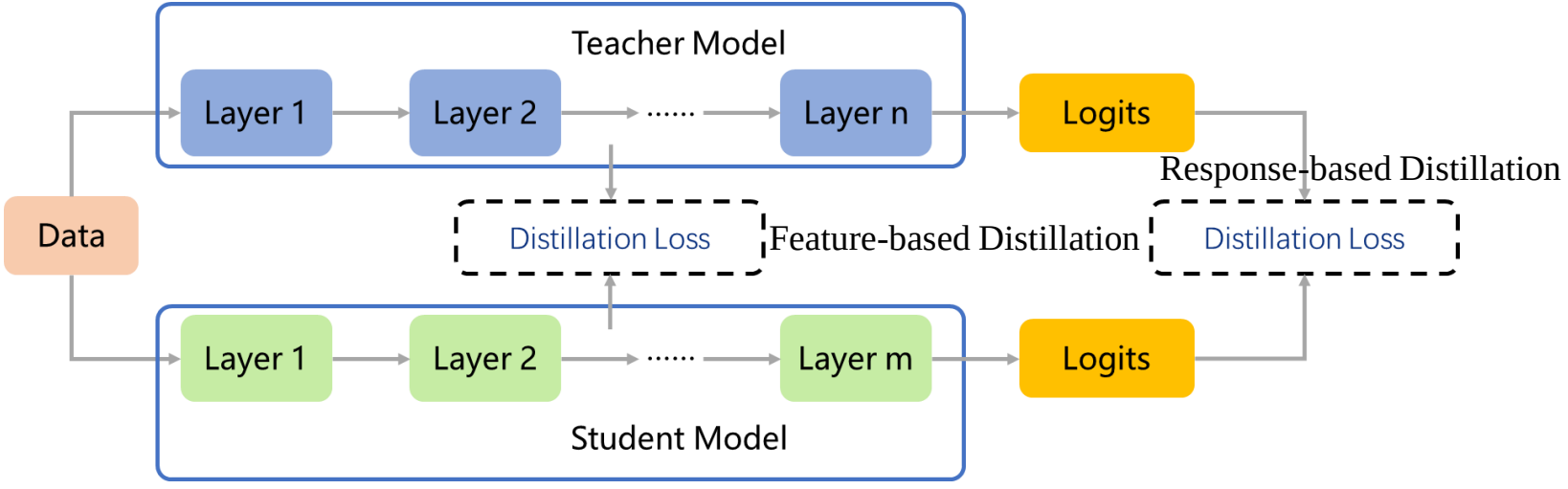
Chonghua Lv[1*], Dong Zhao[2*], Shuang Wang[1✉], Dou Quan[1], Ning Huyan[3], Nicu Sebe[2], Zhun Zhong[4 ✉]

[1]Xidian University, [2]University of Trento, [3]Tsinghua University, [4]Hefei University of Technology

[*]Equal contribution

[✉]Corresponding author

Problem Formulation



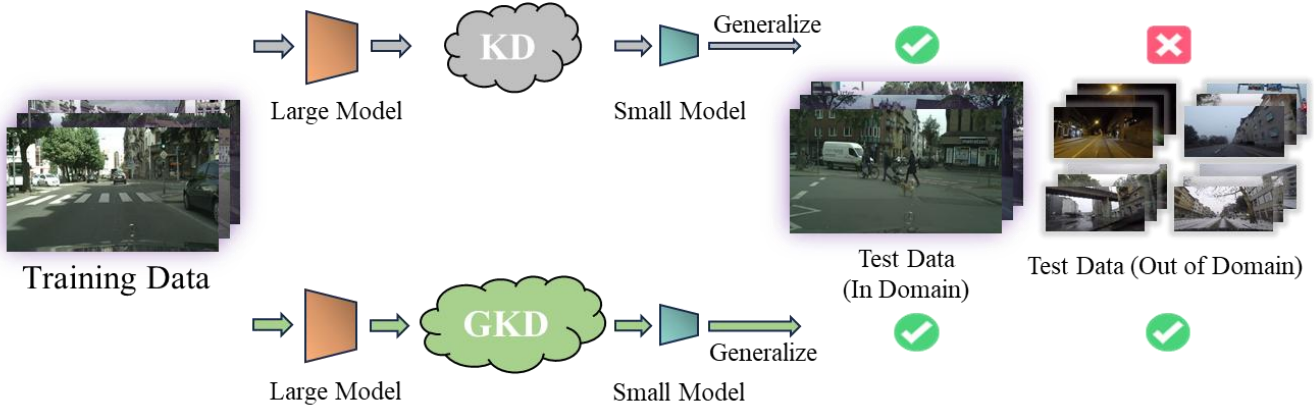
Optimization: $L(x;W) = \alpha * \mathrm{H}\big(y, \sigma(z_s; T=1)\big) + \beta * \mathrm{H}\big(\sigma(z_s; T=\tau), \sigma(z_s; T=\tau)\big)$

