Supplementary Material
GeCo: Geometry-Consistent Regularization for Domain Generalized Semantic Segmentation

Qi Zang[1*], Dong Zhao[2*], Nan Pu[1], Wenjing Li[1], Zhun Zhong[1†], Meng Wang[1†]
[1] School of Computer Science and Information Engineering, Hefei University of Technology, China
[2] Department of Information Engineering and Computer Science, University of Trento, Italy

In this supplementary material, we provide additional details and analyses to complement the main paper. We first describe the implementation settings and training pipeline used for all experiments, including open-set extensions. We then present the full derivation and geometric interpretation of the proposed geodesic regularization, followed by detailed ablation studies that dissect component-wise contributions and hyperparameter sensitivity in both closed-set and open-set DGSS scenarios. Next, we provide extended qualitative comparisons under various domain shifts. Finally, we present additional token visualizations on more source domains and a dedicated comparison of token distributions for the pre-trained DINOv2-L backbone, GeCo, and RP+MSE, which further illustrate the manifold structure of VFM embeddings and the effect of geometry-aware adaptation.

1. Implementation Details

We follow the standard training pipeline used in recent VFM-based segmentation works. The DINOv2-L backbone is frozen except for the LoRA adapters injected into the attention and MLP layers, while the decoder is trained from scratch. We adopt Mask2Former [2] as the default segmentation head unless otherwise specified. For optimization, we use AdamW with a learning rate of 1×10^{-4} for the decoder and 2×10^{-4} for all LoRA [6] parameters, together with a weight decay of 1×10^{-2}. Following common practice, we train for 40k iterations with a batch size of 4 and employ a crop size of 512×512. The data augmentation strictly follows the default configuration of MMSegmentation (random crop, flip, and photometric jitter), without additional domain-specific tuning.

Our method introduces several geometry-related hyperparameters: the neighborhood radius η, perturbation magnitude α, prototype number M_c, and geodesic consistency weight λ_{geo}. To ensure fair comparison and avoid overtuning, all other hyper-parameters are fixed to their recommended default values. We adopt $\eta = 0.75$, $\alpha = 0.5$, $M_c = 50$, and $\lambda_{\text{geo}} = 0.75$ as the default configuration in all reported experiments. Thanks to the lightweight nature of our geometric modules and the limited number of trainable parameters introduced by LoRA, our full model can be trained efficiently on a single RTX 3090Ti GPU, with large backbones such as DINOv2-L remaining fully feasible.

Open-set Details. For the open-set experiments, we reproduce a contrastive-learning baseline following RPL [7], where the source domain (Cityscapes or GTA5) is treated as in-distribution (ID) data and a subset of COCO-Stuff [1] is used as synthetic out-of-distribution (OOD) samples. This setup encourages pixel embeddings of ID and OOD regions to be contrastively separated. During training, ID and OOD batches are jointly sampled: segmentation loss is applied to ID labels, while a pixel-level contrastive loss is applied between ID and OOD embeddings. All optimization settings follow our main implementation details.

2. Detailed Derivation and Analysis of Regularization

For completeness, we provide the explicit derivation of Eq. (12) in the main paper. For a given token t_i, the geodesic distance between prediction vectors $\mathbf{p}_i, \mathbf{p}'_i \in \mathbb{R}^C$ is defined as,

$$\begin{aligned} d_{\text{geo}}(\mathbf{p}_i, \mathbf{p}'_i) &= \arccos\big(u_i\big), \\ u_i \triangleq \langle \sqrt{\mathbf{p}_i}, \sqrt{\mathbf{p}'_i} \rangle &= \sum_{m=1}^{C} \sqrt{p_i^{(m)}\, p_i'^{(m)}}, \end{aligned} \tag{1}$$

and the per-sample squared objective is,

$$\ell_{\text{geo}}(\mathbf{p}_i) = d^2_{\text{geo}}(\mathbf{p}_i, \mathbf{p}'_i). \tag{2}$$

Since $\mathcal{L}_{\text{geo}} = \mathbb{E}_{t_i}[\ell_{\text{geo}}(\mathbf{p}_i)]$, it suffices to compute the gradient of $\ell_{\text{geo}}(\mathbf{p}_i)$ with respect to each class probability $p_i^{(c)}$.

Step 1: Chain rule on the squared distance. We denote,

$$\theta_i \triangleq d_{\text{geo}}(\mathbf{p}_i, \mathbf{p}'_i) = \arccos(u_i). \tag{3}$$

Applying the chain rule to $\ell_{\text{geo}}(\mathbf{p}_i) = \theta_i^2$ gives,

$$\frac{\partial \ell_{\text{geo}}}{\partial p_i^{(c)}} = \frac{\partial(\theta_i^2)}{\partial p_i^{(c)}} = 2\theta_i \frac{\partial \theta_i}{\partial p_i^{(c)}}. \tag{4}$$

Step 2: Derivative of the arccosine. Using the chain rule again,

$$\frac{\partial \theta_i}{\partial p_i^{(c)}} = \frac{\partial \arccos(u_i)}{\partial u_i} \cdot \frac{\partial u_i}{\partial p_i^{(c)}} = -\frac{1}{\sqrt{1-u_i^2}} \cdot \frac{\partial u_i}{\partial p_i^{(c)}}. \tag{5}$$

