FisherTune: Fisher-Guided Robust Tuning of Vision Foundation Models for Domain Generalized Segmentation (Supplementary Material)

Dong Zhao[1], Jinlong Li[2], Shuang Wang[1] ✉, Mengyao Wu, Qi Zang[1] ✉, Nicu Sebe[2], Zhun Zhong[3]
[1] School of Artificial Intelligence, Xidian University, Shaanxi, China
[2] Department of Information Engineering and Computer Science, University of Trento, Italy
[3] School of Computer Science and Information Engineering, Hefei University of Technology, China

A. Detailed Derivation

To simplify the expectation of the loss function in Eq.11, we perform a second-order Taylor expansion of the loss function $\mathcal{L}(\boldsymbol{\theta})$ around the current weight estimate $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

$$\mathcal{L}(\boldsymbol{\theta}) \approx \mathcal{L}(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \nabla_{\boldsymbol{\theta}} \mathcal{L}(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}), \tag{1}$$

where $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\hat{\boldsymbol{\theta}})$ is the First-order derivative of the loss with respect to the weights. $\nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\hat{\boldsymbol{\theta}})$ is the Second-order derivative of the loss (Hessian matrix). Assuming that $\hat{\boldsymbol{\theta}}$ is a local optimum where $\nabla_w \mathcal{L}(\hat{w}) = 0$, the first-order term vanishes. Thus, the Taylor expansion simplifies to:

$$\mathcal{L}(\boldsymbol{\theta}) \approx \mathcal{L}(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}). \tag{2}$$

Taking the expectation over the weight distribution $q(\boldsymbol{\theta})$:

$$\mathbb{E}_{w \sim q(\boldsymbol{\theta})} \left[\mathcal{L}(\boldsymbol{\theta}) \right] \approx \mathcal{L}(\hat{\boldsymbol{\theta}}) + \frac{1}{2} \operatorname{Tr} \left(\nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\hat{\boldsymbol{\theta}}) \Lambda^{-1} \right), \tag{3}$$

using the formula $\mathbb{E}_{\boldsymbol{\theta} \sim q(\boldsymbol{\theta})} \left[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top A (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right] = \operatorname{Tr}(A\Lambda^{-1})$. According to the definition of FIM and its connection with the Hessian matrix, the FIM can be approximated by the Hessian matrix near $\hat{\boldsymbol{\theta}}$: $\nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\hat{\boldsymbol{\theta}}) \approx \mathbf{F}_{\boldsymbol{\theta}}$. Thus,

$$\mathbb{E}_{\boldsymbol{\theta} \sim q(\boldsymbol{\theta})} \left[\mathcal{L}(\boldsymbol{\theta}) \right] \approx \mathcal{L}(\hat{\boldsymbol{\theta}}) + \frac{1}{2} \operatorname{Tr} \left(\mathbf{F}_{\boldsymbol{\theta}} \Lambda^{-1} \right). \tag{4}$$

The KL divergence between two Gaussian distributions is given by:

$$KL(q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta})) = \frac{1}{2} \Big(\operatorname{Tr}(\tau_p^{-1} \tau_q) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{pt}})^\top \tau_p^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{pt}}) - k + \ln \frac{\det \tau_p}{\det \tau_q} \Big). \tag{5}$$

where $\tau_q = \Lambda^{-1}$, $\tau_p = \tau^2 I$, k is the dimensionality of the parameters. After simplification:

$$KL(q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta})) = \frac{1}{2} \Big(\tau^{-2} \operatorname{Tr}(\Lambda^{-1}) + \tau^{-2} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{pt}}\|^2 - k + k \ln \tau^2 + \ln \det \Lambda \Big). \tag{6}$$

Substituting the results back into the total loss function:

$$L(\hat{\boldsymbol{\theta}}, \Lambda^{-1}) = \mathcal{L}(\hat{\boldsymbol{\theta}}) + \frac{1}{2} \operatorname{Tr} \left(\mathbf{F}_{\boldsymbol{\theta}} \Lambda^{-1} \right) + \gamma \left(\frac{1}{2} \tau^{-2} \operatorname{Tr}(\Lambda^{-1}) + \frac{1}{2} \tau^{-2} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{pt}}\|^2 + \frac{1}{2} \ln \det \Lambda \right) + \text{const}. \tag{7}$$

