Stronger, Fewer, & Superior: Harnessing Vision Foundation Models for Domain Generalized Semantic Segmentation

Zhixiang Wei, Lin Chen*, Yi Jin*, Xiaoxiao Ma, Tianle Liu, Pengyang ling, et al.*

Motivation

Leveraging **Stronger** pre-trained models and **Fewer** trainable parameters for **Superior** generalizability

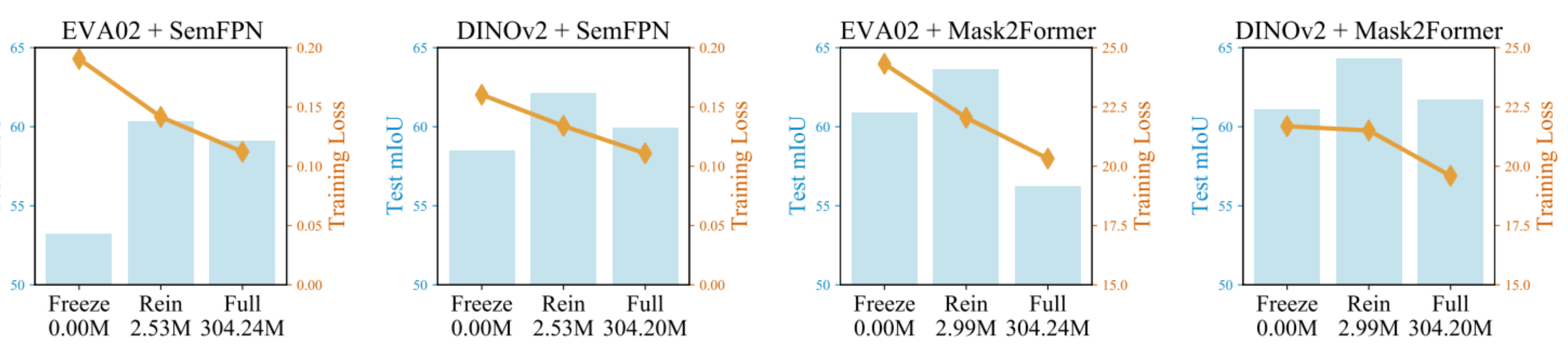
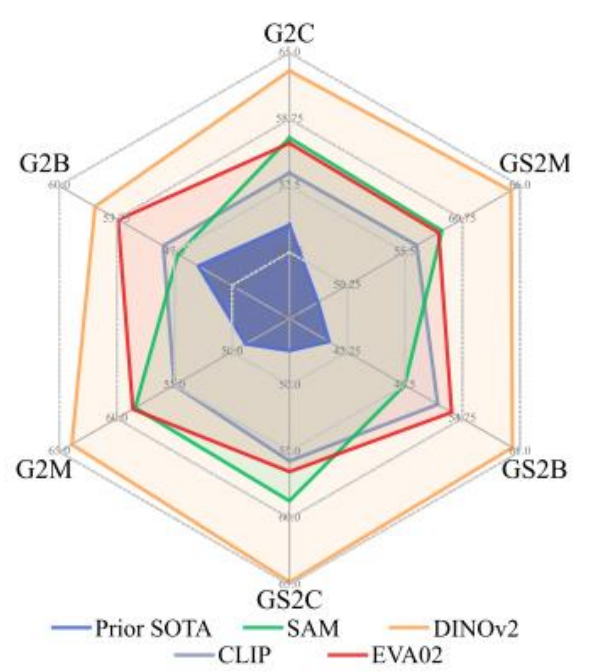
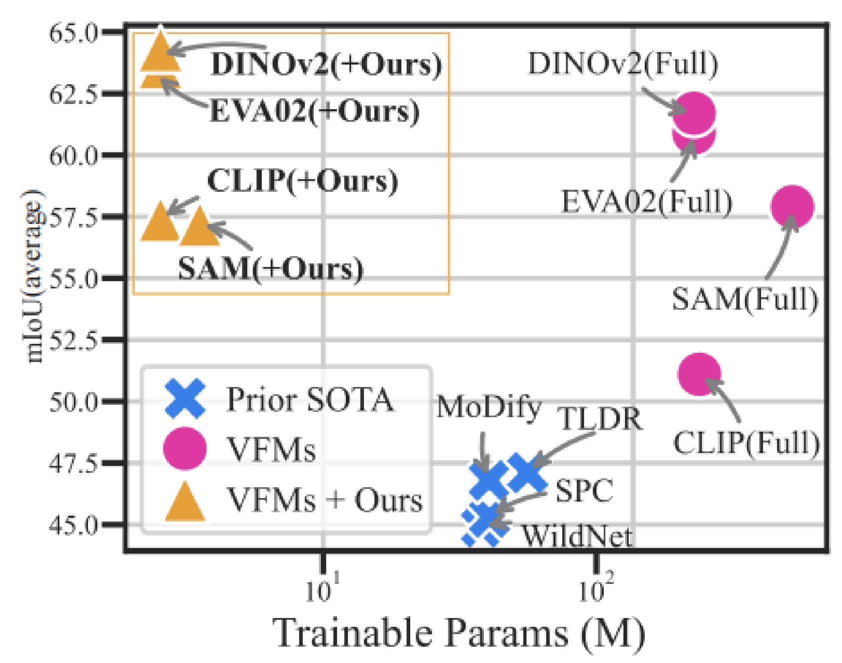