Task Loss

Distillation Loss

Motivation

- Knowledge Distillation (KD) compresses large models for semantic segmentation but overlooks domain generalization.
- Vision Foundation Models (VFMs) are robust on unseen domains, but conventional KD fails to transfer this generalization.



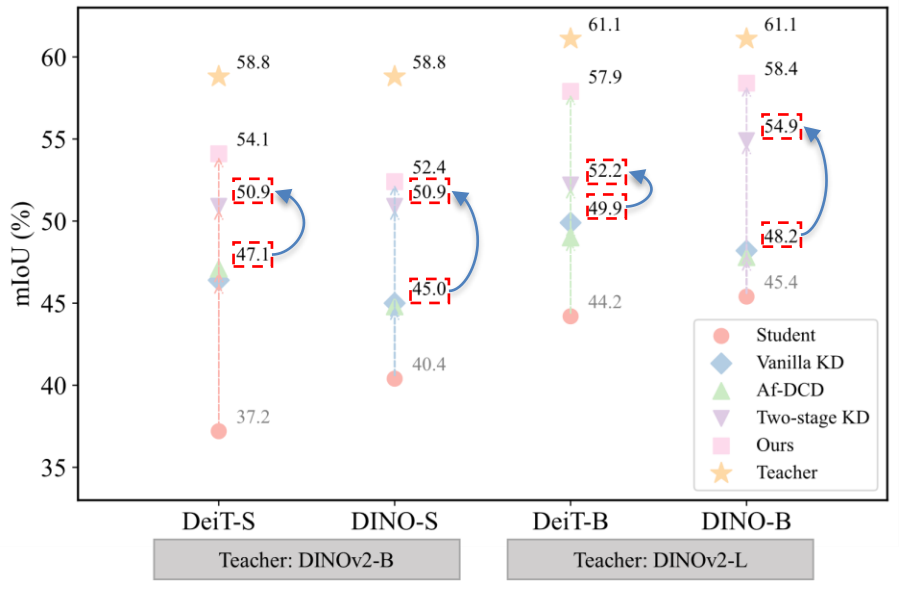
Motivation

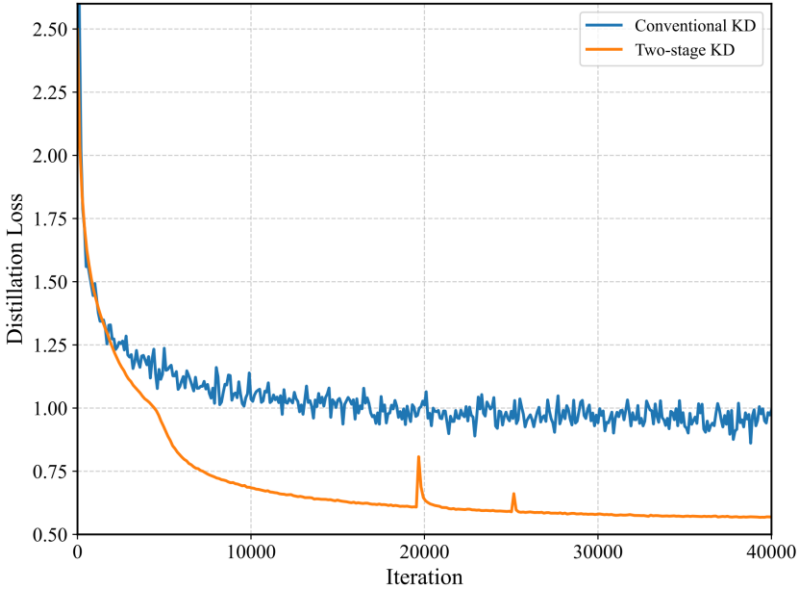
$$\min_{\theta_s,\theta_h} \mathbb{E}_{(x_S,y_S)\sim D_S}[\underbrace{\mathcal{L}(\mathcal{H}_{\theta_h}(\mathcal{F}_{\theta_s}(x_S)),y_S)}_{\text{Task