Step 3: Derivative of the inner product u_i. Recall that,

$$u_i = \sum_{m=1}^{C} \sqrt{p_i^{(m)}\, p_i'^{(m)}}, \tag{6}$$

where m is a dummy index over classes. When differentiating with respect to $p_i^{(c)}$, all terms with $m \neq c$ do not involve $p_i^{(c)}$ and are therefore constants (their derivatives are zero). Only the term with $m = c$ depends on $p_i^{(c)}$,

$$\frac{\partial u_i}{\partial p_i^{(c)}} = \frac{\partial}{\partial p_i^{(c)}} \sqrt{p_i^{(c)}\, p_i'^{(c)}} = \frac{1}{2} \sqrt{\frac{p_i'^{(c)}}{p_i^{(c)}}}. \tag{7}$$

Step 4: Putting everything together. Substituting Eq. (5) and Eq. (7) into Eq. (4), we obtain,

$$\frac{\partial \ell_{\text{geo}}}{\partial p_i^{(c)}} = 2\theta_i \left(-\frac{1}{\sqrt{1-u_i^2}} \cdot \frac{1}{2} \sqrt{\frac{p_i'^{(c)}}{p_i^{(c)}}} \right) \tag{8}$$

$$= -\frac{\theta_i}{\sqrt{1-u_i^2}} \sqrt{\frac{p_i'^{(c)}}{p_i^{(c)}}}. \tag{9}$$

Finally, replacing $u_i = \langle \sqrt{\mathbf{p}_i}, \sqrt{\mathbf{p}_i'} \rangle$ and $\theta_i = d_{\text{geo}}(\mathbf{p}_i, \mathbf{p}_i')$ yields,

$$\frac{\partial \ell_{\text{geo}}}{\partial p_i^{(c)}} = -\frac{d_{\text{geo}}(\mathbf{p}_i, \mathbf{p}_i')}{\sqrt{1 - \langle \sqrt{\mathbf{p}_i}, \sqrt{\mathbf{p}_i'} \rangle^2}} \frac{\sqrt{p_i'^{(c)}}}{\sqrt{p_i^{(c)}}}, \tag{10}$$

which matches Eq. (12) in the main paper. In practice, when $\mathbf{p}_i$ and $\mathbf{p}_i'$ are nearly identical, the term $1 - \langle \sqrt{\mathbf{p}_i}, \sqrt{\mathbf{p}_i'} \rangle^2$ can become numerically very close to zero, leading to unstable gradients. We therefore use a stabilized denominator $\sqrt{1 - \langle \sqrt{\mathbf{p}_i}, \sqrt{\mathbf{p}_i'} \rangle^2 + \varepsilon}$, with a small constant $\varepsilon = 10^{-6}$ in all experiments.

To further clarify the behavior of this regularization, we next provide a more detailed geometric interpretation of the above distance than the main paper.

Geometric Remarks. Our geodesic distance is defined on the probability simplex. Since $\mathbf{p}_i$ and $\mathbf{p}_i'$ are probability vectors with non-negative entries summing to one, their square-root embeddings $\sqrt{\mathbf{p}_i}$ and $\sqrt{\mathbf{p}_i'}$ lie on the positive orthant of the unit ℓ_2 sphere, *i.e.*,

$$\left\| \sqrt{\mathbf{p}_i} \right\|_2^2 = \sum_c p_i^{(c)} = 1, \quad \left\| \sqrt{\mathbf{p}_i'} \right\|_2^2 = \sum_c p_i'^{(c)} = 1. \tag{11}$$

Therefore, the inner product

$$u_i = \left\langle \sqrt{\mathbf{p}_i}, \sqrt{\mathbf{p}_i'} \right\rangle \tag{12}$$

is bounded in $[0, 1]$, and the geodesic angle

$$d_{\text{geo}}(\mathbf{p}_i, \mathbf{p}_i') = \arccos(u_i) \tag{13}$$

satisfies

$$d_{\text{geo}}(\mathbf{p}_i, \mathbf{p}_i') \in [0, \tfrac{\pi}{2}]. \tag{14}$$

In other words, the disagreement between two predictions is measured as a bounded angular deviation on the unit hypersphere, rather than an unbounded Euclidean distance.

Moreover, in the small-angle regime where $\mathbf{p}_i$ and $\mathbf{p}_i'$ are close, we have $u_i \approx 1$ and the standard Taylor expansion of $\arccos(\cdot)$ yields

$$d_{\text{geo}}(\mathbf{p}_i, \mathbf{p}_i') = \arccos(u_i) \approx \sqrt{2\,(1-u_i)}. \tag{15}$$

Since

$$1 - u_i = 1 - \sum_c \sqrt{p_i^{(c)} p_i'^{(c)}} \tag{16}$$

is closely related to the Hellinger divergence between $\mathbf{p}_i$ and $\mathbf{p}_i'$, this shows that our geodesic consistency behaves like a Euclidean-style consistency loss on the square-root embeddings in the vicinity of the optimum, ensuring stable and familiar optimization behaviour near convergence, while for larger prediction discrepancies the full geodesic distance still preserves a curvature-aware angular geometry on the probability simplex and avoids the distortions introduced by treating probabilities as flat Euclidean vectors.

3. Detailed Ablation Studies

In this section, we provide additional ablations to better understand GeCo. We first examine component-wise variants in the open-set DGSS setting, then present class-wise results for closed-set DGSS on GTA5 $\rightarrow$ BDD-100K, and finally analyze the sensitivity to key hyperparameters (η, α, M_c, and λ_{geo}).