Taking the derivative of the loss function with respect to Λ:

$$\frac{\partial L}{\partial \Lambda} = -\frac{1}{2} \Lambda^{-1} \mathbf{F}_{\boldsymbol{\theta}} \Lambda^{-1} - \frac{\gamma}{2} \tau^{-2} \Lambda^{-1} \Lambda^{-1} + \frac{\gamma}{2} \Lambda^{-1}. \tag{8}$$

using the matrix derivative formulas, $\frac{\partial}{\partial X} \operatorname{Tr}(AX^{-1}) = -X^{-1} A X^{-1}$, and $\frac{\partial}{\partial X} \ln \det X = X^{-1}$. Setting $\frac{\partial L}{\partial \Lambda} = 0$:

$$-\left(\frac{1}{2} \Lambda^{-1} \mathbf{F}_{\boldsymbol{\theta}} \Lambda^{-1} + \frac{\gamma}{2} \tau^{-2} \Lambda^{-1} \Lambda^{-1} \right) + \frac{\gamma}{2} \Lambda^{-1} = 0. \tag{9}$$

Multiplying both sides by 2Λ:

$$-\left(\Lambda^{-1} \mathbf{F}_{\boldsymbol{\theta}} + \gamma \tau^{-2} \Lambda^{-1} \right) + \gamma I = 0, \tag{10}$$

which simplifies to:

$$\gamma I = \Lambda^{-1} \left(\mathbf{F}_{\boldsymbol{\theta}} + \gamma \tau^{-2} I \right). \tag{11}$$

Multiplying both sides by Λ:

$$\gamma \Lambda = \mathbf{F}_{\boldsymbol{\theta}} + \gamma \tau^{-2} I. \tag{12}$$

We can thus estimate the Fisher Information Matrix $\mathbf{F}_{\boldsymbol{\theta}}$:

$$\mathbf{F}_{\boldsymbol{\theta}} = \gamma \Lambda - \gamma \tau^{-2} I, \tag{13}$$

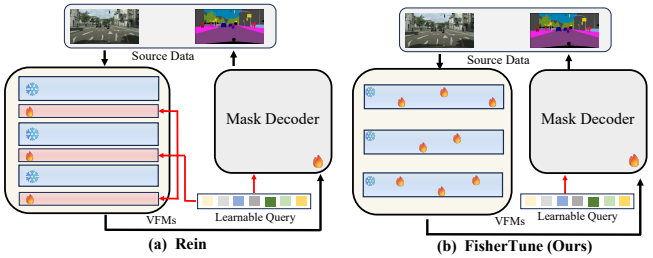

Figure 1. Structural differences between Rein [9] and our method.

As DR-FIM is calculated as,

$$\mathbf{DRF}_\theta = \mathbf{F}_\theta(x,y) + e^{-(\epsilon_\mu+\epsilon_\sigma)} \frac{|\mathbf{F}_{\boldsymbol{\theta}}(x,y) - \mathbf{F}_{\boldsymbol{\theta}}(x',y)|}{\min(\mathbf{F}_{\theta_i}(x), \mathbf{F}_{\theta_i}(x')) + \epsilon}. \tag{14}$$

After combining Eq. 13 with Eq. 14,

$$\mathbf{DRF}_\theta = \gamma \left(\Lambda_x - \tau^{-2} I + e^{-(\epsilon_\mu+\epsilon_\sigma)} \frac{|\Lambda_x - \Lambda_{x'}|}{\min(\Lambda_x, \Lambda_{x'}) + \frac{\epsilon}{\gamma}} \right)., \tag{15}$$

This shows that the DR-FIM $\mathbf{DRF}_\theta$ can be estimated from the covariance matrix Λ^{-1}, with γ and τ^2 as hyperparameters.

B. Datasets and Setup.

Datasets *GTA5* [5]: Derived from the Grand Theft Auto V video game, this dataset provides 24,966 synthetic urban images with pixel-level annotations for 19 categories. The images, at a resolution of 1914x1052 pixels, simulate diverse urban driving scenarios. *Cityscapes* [2]: This dataset contains 5,000 finely annotated real-world urban images from 50 German cities, with a resolution of 2048x1024 pixels. It includes pixel-level annotations for 19 categories. *ACDC* [8]: A real-world dataset designed for adverse visual conditions (fog, nighttime, rain, snow), featuring diverse weather scenes with a resolution of 1280x720 pixels. It is crucial for evaluating models under challenging conditions. *BDD100K* [10]: A large-scale dataset comprising 8,000 training images and 1,000 validation images, each at 1280x720 pixels. Covering various weather and lighting scenarios, it is widely used for domain generalization and segmentation tasks. *Foggy Zurich* [6]: Contains 1,552 unlabeled light and medium fog images and 40 labeled foggy scenes for evaluation. *Foggy Driving* [6]: Comprises 33 finely-annotated and 68 coarsely-annotated images under foggy conditions. *Dark Zurich* [7]: Includes 8,779 images from daytime, twilight, and nighttime conditions, with 50 annotated images for nighttime validation. *Nighttime Driving* [3]: Features 50 annotated nighttime driving images, serving as a benchmark for nighttime segmentation tasks.