Learning}} + \underbrace{\left\|\mathcal{F}_{\theta_t}(x_S)-\mathcal{F}_{\theta_s}(x_S)\right\|_2^2}_{\text{Representation Learning}}]$$

Key Insights ⇒ **decoupling**

Two-stage KD

- Performing feature distillation on source images.
- Freezing the encoder to train the decoder with standard task supervision.





Our approach

- **Domain-general Distillation:**

(1) Task-agnostic Distillation

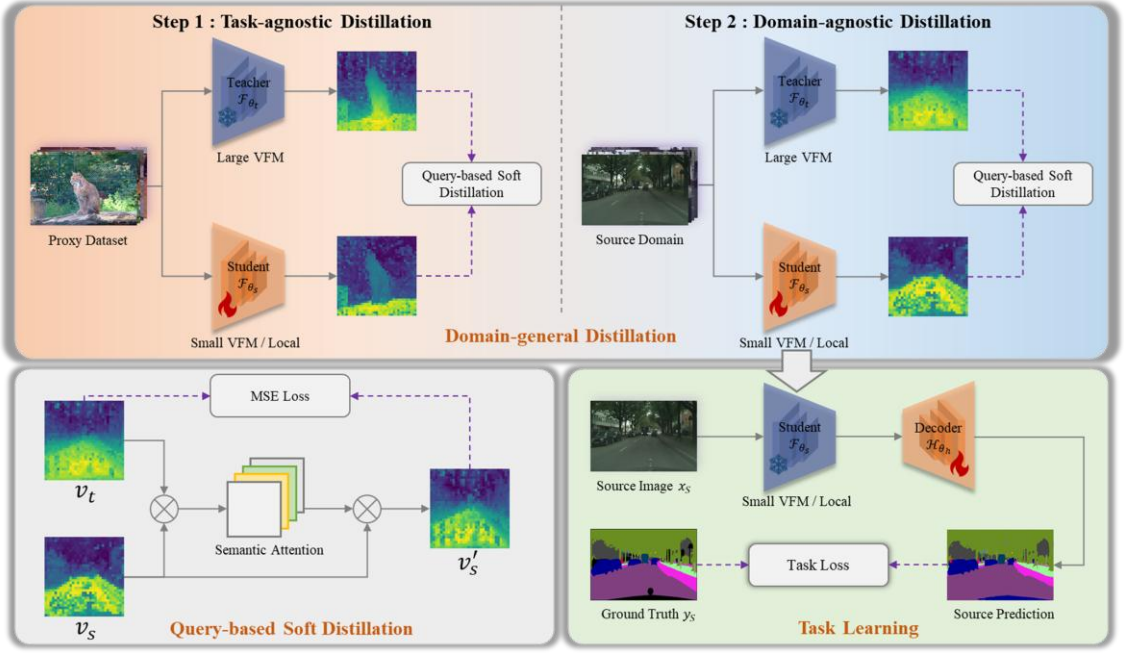
$$\min_{\theta_s} \mathbb{E}_{x_P \sim D_P}[\mathcal{L}_{QSD}(\mathcal{F}_{\theta_t}(x_P), \mathcal{F}_{\theta_s}(x_P))]$$

