Generalizable Knowledge Distillation from Vision Foundation Models for Semantic Segmentation

Supplementary Material

Table 6. Performance comparison between proposed GKD and previous DGSS methods in the F2L setting. All models trained on GTAV and evaluated on Citys, BDD, Map and ACDC.

Method	Arch	Decoder	Params	Citys	BDD	Map	Avg.	ACDC
AdvStyle [65]	ResNet-101 [15]	DeepLabv3 [4]	60.2M	44.5	39.3	43.5	42.4	-
SHADE [64]	ResNet-101 [15]	DeepLabv3 [4]	60.2M	46.7	43.7	45.5	45.3	29.1
TLDR [24]	ResNet-101 [15]	DeepLabv3 [4]	60.2M	47.6	44.9	48.8	47.1	-
GKD	ViT-S	Mask2Former	41.9M	54.9	49.8	57.8	**54.1**	**46.5**
DAFormer [17]	MiT-B5	SegFormer [52]	82.0M	52.7	47.9	54.7	51.7	38.3
HRDA [18]	MiT-B5	SegFormer [52]	82.0M	57.4	49.1	61.2	55.9	44.0
GKD	ViT-B	Mask2Former	106.8M	58.3	54.2	61.3	**57.9**	**51.2**

7. Comparison with DGSS Methods

As shown in Tab. 6, GKD achieves substantial improvements over prior domain generalization for semantic segmentation (DGSS) methods such as AdvStyle [65], SHADE [64], and TLDR [24]. When using ViT-S as the student, GKD surpasses these approaches by a large margin across all target domains, even though it employs a simpler training protocol without domain-specific augmentations or style perturbations. Moreover, when scaling to a ViT-B backbone, GKD further boosts performance to 57.9% average mIoU and 51.2% on the challenging ACDC benchmark, outperforming strong domain adaptation baselines like DAFormer [17] and HRDA [18]. These results highlight the potential of GKD to bridge knowledge distillation and domain generalization. We hope it offers a new perspective on leveraging VFMs for efficient and generalizable semantic segmentation.

Table 7. Results on different decoders in the F2L setting.

Arch	Decoder	Params	GTAV				Cityscapes				
			Citys	BDD	Map	Avg.	Night	Snow	Fog	Rain	Avg.
ViT-B	SemFPN [25]	99.7M	55.5	54.0	59.5	56.3	43.2	67.5	75.3	67.8	63.5
ViT-B	Mask2Former	106.8M	58.3	54.2	61.3	**57.9**	43.8	69.4	76.7	68.4	**64.6**
ViT-S	SemFPN [25]	33.2M	54.7	50.6	56.6	54.0	38.6	61.5	71.7	61.9	**58.4**
ViT-S	Mask2Former	41.9M	54.9	49.8	57.8	**54.1**	39.3	60.4	72.7	58.4	57.7

8. Ablation on Other Decoder

To examine whether the observed gains originate from specific decoder choices, we evaluate GKD with SemFPN [25] and Mask2Former [5], as shown in Tab. 7. GKD consistently enhances generalization across both ViT-B and ViT-S backbones, independent of the decoder design, with Mask2Former yielding slightly better results. Therefore, we use Mask2Former as the default decoder in the main experiments.

Table 8. Performance comparison between proposed GKD and various KD methods in the F2L setting. Tea:Teacher. Stu: student. We use the self-supervised method DINO to initialize the student.

Method	Arch	GTAV				Cityscapes				
		Citys	BDD	Map	Avg.	Night	Snow	Fog	Rain	Avg.
Tea: DINOv2	ViT-L	63.3	56.1	63.9	61.1	54.6	69.4	78.9	72.6	68.9
DINOv2	ViT-B	59.6	54.3	62.6	58.8	49.9	67.6	77.5	69.9	66.2
Stu: DINO	ViT-B	44.5	42.7	49.2	45.4	29.0	51.1	65.6	46.7	48.1
+Vanilla KD [38]	ViT-B	47.9	45.0	51.7	48.2	32.5	53.2	63.3	52.4	50.3
+CWD [42]	ViT-B	45.0	44.8	49.5	46.4	29.3	52.0	67.2	55.9	51.1
+Af-DCD [9]	ViT-B	47.1	46.4	50.0	47.8	31.1	53.9	66.1	52.6	50.9
+GKD	ViT-B	60.5	53.6	61.1	**58.4**	43.2	63.2	75.6	61.0	**60.7**
Tea: DINOv2	ViT-B	59.6	54.3	62.6	58.8	49.9	67.6	77.5	69.9	66.2
DINOv2	ViT-S	53.2	51.3	57.1	53.9	39.3	64.1	68.7	61.0	58.3
Stu: DINO	ViT-S	39.7	36.9	44.7	40.4	27.8	48.0	66.4	44.5	46.6
+Vanilla KD [38]	ViT-S	45.4	41.5	47.9	45.0	30.8	52.7	67.9	50.6	50.5
+CWD [42]	ViT-S	44.0	41.4	47.1	44.1	30.7	51.4	62.1	51.4	48.9
+Af-DCD [9]	ViT-S	44.5	42.6	47.3	44.8	28.7	52.0	68.2	50.4	49.8
+GKD	ViT-S	53.2	48.6	55.5	**52.4**	35.6	59.0	72.4	56.8	**55.9**