Setup Following Rein [9], we integrated Mask2Former [1] with various Vision Foundation Models (VFMs) as backbones. Unlike Rein, we adopted the original decoding mechanism of Mask2Former, with slight differences from the modified version in Rein, as detailed in Fig. 1. For training, we used the AdamW optimizer [4] with a backbone learning rate of 1×10^{-5} and a decoder learning rate of 1×10^{-4}, consistent with the suggested configuration in Rein. We conducted 40,000 iterations with a batch size of 4, where input images were cropped to a resolution of 512×512. The training process was divided into three stages: a warm-up phase lasting T_1 = 10,000 iterations to adapt the decoder, the DR-FIM estimation phase over T_2 = 2,000 iterations, and a fine-tuning phase spanning T_3 = 28,000 iterations. Regularization and variance coefficients, γ and τ, were set to 0.01 and 0.01, respectively. ϵ_μ and ϵ_σ are randomly sampled from a Gaussian distribution $\sim \mathcal{N}(0,1)$. The dynamic parameter selection strategy began with fine-tuning the top 1% ($\delta_{\min}$) of parameters ranked by DR-FIM and expanded progressively to 15% ($\delta_{\max}$) throughout the training process.

C. The Ratio of Fine-tuned Parameters.

The impact of fine-tuning parameter proportions on performance for DINOV2 and SAM VFMs is illustrated in Fig. 2. We controlled $\delta_{\max}$ to vary from 0.01 (approximately 3M parameters) to 0.1 (approximately 30M parameters). The experiment reveals the following key findings: (1). When fine-tuning a small number of parameters (around 3M), our method performs better than the Freeze approach but falls short compared to Rein. (2). As the number of fine-tuned parameters increases to approximately 15M, our method shows a significant improvement in performance, surpassing Rein. (3). Further increasing the fine-tuned parameters to approximately 30M results in no substantial performance gains, with the performance tending to stabilize. We attribute these results to the following analysis: Rein improves task adaptability by introducing additional structures to the model, while our method does not add any extra structures. Consequently, our approach requires fine-tuning more VFM parameters to effectively activate the task-specific adaptability of the model.

D. Segmentation Result Visualization.

As shown in Fig.4-Fig.6, our method achieves better segmentation structure and semantic prediction than Rein under various severe weather and extreme conditions.

E. Influence of Hyper-parameters.

Coefficient γ and τ in Eq.(15). The regularization coefficient γ and variance coefficient τ are the most influential parameters in our method, as they jointly determine the adjustment of the expected parameter distribution and directly influence the prediction of the DR-FIM. We explore the combined impact of these parameters on performance,

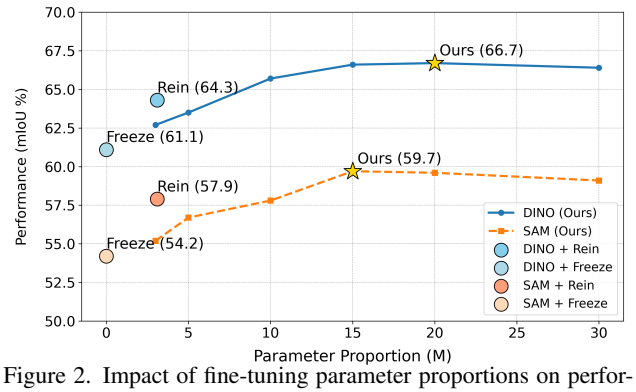

Figure 2. Impact of fine-tuning parameter proportions on performance for DINOV2 and SAM VFMs, reported as the average metric across GTAV → Cityscapes, BDD100K, and Mapillary tasks.

reporting the average mIoU across GTAV → Cityscapes, → BDD100K, and → Mapillary using DINOv2, as shown in Fig. 3. Overall, the results indicate that the impact of both parameters remains within an acceptable range when γ and τ are set between 0.005 and 0.02. Specifically: When γ becomes excessively large, the performance drops significantly, reducing the mIoU score by 1.7%. This demonstrates that overly enforcing the target distribution to align with the pre-trained distribution decreases the model's adaptability. When τ becomes excessively large, the performance also drops significantly, reducing the mIoU score by 1.3%. This indicates that a large expected fluctuation in the target distribution reduces the model's generalization ability.