(2) Domain-agnostic Distillation

$$\min_{\theta_s} \mathbb{E}_{x_S \sim D_S}[\mathcal{L}_{QSD}(\mathcal{F}_{\theta_t}(x_S), \mathcal{F}_{\theta_s}(x_S))]$$

- **Task Learning:**

$$\min_{\theta_h} \mathbb{E}_{(x_S, y_S) \sim D_S}[\mathcal{L}(\mathcal{H}_{\theta_h}(\mathcal{F}_{\theta_s}(x_S)), y_S)]$$



Experiment Results

Method	Arch	Params	GTAV				Cityscapes					P-R		
			Citys	BDD	Map	Avg.	Night	Snow	Fog	Rain	Avg.	P-I	V-I	Avg.
Tea: DINOv2	ViT-L	324.8M	63.3	56.1	63.9	61.1	54.6	69.4	78.9	72.6	68.9	76.7	63.4	70.1
DINOv2	ViT-B	106.8M	59.6	54.3	62.6	58.8	49.9	67.6	77.5	69.9	66.2	72.3	55.9	64.1
Stu: DeiT	ViT-B	106.8M	43.1	41.8	47.7	44.2	28.1	47.2	64.1	48.5	47.0	67.5	34.1	50.8
+Vanilla KD [38]	ViT-B	106.8M	48.5	48.2	53.2	49.9	33.2	55.7	71.3	57.0	54.3	69.0	42.9	56.0
+CWD [42]	ViT-B	106.8M	49.3	46.7	51.8	49.3	33.1	53.8	70.5	54.1	52.9	70.0	41.5	55.8
+Af-DCD [9]	ViT-B	106.8M	48.4	45.2	53.4	49.0	31.7	54.3	67.3	55.3	52.1	69.5	42.5	56.0
+G2SD [21]	ViT-B	106.8M	50.8	49.1	53.4	51.1	33.1	55.7	69.5	56.8	53.8	72.4	46.7	59.5
+Vitkd [56]	ViT-B	106.8M	45.1	45.6	49.8	46.8	31.1	55.2	66.3	52.8	51.4	67.8	39.8	53.8
+Proteus [60]	ViT-B	106.8M	48.1	46.4	52.8	49.1	32.5	54.6	69.4	54.6	52.8	70.1	43.5	56.8
+GKD	ViT-B	106.8M	58.3	54.2	61.3	**57.9**	43.8	69.4	76.7	68.4	**64.6**	74.5	55.6	**65.1**
Tea: DINOv2	ViT-B	106.8M	59.6	54.3	62.6	58.8	49.9	67.6	77.5	69.9	66.2	72.3	55.9	64.1
DINOv2	ViT-S	41.9M	53.2	51.3	57.1	53.9	39.3	64.1	68.7	61.0	58.3	73.9	54.0	64.0
Stu: DeiT	ViT-S	41.9M	34.9	33.8	42.8	37.2	22.7	43.0	55.0	42.2	40.7	67.6	28.7	48.2
+Vanilla KD [38]	ViT-S	41.9M	45.0	44.2	49.9	46.4	31.4	51.3	63.6	50.1	49.1	70.3	37.6	54.0
+CWD [42]	ViT-S	41.9M	45.9	44.5	49.8	46.7	31.7	51.0	64.7	52.6	50.0	70.4	38.6	54.5
+Af-DCD [9]	ViT-S	41.9M	44.7	45.6	50.9	47.1	31.6	49.9	70.1	49.9	50.4	71.2	38.2	54.7
+G2SD [21]	ViT-B	41.9M	45.2	45.9	52.3	47.8	33.5	51.4	65.6	54.2	51.2	72.7	40.2	56.5
+Vitkd [56]	ViT-S	41.9M	42.5	42.5	48.2	44.4	28.0	51.3	65.1	45.9	47.6	62.9	34.9	48.9
+Proteus [60]	ViT-S	41.9M	47.4	44.6	50.2	47.4	32.8	51.3	62.1	49.3	48.9	70.8	38.5	54.7
+GKD	ViT-S	41.9M	54.9	49.8	57.8	**54.1**	39.3	60.4	72.7	58.4	**57.7**	73.8	48.7	**61.3**

Table 1. Performance comparison between proposed GKD and various KD methods in the F2L setting. P-R: Potsdam-RGB, P-I: PotsdamIRRG, V-I: Vaihingen-IRRG.