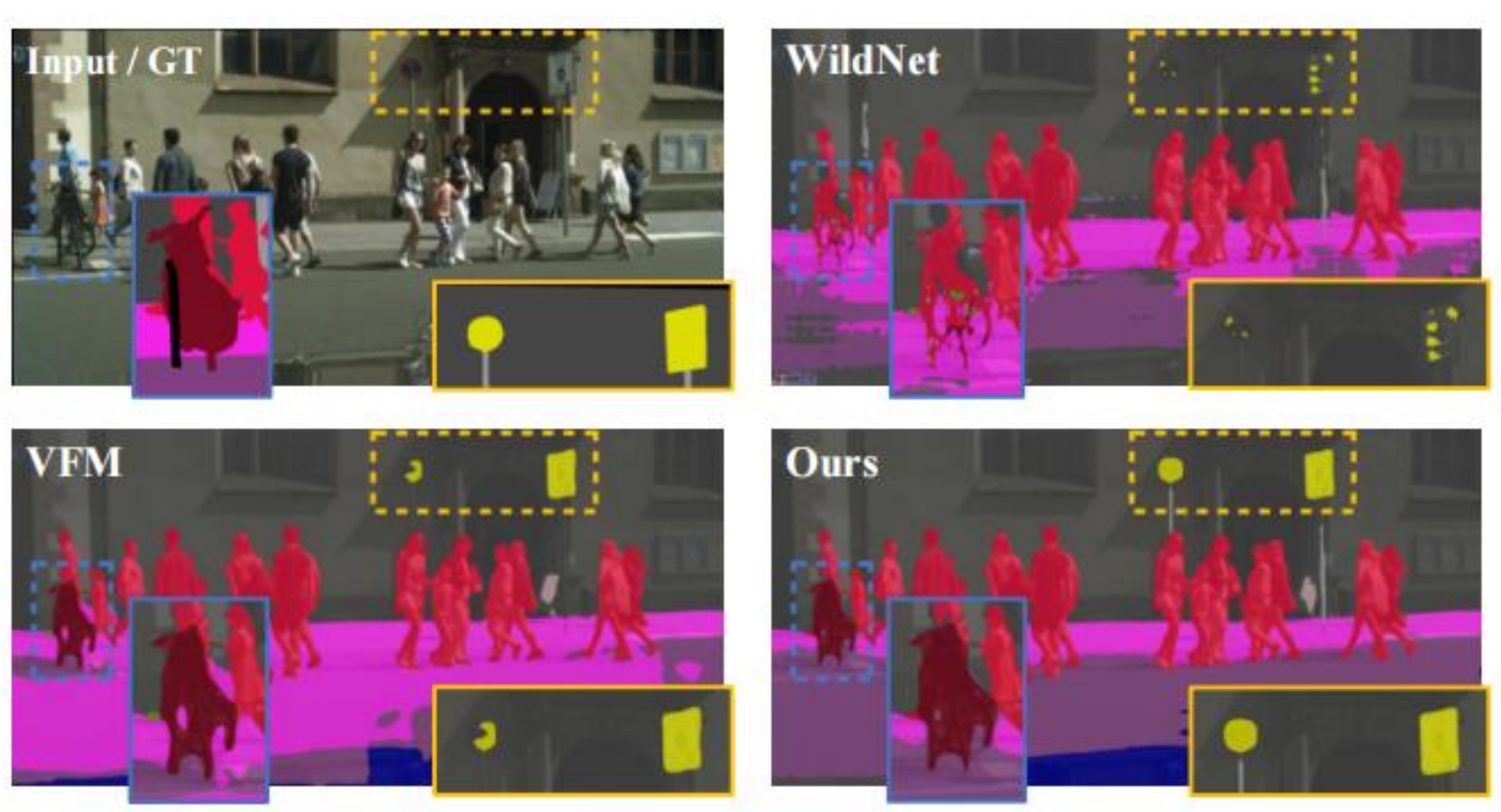
Figure 1: The curves of training loss and test metrics display consistent trends across different VFMs and decode heads.