The above parameter analysis demonstrates that setting γ and τ within the range of 0.005 to 0.02 ensures acceptable performance. Furthermore, this highlights the importance of maintaining a smoothly adjusted posterior parameter distribution to preserve the effectiveness of our method.

Combination coefficient of Eq.(7). We conducted an ablation study on the coefficient λ in Eq. (7), formulated as $(1-\lambda)\mathbf{F}_{\boldsymbol{\theta}} + \lambda\Delta\mathbf{F}_{\boldsymbol{\theta}}$. As shown in the table, FisherTune remains stable within $\lambda \in [0.4, 0.7]$, indicating insensitivity to this parameter. When λ is too small, domain-sensitive Fisher information is underutilized, and when λ is too large, overemphasizing $\Delta\mathbf{F}_{\boldsymbol{\theta}}$ slightly reduces task relevance.

λ	0	0.1	0.3	0.4	0.5	0.6	0.7	0.9	1
CS → ACDC (EVA02)	69.5	70.1	71.5	72.5	**72.9**	72.7	72.6	71.7	71.3
CS → ACDC (DINOV2)	71.4	72.8	74.9	76.9	**77.5**	77.2	77.0	76.6	76.1

Table 1. Impact of Combination coefficient.

Training Strategy We use sequential execution (SE) of DR-FIM and PEFT, as described in the Step 2 and Step 3 of Algorithm 1. We further tested different progressive training strategies (linear, exponential, and cosine annealing) for parameter scheduling, and found that linear and exponential approaches perform better. We explored alternating execution (AE) of DR-FIM and PEFT with 2k, 4k, and 6k iterations after warm-up, but SE outperforms AE, likely due to conflicts between DR-FIM and PEFT.

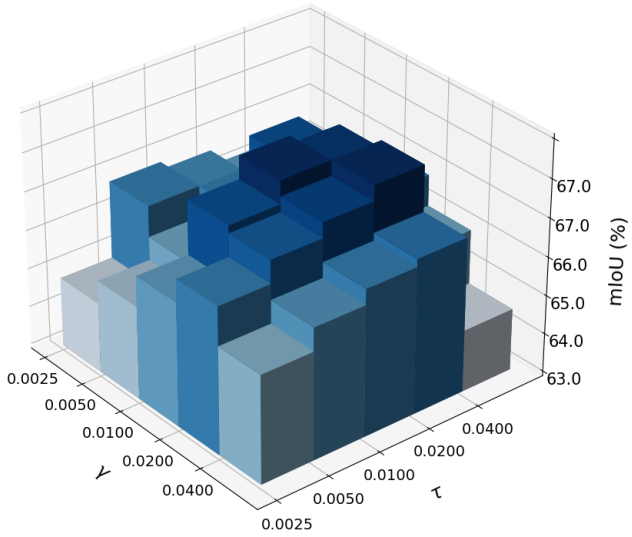

Figure 3. Impact of the regularization coefficient γ and variance coefficient τ on generalization performance.

VFM	Task	Linear	Exp (Ours)	Cosine	Alter-2K	Alter-4K	Alter-6K
DINOv2	CS→BDD	67.6	**67.7**	66.1	65.2	65.4	65.9
DINOv2	CS→ACDC	77.3	**77.5**	76.2	76.1	76.2	76.8

Table 2. Impact of fine-tuning Strategy.

F. GPU memory.

As shown in Table below, our method is more efficient than Full Tuning, slightly costlier than Rein, but outperforms both in performance, achieving a better balance between cost and performance.

	Method	GPU Memory (GB)	Training Time (Hours)
DINOv2-L	Full Tuning	14.7	11.2
	Rein	10.0	9.5
	FisherTune (Ours)	12.5	10.3
EVA02-L	Full Tuning	15.9	11.8
	Rein	12.5	10.0
	FisherTune (Ours)	13.5	10.5

Table 3. GPU Memory and Training Time.