Method	Arch	Params	GTAV				Cityscapes					P-R		
			Citys	BDD	Map	Avg.	Night	Snow	Fog	Rain	Avg.	P-I	V-I	Avg.
Tea: DINOv2	ViT-L	324.8M	63.3	56.1	63.9	61.1	54.6	69.4	78.9	72.6	68.9	76.7	63.4	70.1
Stu: DINOv2	ViT-B	106.8M	59.6	54.3	62.6	58.8	49.9	67.6	77.5	69.9	66.2	72.3	55.9	64.1
+Vanilla KD [38]	ViT-B	106.8M	59.9	54.5	60.2	58.2	48.6	68.0	79.4	70.5	66.6	75.9	52.9	64.4
+Af-DCD [9]	ViT-B	106.8M	59.5	53.0	60.0	57.5	48.9	68.0	79.1	71.0	66.7	76.2	52.3	64.3
+Vitkd[56]	ViT-B	106.8M	58.0	53.0	59.3	56.7	46.6	67.6	77.1	69.6	65.2	75.5	51.6	63.6
+Proteus [60]	ViT-B	106.8M	60.1	54.6	61.4	58.7	48.3	67.6	79.7	71.1	66.7	75.6	53.5	64.6
+GKD	ViT-B	106.8M	62.6	55.0	61.8	**59.8**	48.3	71.3	80.3	72.0	**68.0**	75.4	56.4	**65.9**
Tea: DINOv2	ViT-B	106.8M	59.6	54.3	62.6	58.8	49.9	67.6	77.5	69.9	66.2	72.3	55.9	64.1
Stu: DINOv2	ViT-S	106.8M	53.2	51.3	57.1	53.9	39.3	64.1	68.7	61.0	58.3	73.9	54.0	64.0
+Vanilla KD [38]	ViT-S	41.9M	52.9	49.4	56.3	52.9	38.6	62.6	73.8	61.8	59.2	76.5	48.9	62.7
+Af-DCD [9]	ViT-S	41.9M	54.0	50.2	55.7	53.3	37.5	63.2	75.4	59.3	58.8	74.2	42.6	58.4
+Vitkd [56]	ViT-S	41.9M	49.9	49.1	55.7	51.6	37.8	62.9	73.5	60.3	58.6	71.7	43.7	57.7
+Proteus [60]	ViT-S	41.9M	53.5	49.7	56.9	53.4	37.6	62.5	74.9	62.6	59.4	76.0	42.5	59.3
+GKD	ViT-S	41.9M	57.1	51.3	58.4	**55.6**	39.2	62.6	75.5	62.2	**59.9**	74.0	53.5	**63.8**
Tea: EVA02	TrV-L	324.8M	58.4	52.5	59.0	56.7	39.1	64.9	73.3	62.6	60.0	74.8	48.8	61.8
Stu: EVA02	TrV-B	106.8M	56.2	53.0	59.4	56.2	46.1	65.1	76.7	62.6	62.6	74.7	51.6	63.2
+Vanilla KD [38]	TrV-B	106.8M	54.4	53.2	59.4	55.7	43.5	65.2	75.3	62.5	61.6	71.4	47.4	59.4
+Af-DCD [9]	TrV-B	106.8M	55.8	52.9	58.0	55.6	46.4	65.8	75.4	63.5	62.7	72.8	47.5	60.2
+Vitkd [56]	TrV-B	106.8M	48.6	50.2	55.2	51.3	32.5	59.2	68.4	56.6	54.2	71.2	44.2	57.7
+Proteus [60]	TrV-B	106.8M	53.7	52.8	59.4	55.3	45.4	64.7	74.2	61.1	61.4	73.5	48.6	61.1
+GKD	TrV-B	106.8M	59.0	54.5	61.0	**58.2**	46.9	67.8	77.1	65.8	**64.4**	76.4	57.9	**67.2**
Tea: EVA02	TrV-B	106.8M	45.9	44.1	49.8	46.6	20.9	54.6	63.3	48.7	46.9	66.4	34.0	50.2
Stu: EVA02	TrV-S	41.9M	48.5	47.0	52.8	49.4	37.6	56.4	70.8	54.4	54.8	69.4	42.8	56.1
+Vanilla KD [38]	TrV-S	41.9M	47.5	46.2	52.2	48.6	34.1	57.1	69.7	52.1	53.2	69.9	42.2	56.1
+Af-DCD [9]	TrV-S	41.9M	48.3	47.2	52.1	49.2	36.1	56.8	70.9	55.0	54.7	70.3	42.6	56.5
+Vitkd [56]	TrV-S	41.9M	43.3	42.1	47.7	44.4	28.9	52.2	66.1	49.7	49.2	66.0	34.8	50.4
+Proteus [60]	TrV-S	41.9M	47.1	45.1	51.1	47.8	34.0	56.3	70.8	50.6	52.9	68.4	41.8	55.1
+GKD	TrV-S	41.9M	51.1	45.9	53.7	**50.2**	36.0	59.0	71.2	55.8	**55.5**	71.7	45.7	**58.7**