3.1. Ablation on Open-Set DGSS

As shown in Table 2, simply combining Gaussian random perturbation (RP) with the MSE consistency loss brings only modest gains over full fine-tuning in the open-set

Method	road	side.	build.	wall	fence	pole	light	sign	vege.	terr.	sky	pers.	rider	car	truck	bus	train	motor.	bicy.	mIoU
Full fine-tuning	89.1	60.0	83.9	31.9	47.3	43.8	54.5	42.4	74.8	46.6	85.9	67.0	19.7	81.4	66.1	79.5	2.8	62.3	51.4	57.4
RP+MSE	94.8	62.7	86.7	32.3	48.9	46.1	54.6	45.3	76.0	48.1	87.4	69.4	20.8	87.3	68.1	79.7	7.1	66.4	53.7	59.9
CGP+MSE	92.3	66.7	89.1	32.5	52.1	47.6	58.3	46.0	80.7	48.2	87.5	70.0	22.5	91.7	72.7	**83.0**	17.2	68.2	**56.4**	62.2
RP+GBR	**93.4**	65.9	84.7	32.5	49.2	46.1	55.1	44.4	75.4	47.8	**88.0**	67.6	20.5	85.3	66.7	79.3	7.1	69.1	54.0	59.5
GeCo (Ours)	93.0	**66.8**	**93.9**	**32.6**	**53.1**	**49.5**	**59.1**	**46.9**	**84.1**	**48.7**	87.0	**70.4**	**23.7**	**95.4**	**74.6**	81.8	**38.8**	**70.3**	56.1	**64.6**

Table 1. Ablation results on the GTA5 → BDD-100K closed-set domain generalization task using the DINOv2-L backbone. The IoU score for each category is reported. Full fine-tuning is the baseline.

Method	ACDC-POC			MUAD		
	AP↑	FPR_{95}↓	mIoU	AP↑	FPR_{95}↓	mIoU
Full fine-tuning	90.6	0.5	61.6	51.8	20.7	37.7
RP+MSE	91.5	0.4	62.1	54.3	18.8	38.3
CGP+MSE	92.5	0.3	63.6	55.2	16.9	40.2
RP+GBR	89.9	0.4	62.0	53.3	19.5	37.9
GeCo (Ours)	**93.5**	**0.3**	**64.7**	**58.6**	**15.1**	**41.5**

Table 2. Ablation results on the Cityscapes → ACDC-POC + MUAD open-set domain generalization task using the DINOv2-B backbone. Full fine-tuning is the baseline. The results report scores for both known (AP↑, FPR_{95}↓) and unknown (mIoU↑) classes. All results are obtained by combining the above variants with the RPL [7] to perform contrastive learning between in-domain and out-of-domain data.

η	Test Domains (mIoU)			
	Cityscapes	BDD-100K	Mapillary	Average
0.25	66.8	60.8	65.7	64.4
0.50	68.5	61.8	67.4	65.9
0.75	68.5	**64.6**	**69.9**	**67.7**
1.00	**68.8**	63.7	68.4	67.0
1.25	68.3	61.6	67.1	65.7
1.50	64.7	61.8	67.5	64.7

Table 3. Sensitivity analysis of the hyper-parameter η on the GTA5 → Cityscapes + BDD-100K + Mapillary domain generalization task using the DINOv2-L backbone.

DGSS setting: AP and mIoU on ACDC-POC [4] and MUAD [5] improve slightly, while FPR_{95} remains relatively high, indicating that unknown-region responses are still poorly calibrated. Replacing RP with the proposed curvature-guided perturbation (CGP) yields consistent improvements on both datasets, with higher AP and mIoU and notably lower FPR_{95}, showing that geometry-aware perturbations generate more informative variations for separating in-domain and out-of-domain samples without breaking the pretrained representation manifold. In contrast, using the geodesic-based regularization (GBR) alone (RP+GBR) stabilizes FPR_{95} and slightly improves mIoU, but does not match the gains achieved by CGP, suggesting that curvature-aware perturbation is the primary driver for enhancing open-set discrimination. The full configuration GeCo (CGP+GBR) achieves the best performance on both ACDC-POC and MUAD, attaining the highest AP and mIoU together with the lowest FPR_{95}. These results indicate that combining curvature-guided perturbations with geodesic-consistent regularization is particularly effective for open-set DGSS, as it yields more reliable detection of unknown regions while maintaining strong segmentation quality on known classes.

3.2. Class-wise Component Analysis

As shown in Table 1, the per-class ablation on GTA5 → BDD-100K confirms that GeCo improves not only the overall mIoU but also the performance on most individual categories. Naive random perturbation with MSE (RP+MSE) mainly boosts large, frequent classes such as *road*, *building* and *car*, but brings limited gains on structure-sensitive or rare categories. Introducing curvature-guided perturbation (CGP+MSE) yields broader benefits, notably on thin objects and dynamic classes such as *pole*, *traffic light*, *truck*, *bus* and *train*, indicating that geometry-aware perturbations help the model better preserve fine structures under domain shift. Geodesic-based regularization (RP+GBR) further stabilizes several large-scale categories, including *road* and *sky*, reflecting its effect in constraining feature updates along manifold-consistent directions. The full configuration GeCo (CGP+GBR) achieves the highest overall mIoU (64.6%) and secures the best IoU on a large portion of classes, especially for challenging categories such as *building*, *fence*, *pole*, *traffic light*, *traffic sign*, *vegetation*, *person*, *rider*, *car*, *truck*, *train* and *motorcycle*. These results demonstrate that combining curvature-guided perturbation with geodesic-consistent regularization leads to more uniformly improved generalization across both dominant and rare classes in the closed-set DGSS setting.