G. Limitations.

While our method provides a novel and feasible approach to estimating domain- and task-sensitive parameters in VFMs, effectively activating their adaptability for DGSS tasks while preserving their strong generalization capabilities, it also introduces a more complex training process and increases the number of optimization parameters. Addressing these complexities and optimizing the training efficiency will be the focus of our future efforts.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask

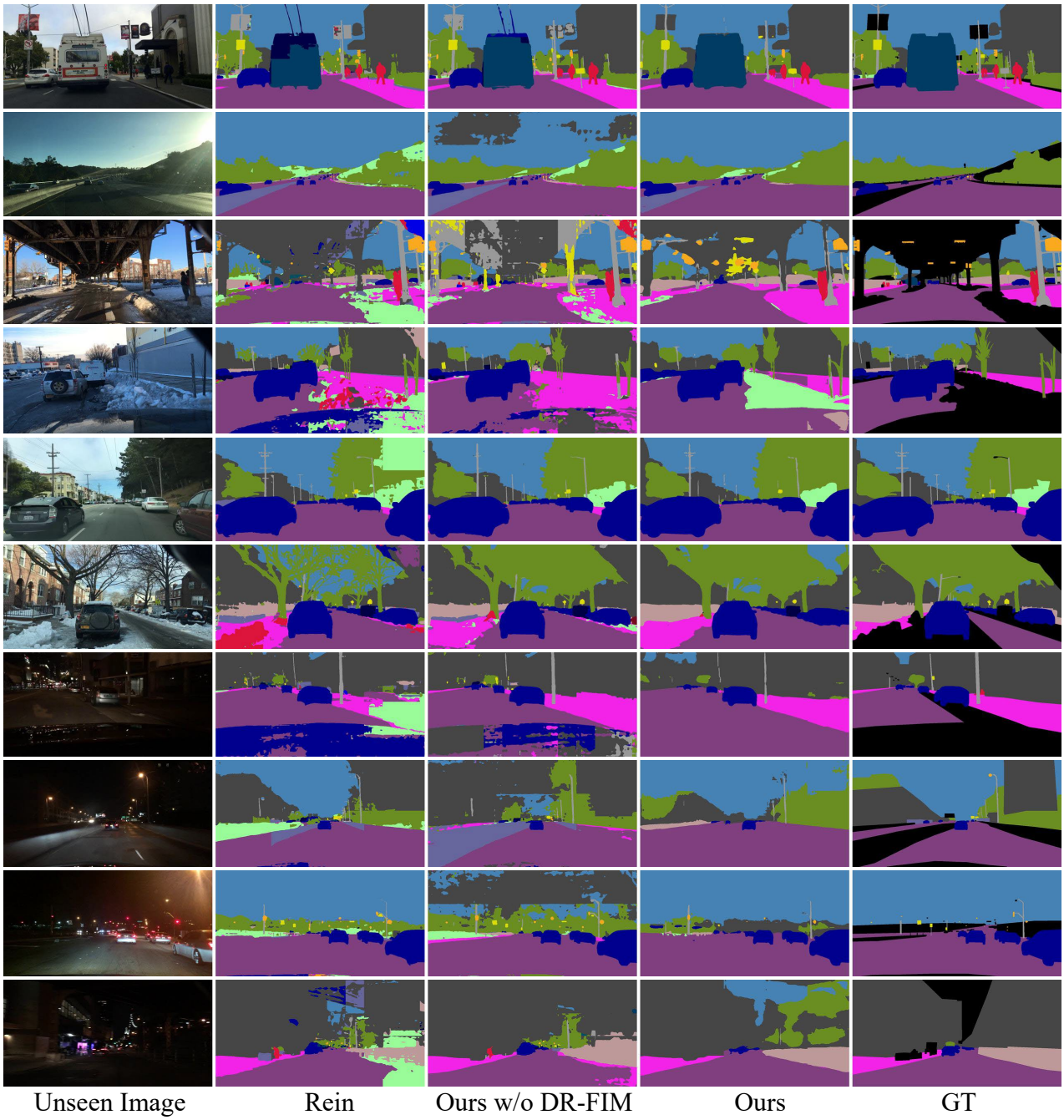

Figure 4. Domain generalization segmentation visualization results on GTAV → BDD100K.

transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2

[3] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *IEEE TPAMI*, pages 3819–3824, 2018. 2