Table 2. Performance comparison between proposed GKD and various KD methods in the F2F setting.

Experiment Results

Method	GTAV				Cityscapes			
	1/16	1/8	1/4	full	1/16	1/8	1/4	full
F2F								
Stu: DINOv2-B	58.3	58.4	58.7	58.8	62.1	63.9	64.1	66.2
+Af-DCD	56.2	56.6	56.9	57.5	60.9	62.2	63.9	66.7
+GKD	**58.4**	**58.6**	**59.1**	**59.8**	**62.3**	**64.7**	**64.9**	**67.2**
Stu: DINOv2-S	52.4	53.0	53.5	53.9	53.0	55.0	56.8	58.3
+Af-DCD	48.9	49.7	52.4	53.3	52.3	55.2	57.4	58.8
+GKD	**54.7**	**55.0**	**55.3**	**55.6**	**54.7**	**56.5**	**57.6**	**59.9**
F2L								
Stu: DeiT-B	42.1	42.1	43.2	44.2	39.1	42.6	43.4	47.0
+Af-DCD	47.4	48.4	49.0	49.0	45.2	48.9	50.4	52.1
+GKD	**56.5**	**56.8**	**56.8**	**57.9**	**59.1**	**60.0**	**61.2**	**64.6**
Stu: DeiT-S	35.7	36.7	37.5	37.2	32.7	38.0	38.2	40.7
+Af-DCD	46.0	46.1	46.2	47.1	43.6	46.5	49.0	50.4
+GKD	**51.4**	**51.5**	**53.6**	**54.1**	**54.6**	**54.8**	**57.0**	**57.7**

Table 3. Performance comparison under different labeled data fractions.