3.3. Impact of Hyperparameters

Sensitivity to the neighborhood radius η. As shown in Table 3, the performance of GeCo is sensitive to the choice of the neighborhood radius η used to construct $\mathcal{N}_{\text{proto}}(t_i)$ and estimate local curvature. For very small values ($\eta \leq 0.25$), the neighborhood around each token becomes too sparse, so the PCA-based tangent space and curvature proxy (Eq.(4)) are estimated from a few nearly collinear samples. This leads to unstable curvature estimates and noisy perturbation magnitudes, resulting in degraded average mIoU (64.4%). Increasing η to the medium range (0.5–1.0) yields more re-

α	Test Domains (mIoU)			
	Cityscapes	BDD-100K	Mapillary	Average
0.00	63.7	57.4	64.2	61.8
0.10	64.0	62.4	66.0	64.1
0.25	67.8	64.6	68.9	67.1
0.50	**68.5**	**64.6**	**69.9**	**67.7**
0.75	67.7	63.9	68.8	66.8
0.90	67.5	60.3	65.3	64.4
1.00	64.4	60.7	65.2	63.4

Table 4. Sensitivity analysis of the hyper-parameter α on the GTA5 $\rightarrow$ Cityscapes + BDD-100K + Mapillary domain generalization task using the DINOv2-L backbone.

liable local geometry: neighborhoods are large enough to capture meaningful semantic variation while still dominated by tokens with similar semantics, and the performance steadily improves, with the best average mIoU achieved at η=0.75 (67.7%) and a comparable score at η=1.0 (67.0%). When η is further enlarged ($\eta > 1.0$), the neighborhood starts to mix tokens from visually and semantically different regions, which blurs manifold boundaries and weakens the contrast between high- and low-curvature areas; as a consequence, the curvature-guided perturbation becomes less discriminative and the average mIoU drops again (65.7% at η=1.25, 64.7% at η=1.5). Taken together, these observations indicate that a moderate neighborhood scale is crucial: it yields a stable yet semantically coherent local manifold for curvature estimation, which in turn makes the curvature-guided perturbation more informative.

Sensitivity to the perturbation scaling factor α. As shown in Table 4, setting α=0 (*i.e.*, removing the curvature-guided perturbation term so that no geometry-aware perturbation is injected) yields the worst average performance (61.8% mIoU), which confirms that the perturbation branch provides a necessary complementary signal on top of the geodesic module. As we gradually increase α from 0.1 to 0.5, the performance consistently improves across all three target domains, with the best average mIoU obtained at α=0.5 (67.7%). This trend indicates that using α to moderately scale the curvature-guided perturbation magnitude is beneficial: it amplifies meaningful local variations on the token manifold (as determined by Eq. (6) in the main paper) and helps learn decision boundaries that are more robust to domain shift. However, when α becomes too large ($\alpha \geq 0.75$), the gains start to saturate and eventually degrade, as overly amplified perturbations distort the underlying semantics and mix tokens from different regions of the manifold, turning the perturbation signal into harmful noise. Hence, curvature-guided perturbations are effective only when the scaling factor α stays in a moderate range, where the perturbation branch strengthens rather than undermines the geodesic-based regularization module.

Sensitivity to the number of prototypes M_c. As shown in Table 5, the number of prototypes M_c per class, which controls how many dominant shapes/modes are captured for each category, has a clear impact on domain generalization. When M_c=0, the model degenerates to using only a global representation per class (without explicit multi-prototype modeling), leading to the lowest average mIoU (64.3%) and indicating that the popular shapes within each class are poorly characterized and the resulting curvature-guided perturbations are inaccurate. Increasing M_c from 10 to 20 gradually improves performance, as more prototypes allow the model to better approximate the diverse modes of each semantic class and provide more reliable local neighborhoods for curvature estimation. The best average mIoU is achieved at M_c=50 (67.7%), which strikes a balance between capturing sufficient intra-class diversity and keeping the prototype bank compact. Further increasing M_c to 100 brings only marginal changes and slightly reduces the average score, while incurring higher computational and memory costs. In summary, using a moderate number of prototypes per class (*e.g.*, M_c=50) is sufficient to model prevalent shapes and support effective geometry-aware perturbations, whereas too few prototypes underfit the manifold and too many mainly add overhead without clear gains.

M_c	Test Domains (mIoU)			
	Cityscapes	BDD-100K	Mapillary	Average
0.00	65.9	59.6	67.3	64.3
10.00	67.0	61.9	67.6	65.5
20.00	67.0	63.5	69.0	66.5
50.00	**68.5**	**64.6**	69.9	**67.7**
100.00	68.5	64.1	**70.0**	67.5

Table 5. Sensitivity analysis of the hyper-parameter M_c on the GTA5 $\rightarrow$ Cityscapes + BDD-100K + Mapillary domain generalization task using the DINOv2-L backbone.