[4] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[5] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 102–118, 2016. 2

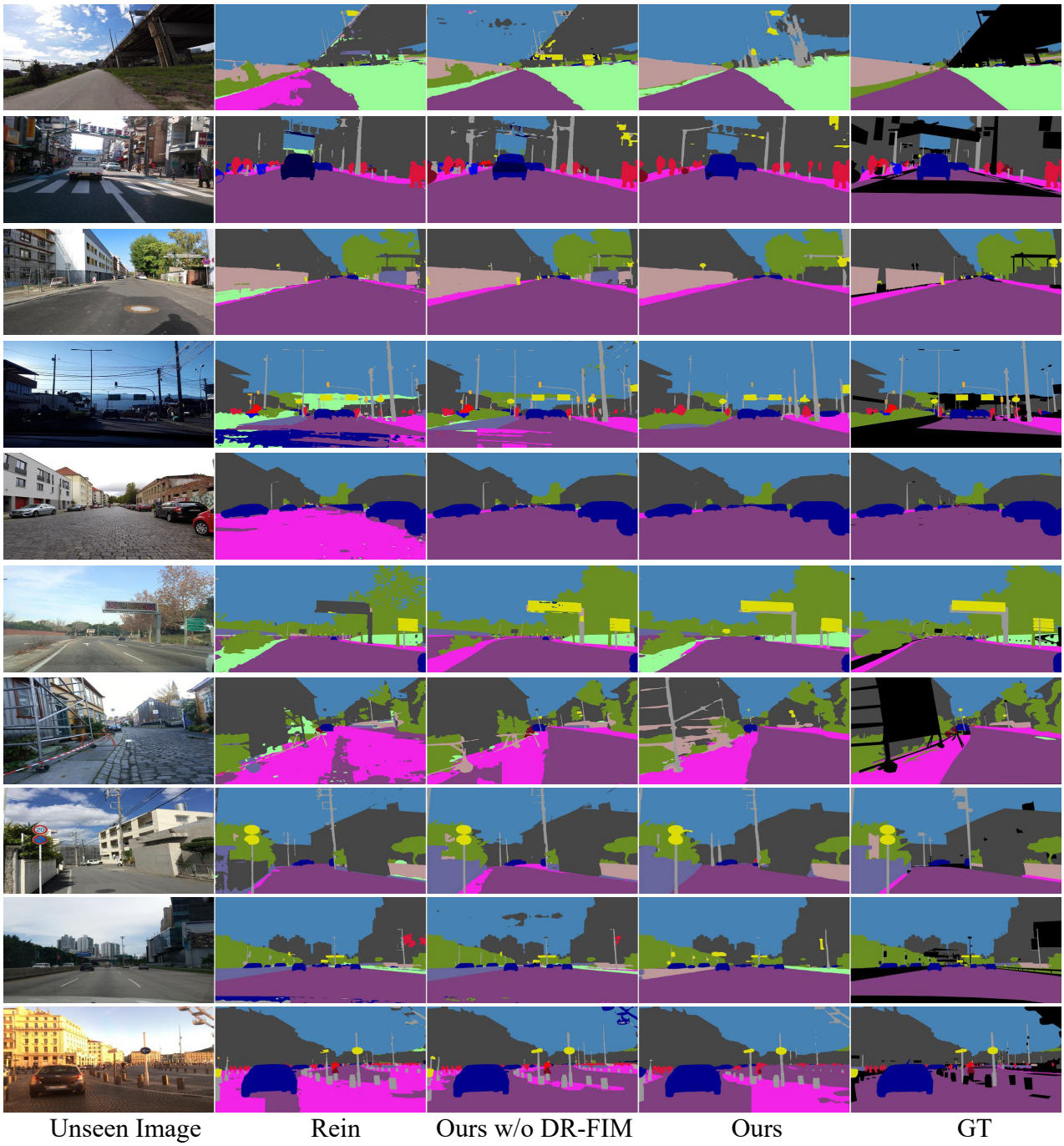

Figure 5. Domain generalization segmentation visualization results on GTAV → Mapillary.

[6] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *ECCV*, pages 687–704, 2018. 2

[7] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, pages 7374–7383, 2019. 2

[8] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 2

[9] Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. Stronger fewer & superior: Harnessing vision foundation

Figure 6. Domain generalization segmentation visualization results on Cityscapes → ACDC.

models for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28619–28630, June 2024. 2

[10] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 2