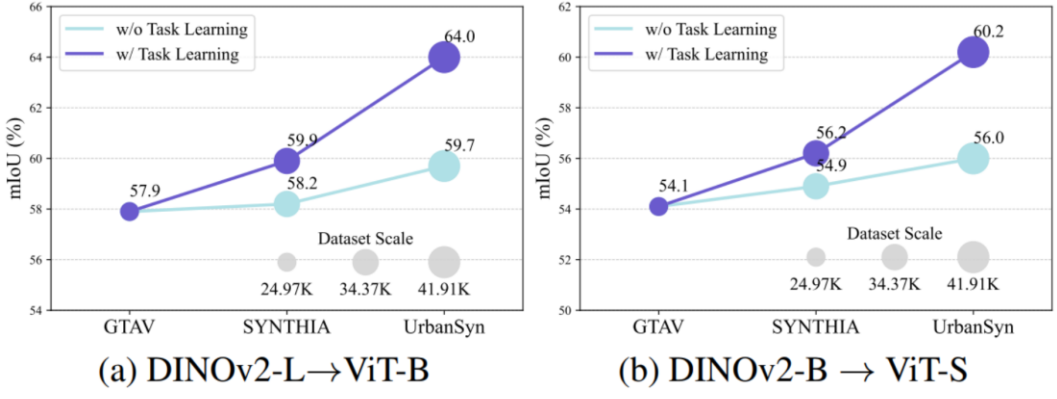


Figure 1. Performance comparison on more source domains under Citys + BDD + Map generalization setting.

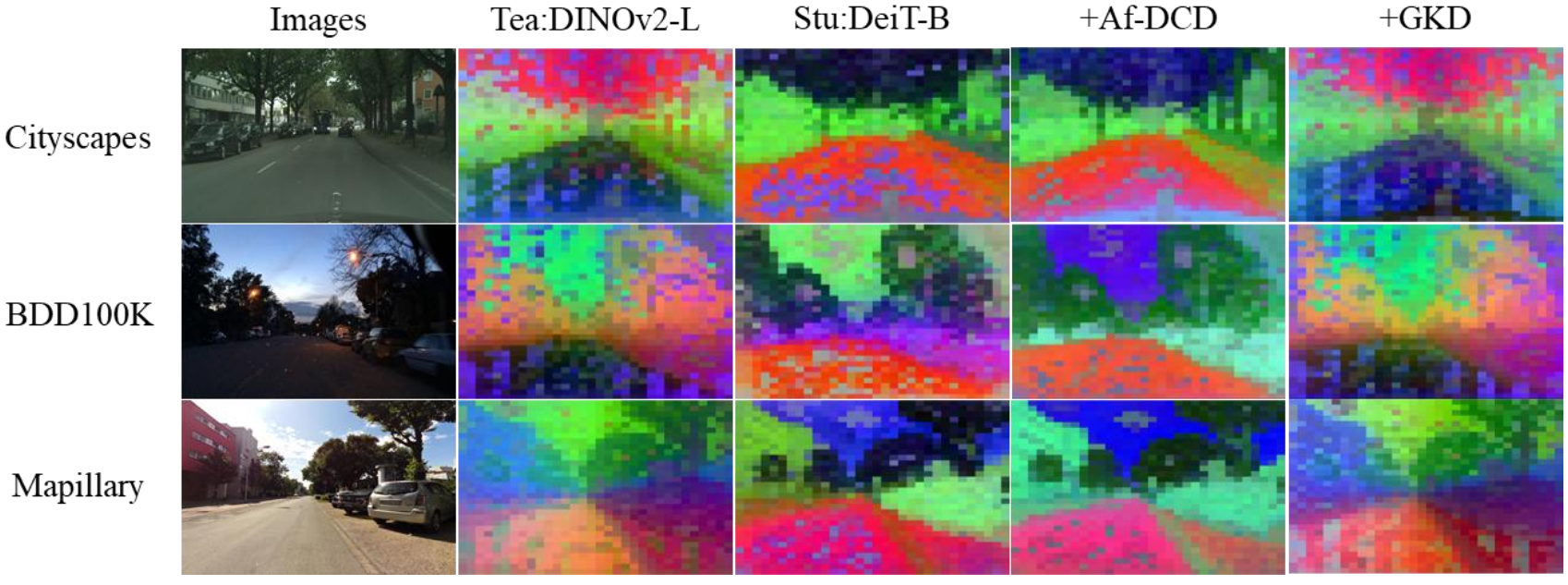
Methods	Citys	BDD	Map	Avg.
MSE†	45.0	44.2	49.9	46.4
QSD†	48.9	46.5	51.1	48.8
MSE	54.2	49.0	56.1	53.1
CWD	53.0	48.9	53.8	51.9
Vitkd	53.2	48.7	55.0	52.3
QSD	**54.9**	**49.8**	**57.8**	**54.1**

Table 4. Ablation study on distillation strategies.