9. Initialize with Self-supervised Methods

We further investigate whether GKD can benefit from self-supervised initializations and maintain its robustness across different pre-trained strategies (Tab. 8). When initializing the student with DINO [3], conventional KD methods (Vanilla KD, CWD, Af-DCD) provide only marginal improvements over the baseline. In contrast, GKD achieves substantial gains, narrowing the teacher–student gap and reaching 58.4% mIoU for ViT-B and 52.4% mIoU for ViT-S on the GTAV $\rightarrow$ Citys + BDD + Map generalization setting, while also delivering 60.7%/55.9% on the Citys $\rightarrow$ Night + Snow + Fog + Rain generalization setting. These results indicate that GKD effectively transfers generalizable knowledge independent of the pretraining recipe, underscoring its robustness and versatility.

10. Generalization on Medical Scene

We additionally validate the effectiveness of GKD on the medical scene. The EndoVis18 [1] comprises 7 real robotic surgery video sequences with over 8,000 frames and pixel-level annotations with 12 instrument categories. We use training sets 1–4 as the source domain and training sets 5–7 as unseen target domain, with clear domain shifts in lighting, viewpoint, surgical context[63]. As shown in Tab. 9, GKD still achieves significant improvements over prior KD methods, only a slight difference compared to the teacher.

11. Generalization on Object Decection

GKD decouples representation and task learning, thus it extends beyond segmentation by replacing the task decoder