λ_{geo}	Test Domains (mIoU)			
	Cityscapes	BDD-100K	Mapillary	Average
0.00	63.7	57.4	64.2	61.8
0.25	64.8	63.1	67.3	65.1
0.50	67.6	**64.9**	68.7	67.1
0.75	**68.5**	64.6	**69.9**	**67.7**
1.00	68.3	64.1	69.3	67.2

Table 6. Sensitivity analysis of the hyper-parameter λ_{geo} on the GTA5 $\rightarrow$ Cityscapes + BDD-100K + Mapillary domain generalization task using the DINOv2-L backbone.

Sensitivity to the geodesic weight λ_{geo}. As shown in Table 6, the geodesic weight λ_{geo} in Eq. (11) of the main paper, which balances the geodesic consistency loss against the main segmentation loss, critically influences the final performance. When λ_{geo}=0, the geodesic term is removed and the model relies solely on the segmentation loss (plus the perturbation branch), yielding the lowest average mIoU (61.8%). This indicates that geometry-

aware consistency is an essential component for stabilizing adaptation on target domains. As λ_{geo} increases from 0.25 to 0.75, the performance steadily improves across all three test sets, with the best average mIoU obtained at λ_{geo}=0.75 (67.7%). This suggests that a moderately weighted geodesic loss effectively constrains feature updates along manifold-consistent directions without excessively restricting task-driven optimization. Further increasing λ_{geo} to 1.0 produces similar but slightly lower scores, implying that over-emphasizing the geodesic term can lead to over-regularization and weaken the model's ability to fit discriminative segmentation cues. Overall, λ_{geo}=0.75 offers a favorable trade-off between preserving the pretrained geometry and learning domain-robust decision boundaries.

4. Qualitative Comparison Results

To complement the qualitative examples in Fig. 5 of the main paper, we present additional visual comparisons on both closed-set and open-set DGSS settings. These results provide a more complete picture of how GeCo affects segmentation quality under challenging domain shifts.

On the Closed-Set DGSS. In the closed-set DGSS setting, we compare GeCo with Rein [11], SoMA [13] and GT (Ground Truth) on multi-target generalization across real-world datasets such as BDD-100K [12], Mapillary [8], and ACDC [10] (See Figs. 1-3). Across diverse scenes, GeCo produces more structurally coherent segmentations, with sharper object boundaries and fewer spurious label fragments in texture-rich regions (*e.g.*, building façades, vegetation, traffic signs).

On the Open-Set DGSS. In the open-set DGSS setting, we extend the qualitative comparison of Fig. 5 by showing more examples of both segmentation maps and unknown-region response maps (RM) on the unseen open-set dataset MUAD [5] (See Fig. 4). By comparing with Rein [11] and our proposed curvature-guided perturbation (CGP), GeCo produces cleaner and more localized unknown responses that align well with visually novel objects, while keeping the confidence on known categories stable. The corresponding semantic predictions exhibit fewer category collapses and less label bleeding at the interface between known and unknown regions.

5. More Token Visualization

To further support the observations in Fig. 2 of the main paper, we provide more visualizations of the frozen DINOv2-L [9] backbone on Cityscapes [3] in Fig. 5. We aggregate pixel-level token features from Cityscapes images and apply t-SNE to obtain a 2D embedding. The resulting map reveals discernible groups of tokens that roughly correspond to major semantic regions (*e.g.*, road, building, vegetation, car, sky), suggesting that the pretrained model already organizes dense pixel representations into semantically meaningful manifolds across scenes.

We also visualize the spatial structure of these features by projecting the high-dimensional tokens onto the first three principal components and rendering them as pseudo-RGB maps on the image plane. For each image, we further apply unsupervised clustering to its pixel-level token features, and display the resulting cluster assignments in the spatial domain. Pixels within the same semantic region tend to share similar colors and cluster memberships, and vary smoothly inside objects, while showing noticeable transitions at object boundaries. These additional visualizations reinforce our manifold assumption on VFM token embeddings and motivate the curvature-guided perturbation and geodesic regularization proposed in the main paper, which are designed to respect and exploit this structured organization rather than disrupt it with uninformed perturbations.

6. Comparison of Token Distributions

Figure 6 compares the t-SNE distributions of pixel-level token embeddings on ACDC for the pre-trained DINOv2-L backbone, GeCo, and the RP+MSE baseline. The pre-trained VFM already organizes tokens into roughly class-consistent clusters, but several categories remain partially entangled in the target domain. After adaptation with GeCo, the global manifold structure is largely preserved while the class-wise organization becomes richer: clusters are more compact, inter-class margins are clearer, and fine-grained substructures within each semantic region are better separated. This indicates that the geometry-aware perturbation and geodesic consistency refine the token manifold itself and, on top of that, sharpen the decision boundaries that partition it. In contrast, RP+MSE tends to fragment clusters and introduces local mixing between different classes (as highlighted in the insets), suggesting that random perturbations distort the pretrained geometry and lead to a less stable token organization under domain shift.