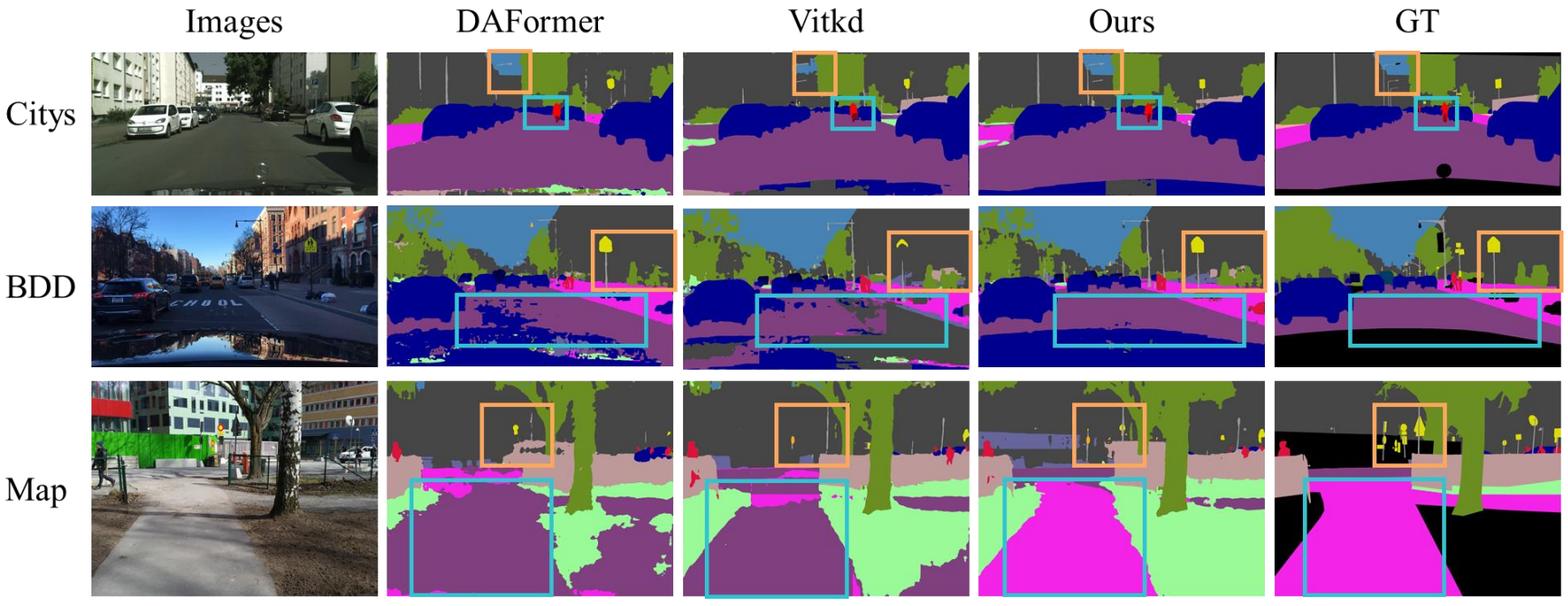
Task-agnostic Distillation	Domain-agnostic Distillation	QSD			Frozen Encoder	mIoU
		CLS Token	Feature	Mask Patch		
✗	✗	✗	✗	✗	✗	46.4
✗	✓	✗	✗	✗	✗	50.9
✓	✓	✗	✗	✗	✗	53.1
✓	✓	✓	✓	✗	✗	53.4
✓	✓	✓	✓	✓	✗	54.0
✓	✓	✓	✓	✓	✓	**54.1**

Table 5. Ablation study for each component.

Experiment Results



Experiment Results



Thank you

Please feel free to contact me if you have any questions: youngerlv@stu.xidian.edu.cn