(a) **Stronger** pre-trained models



(b) **Fewer** trainable parameters



(c) **Superior** generalization ability

Figure 2: Vision Foundation Models (VFMs) are stronger pre-trained models that serve as robust backbones, effortlessly outperforming previous state-of-the-art Domain Generalized Semantic Segmentation (DGSS),

Method

Assess VFMs for DGSS

Previous DGSS methods						
Methods	GTR[49]	AdvStyle[68]	WildNet[37]	SPC[24]	PASTA[4]	TLDR[34]
Publications	TIP21	NIPS22	CVPR22	CVPR23	ICCV23	ICCV23
mIoU (Citys)	43.7	43.4	45.8	46.7	45.3	47.6
mIoU (BDD)	39.6	40.3	41.7	43.7	42.3	44.9
mIoU (Map)	39.1	42.0	47.1	45.5	48.6	48.8
mIoU (Average)	40.8	41.9	44.9	45.3	45.4	47.1

Frozen backbone of VFMs					
Methods	CLIP-ViT-L[51]	MAE-L[21]	SAM-H[35]	EVA02-L[16]	DINOv2-L[46]
Publications	ICML21	CVPR22	ICCV23	arXiv23	arXiv23
mIoU (Citys)	53.7	43.3	57.0	56.5	**63.3**
mIoU (BDD)	48.7	37.8	47.1	53.6	**56.1**
mIoU (Map)	55.0	48.0	58.4	58.6	**63.9**
mIoU (Average)	52.4	43.0	54.2	56.2	**61.1**

Table 1: We begin by comparing the performance of various VFMs against existing DGSS methods, demonstrate the powerful potential of VFMs in DGSS, thereby establishing VFMs as a meaningful benchmark in the field.