References

[1] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. *CoRR*, abs/1612.03716, 2016. 1

[2] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 1

[3] 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*

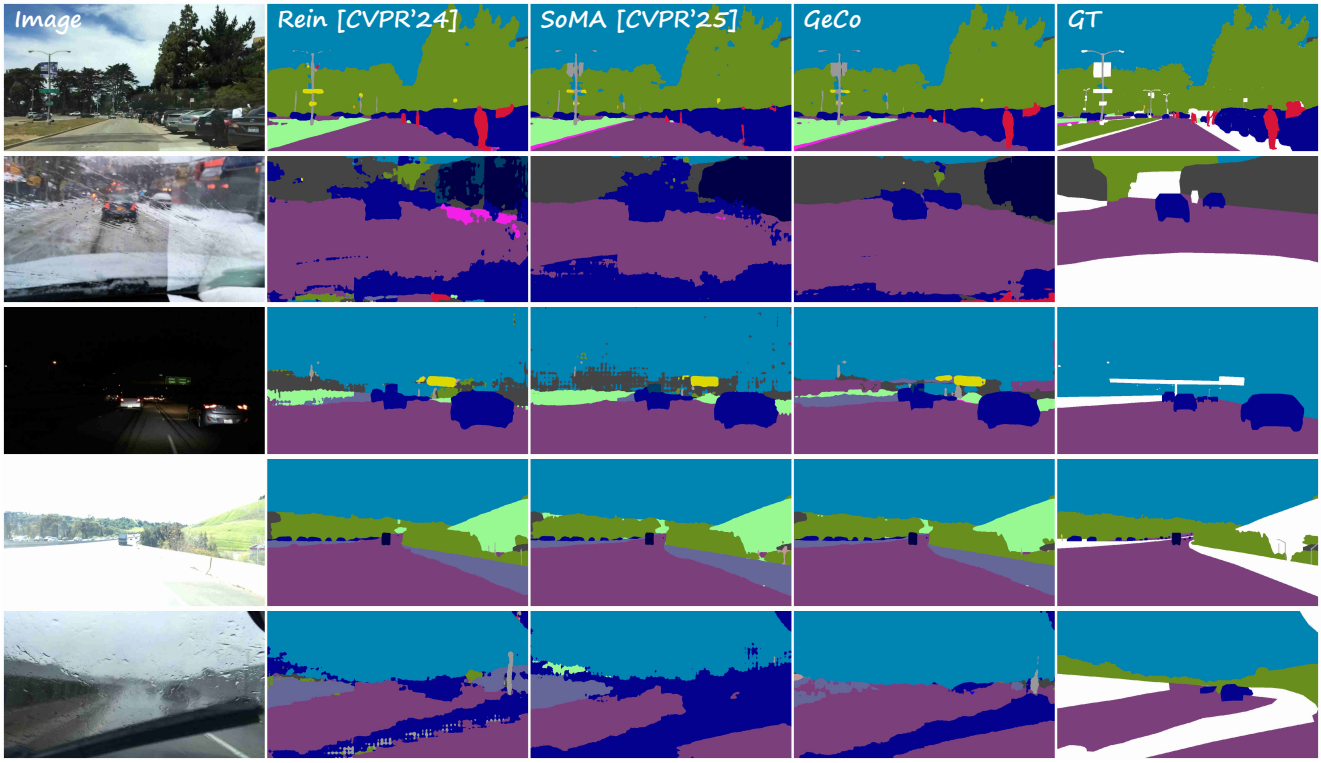


Figure 1. Qualitative comparison on the BDD-100K. The model is trained on GTA5 with DINOv2-L backbone.