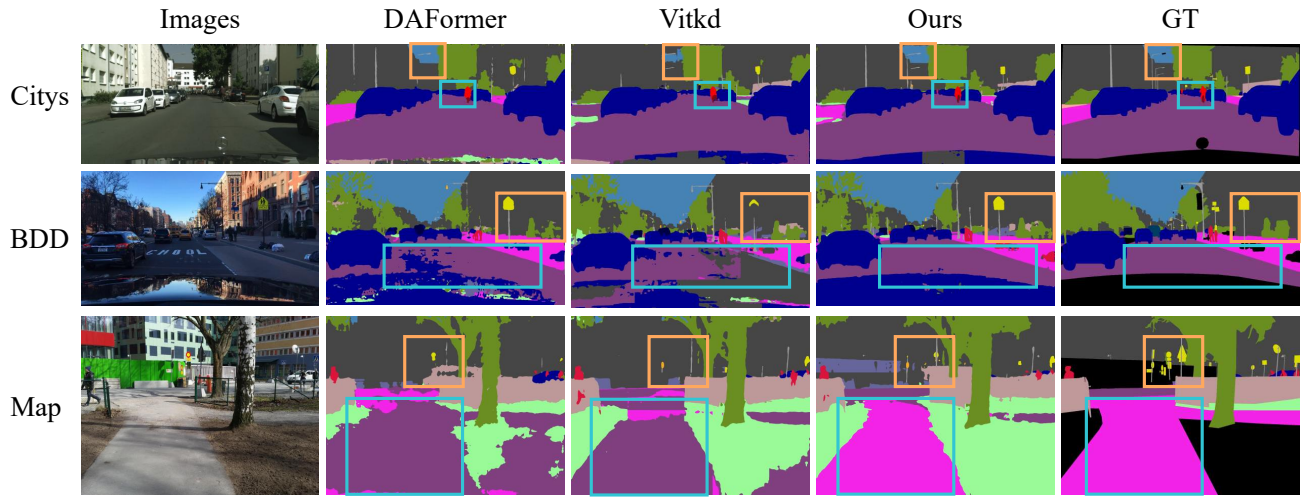
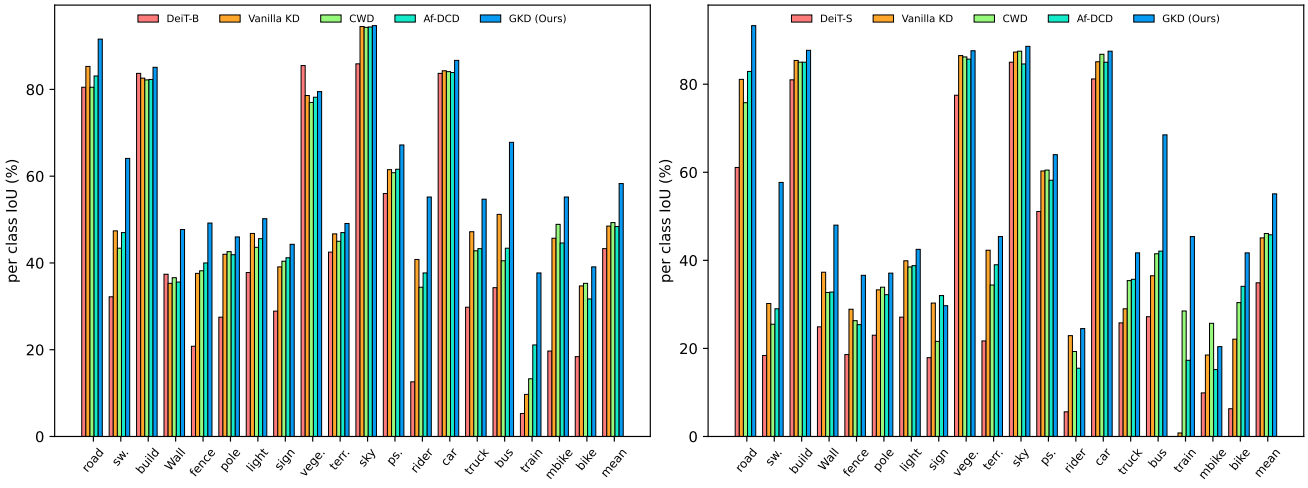


Figure 8. Segmentation results of various KD methods (with ViT-B as backbone) and existing DGSS methods (DAFormer [17] with MiT-B5 [52] as backbone) under GTAV → Citys + BDD + Map generalization setting.



(a) DINOv2-L→ViT-B

(b) DINOv2-B→ViT-S

Figure 9. Performance comparison of each category between GKD and conventional KD on Cityscapes validation set. We adopt DeiT to initialize the student, (a) and (b) represent different distillation architectures.

Table 9. Generalization to medical scene EndoVis18.

Method	Arch	Params	*Source:* EndoVis18 Seq 1-4
			Target: EndoVis18 Seq 5-7 (mIoU)
Tea: DINOv2	ViT-B	106.8M	53.7
Stu: DeiT	ViT-S	41.9M	43.3
+CWD	ViT-S	41.9M	46.5
+Vitkd	ViT-S	41.9M	45.1
+GKD	ViT-S	41.9M	**51.6**

Table 10. Generalization to object detection with Faster-RCNN.

Method	Arch	*Source:* Pascal VOC
		*Target:*Pascal VOC-corruptions (mAP)
Tea: DINOv2	ViT-B	79.9
Stu: DeiT	ViT-S	49.0
+CWD	ViT-S	56.3
+Vitkd	ViT-S	56.6
+GKD	ViT-S	**72.9**

and loss. We additionally validate the effectiveness of GKD on object decection. We train on Pascal VOC[8]. Notably, we introduce image corruption to introduce domain distribution shift. As shown in Tab. 10, GKD still achieves significant improvements over prior KD methods.

12. Qualitative Results

Some segmentation results are compared are provided in Fig. 8 Compared to DAFormer and Vitkd, GKD predicts clearer object boundaries and more consistent predictions across diverse domains. Such as clearer class boundaries (bus, road and sidewalk), and better preservation of small

targets (pole, traffic sign and rider). This visual consistency is quantitatively confirmed by the per-class IoU analysis in Fig. 9. Our query-based soft distillation mechanism enables the student to inherit domain-agnostic semantic relations from the teacher, yielding more coherent predictions and robust generalization under domain shifts.