Harness VFMs for DGSS

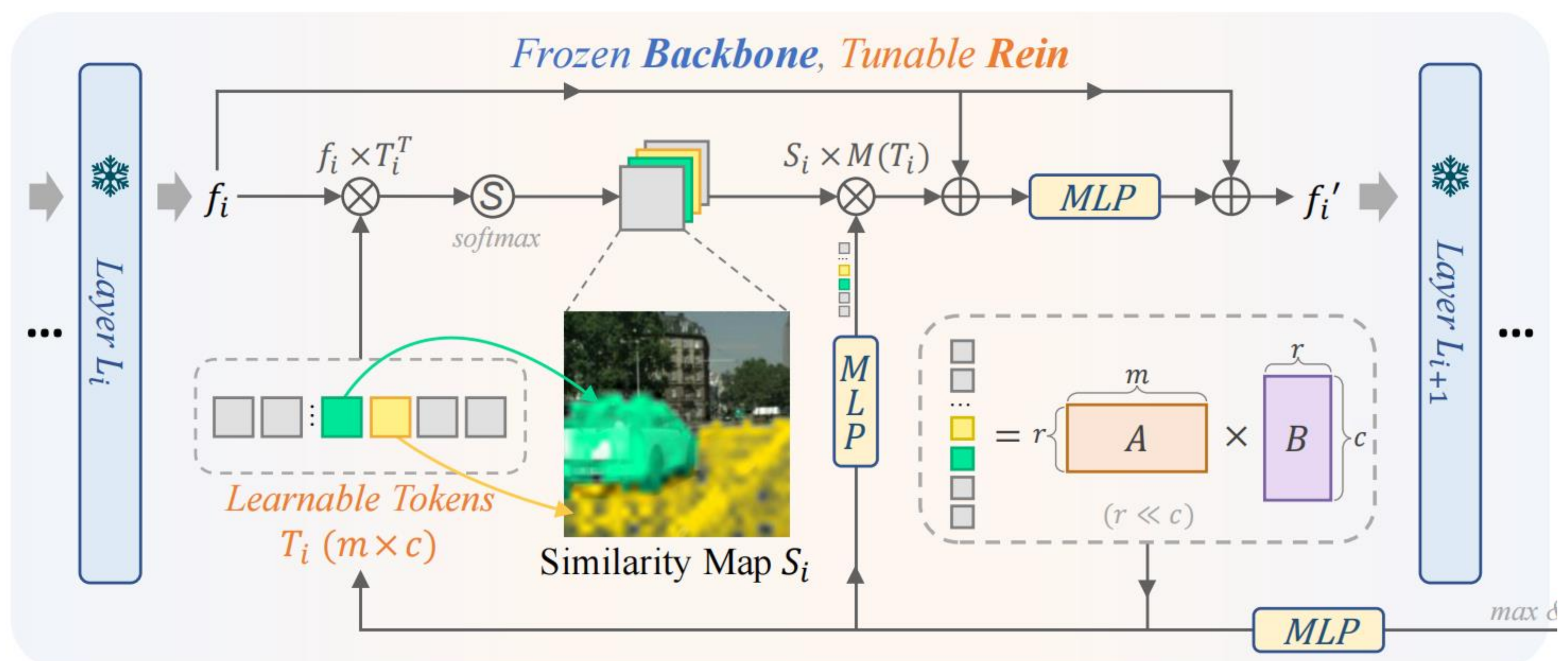


Figure 3. An overview of proposed Rein.

GitHub **WeChat**

Experiments

Backbone	Fine-tune Method	Trainable Params*	mIoU Citys	BDD	Map	Avg.
CLIP [51] (ViT-Large)	Full	304.15M	51.3	47.6	54.3	51.1
	Freeze	0.00M	53.7	48.7	55.0	52.4
	Rein	2.99M	**57.1**	**54.7**	**60.5**	**57.4**
MAE [21] (Large)	Full	330.94M	53.7	**50.8**	58.1	54.2
	Freeze	0.00M	43.3	37.8	48.0	43.0
	Rein	2.99M	**55.0**	49.3	**58.6**	**54.3**
SAM [35] (Huge)	Full	632.18M	57.6	51.7	61.5	56.9
	Freeze	0.00M	57.0	47.1	58.4	54.2
	Rein	4.51M	**59.6**	**52.0**	**62.1**	**57.9**
EVA02 [16, 17] (Large)	Full	304.24M	62.1	56.2	64.6	60.9
	Freeze	0.00M	56.5	53.6	58.6	56.2
	Rein	2.99M	**65.3**	**60.5**	**64.9**	**63.6**
DINOV2 [46] (Large)	Full	304.20M	63.7	57.4	64.2	61.7
	Freeze	0.00M	63.3	56.1	63.9	61.1
	Rein	2.99M	**66.4**	**60.4**	**66.1**	**64.3**

Table 2: Performance Comparison with the proposed Rein across Multiple VFMs as Backbones.