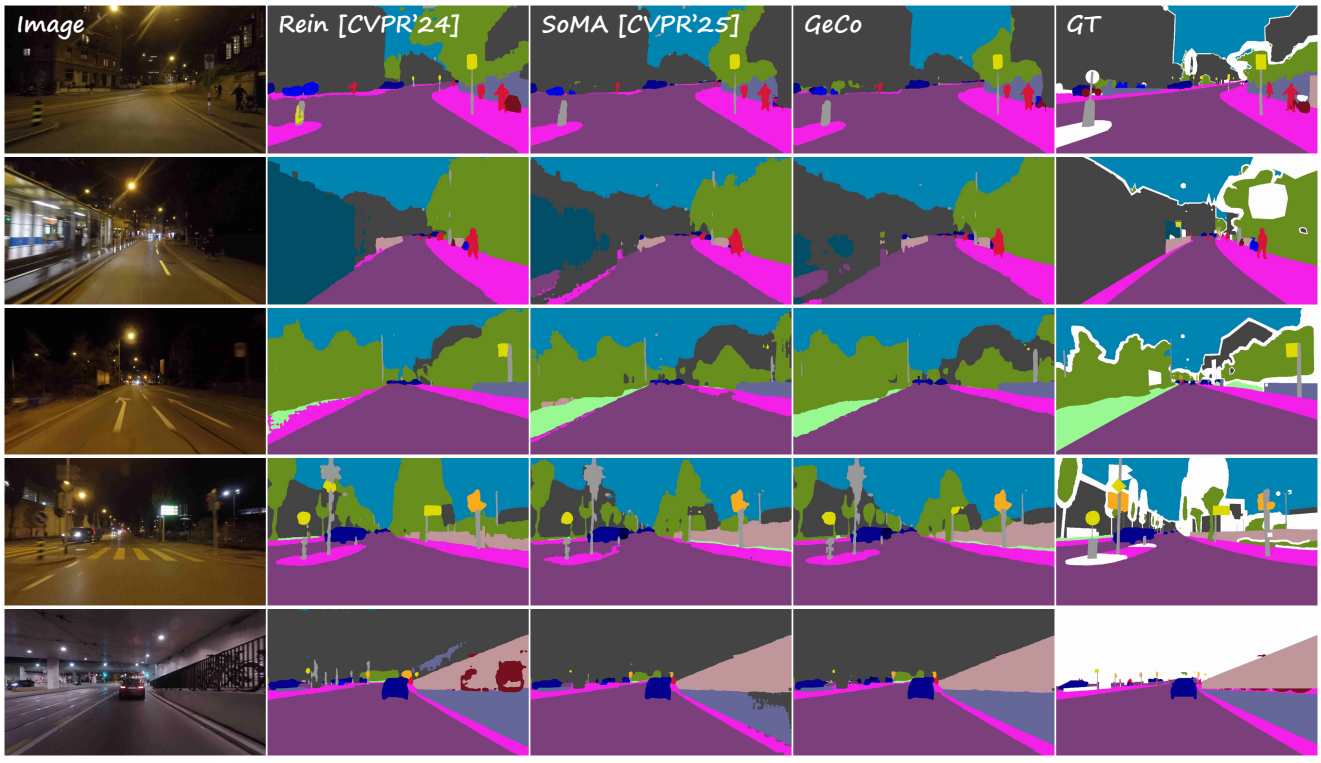


Figure 2. Qualitative comparison on the ACDC. The model is trained on Cityscapes with DINOv2-L backbone.

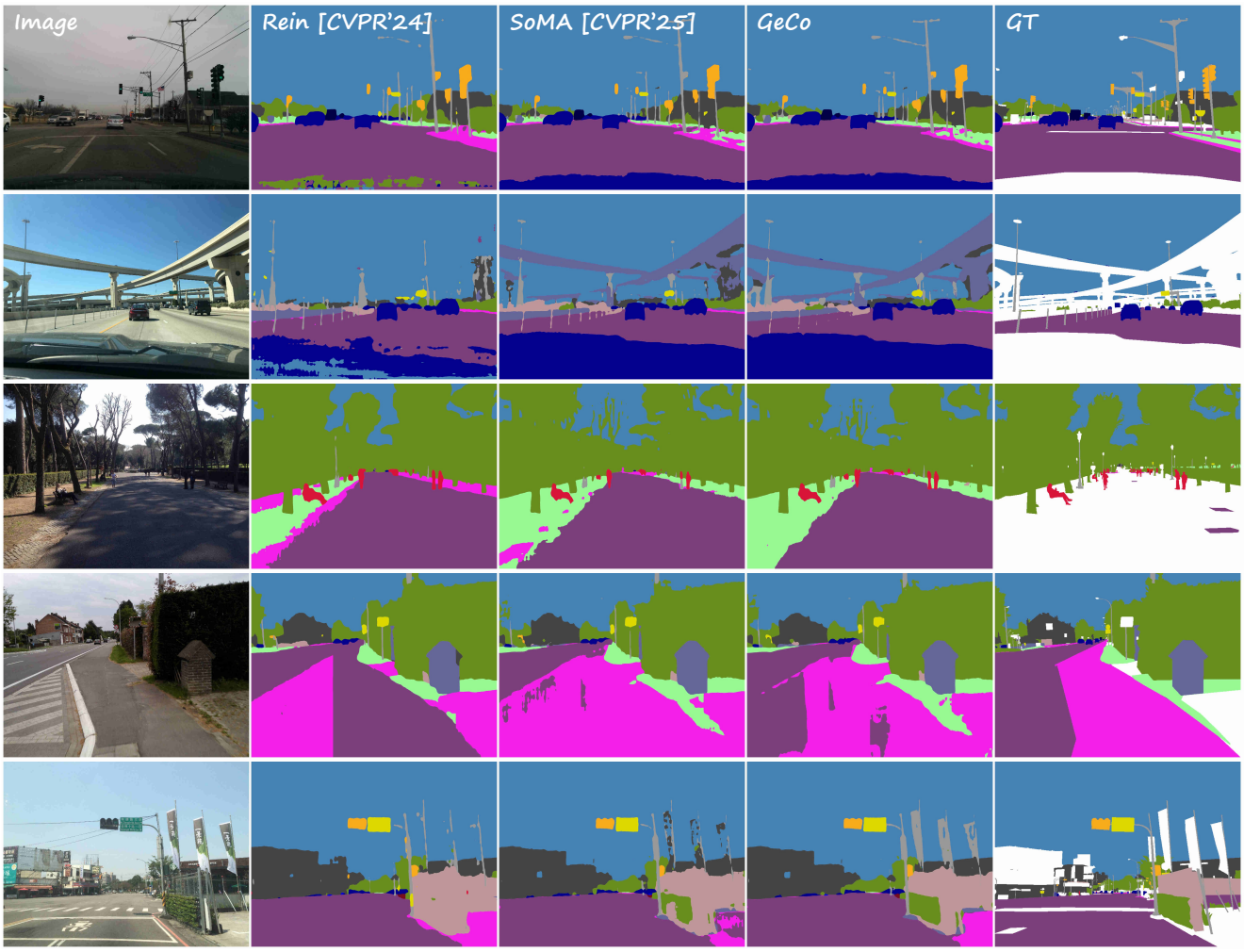


Figure 3. Qualitative comparison on the Mapillary. The model is trained on GTA5 with DINOv2-L backbone.

computer vision and pattern recognition, pages 3213–3223, 2016. 5

[4] Pau de Jorge, Riccardo Volpi, Puneet K Dokania, Philip HS Torr, and Grégory Rogez. Placing objects in context via inpainting for out-of-distribution segmentation. In *European Conference on Computer Vision*, pages 456–473. Springer, 2024. 3