Target	ACDC[55] (test) Night	Snow	Fog	Rain	All
HGFormer	52.7	68.6	69.9	72.0	67.2
Ours	**70.6**	**79.5**	**76.4**	**78.2**	**77.6**

Table 3: Results on Cityscapes → ACDC (test).

Source Domain	Cityscapes mIoU
GTAV	66.4
+Synthia	68.1
+UrbanSyn	**78.4**
+1/16 of Cityscapes Training set	**82.5**

Table 4: Synthetic data + 1/16 of Citys. → Citys. val set.

Backbone	Fine-tune Method	Trainable Params*	mIoU Citys	BDD	Map	Avg.
EVA02 (Large) [16, 17]	Full	304.24M	62.1	56.2	64.6	60.9
	+AdvStyle [68]	304.24M	63.1	56.4	64.0	61.2
	+PASTA [4]	304.24M	61.8	57.1	63.6	60.8
	+GTR-LTR [49]	304.24M	59.8	57.4	63.2	60.1
	Freeze	0.00M	56.5	53.6	58.6	56.2
	+AdvStyle [68]	0.00M	51.4	51.6	56.5	53.2
	+PASTA [4]	0.00M	57.8	52.3	58.5	56.2
	+GTR-LTR [49]	0.00M	52.5	52.8	57.1	54.1
	+LoRA [23]	1.18M	55.5	52.7	58.3	55.5
	+AdaptFormer [5]	3.17M	63.7	59.9	64.2	62.6
	+VPT [25]	3.69M	62.2	57.7	62.5	60.8
	+Rein (ours)	2.99M	**65.3**	**60.5**	**64.9**	**63.6**
DINOv2 (Large) [46]	Full	304.20M	63.7	57.4	64.2	61.7
	+AdvStyle [68]	304.20M	60.8	58.0	62.5	60.4
	+PASTA [4]	304.20M	62.5	57.2	64.7	61.5
	+GTR-LTR [4]	304.20M	62.7	57.4	64.5	61.6
	Freeze	0.00M	63.3	56.1	63.9	61.1
	+AdvStyle [68]	0.00M	61.5	55.1	63.9	60.1
	+PASTA [4]	0.00M	62.1	57.2	64.5	61.3
	+GTR-LTR [4]	0.00M	60.2	57.7	62.2	60.0
	+LoRA [23]	0.79M	65.2	58.3	64.6	62.7
	+AdaptFormer [5]	3.17M	64.9	59.0	64.2	62.7
	+VPT [25]	3.69M	65.2	59.4	65.5	63.3
	+Rein (ours)	2.99M	**66.4**	**60.4**	**66.1**	**64.3**

Table 5: Performance Comparison of the proposed Rein against other DGSS and PEFT methods.

Avg. mIoU	ResNet (50)	ResNet (101)	ConvNeXt (Large)
Full	38.9	46.1	52.2
Ours	**46.6**	**46.3**	**55.5**

Methods	Backbone	Trainable Parameters*	mIoU BDD	Map	Avg.
IBN [47]	ResNet50 [20]	23.58M	48.6	57.0	52.8
DRPC [64]	ResNet50 [20]	23.58M	49.9	56.3	53.1
GTR [49]	ResNet50 [20]	23.58M	50.8	57.2	54.0
SAN-SAW [50]	ResNet50 [20]	23.58M	53.0	59.8	56.4
WildNet [37]	ResNet101 [20]	42.62M	50.9	58.8	54.9
HGFormer [12]	Swin-L [41]	196.03M	61.5	72.1	66.8
Freeze	EVA02-L [16]	0.00M	57.8	63.8	60.8
Rein (Ours)	EVA02-L [16]	2.99M	64.1	69.5	66.8
Freeze	DINOv2-L [46]	0.00M	63.4	69.7	66.7
Rein (Ours)	DINOv2-L [46]	2.99M	**65.0**	**72.3**	**68.7**

Table 7: Results for Cityscapes to BDD100K+Mapillary.