[5] Gianni Franchi, Xuanlong Yu, Andrei Bursuc, Angel Tena, Rémi Kazmierczak, Séverine Dubuisson, Emanuel Aldea, and David Filliat. Muad: Multiple uncertainties for autonomous driving, a benchmark for multiple uncertainty types and tasks. *arXiv preprint arXiv:2203.01437*, 2022. 3, 5

[6] Edward J Hu et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*, 2022. 1

[7] Yuyuan Liu, Choubo Ding, Yu Tian, Guansong Pang, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1151–1161, 2023. 1, 3

[8] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 5

[9] Maxime Oquab et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5

[10] 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. 5

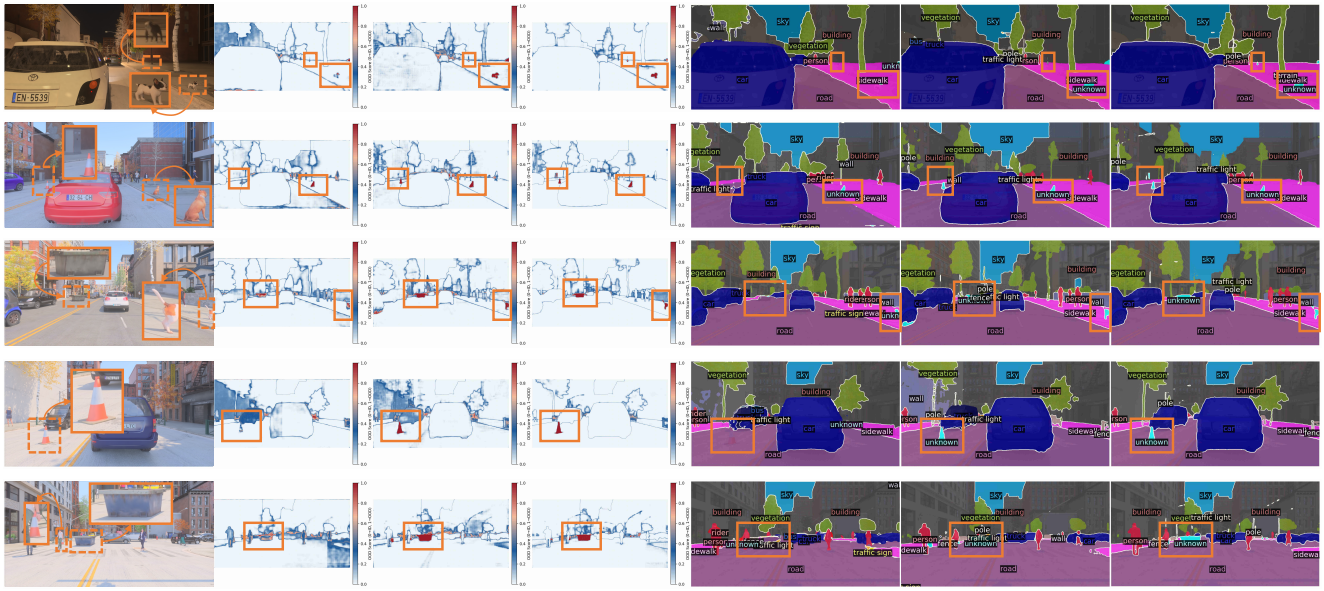
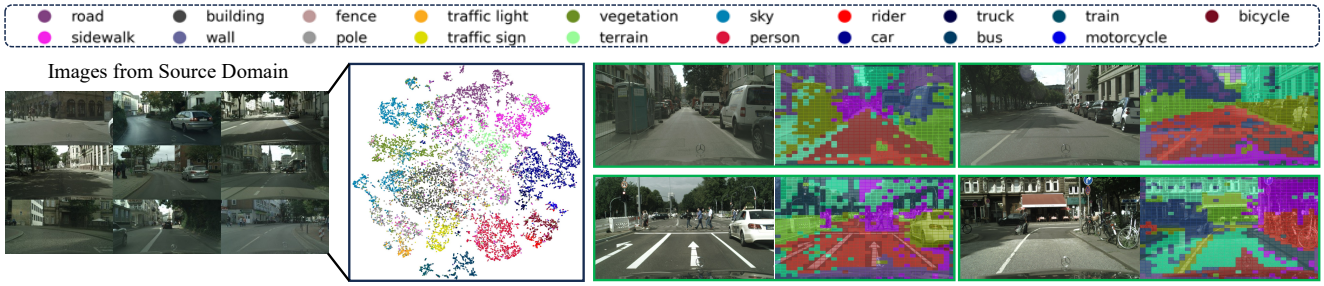


Figure 4. Qualitative comparison on the MUAD. The model is trained on Cityscapes with DINOv2-B backbone.



(a). Pre-train Token Distribution from DINOv2-L (b). Pre-train Token Visualization from DINOv2-L

Figure 5. Visualization of the manifold structure of token embeddings constructed by the pre-trained VFM (DINOv2-L). The T-SNE in (a) shows that the tokens corresponding to the categories of the source domain are regularly projected onto the space, forming distinct cluster structures. (b) shows the pre-trained token similarity across different spatial locations in various images, where the colors of similar tokens are closer. It is obtained by applying PCA dimensionality reduction to the pre-trained tokens and re-mapping them to the RGB space.

[11] 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 models for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 28619–28630, 2024. 5

[12] 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. 5

[13] Seokju Yun, Seunghye Chae, Dongheon Lee, and Youngmin Ro. Soma: Singular value decomposed minor components adaptation for domain generalizable representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25602–25612, 2025. 5

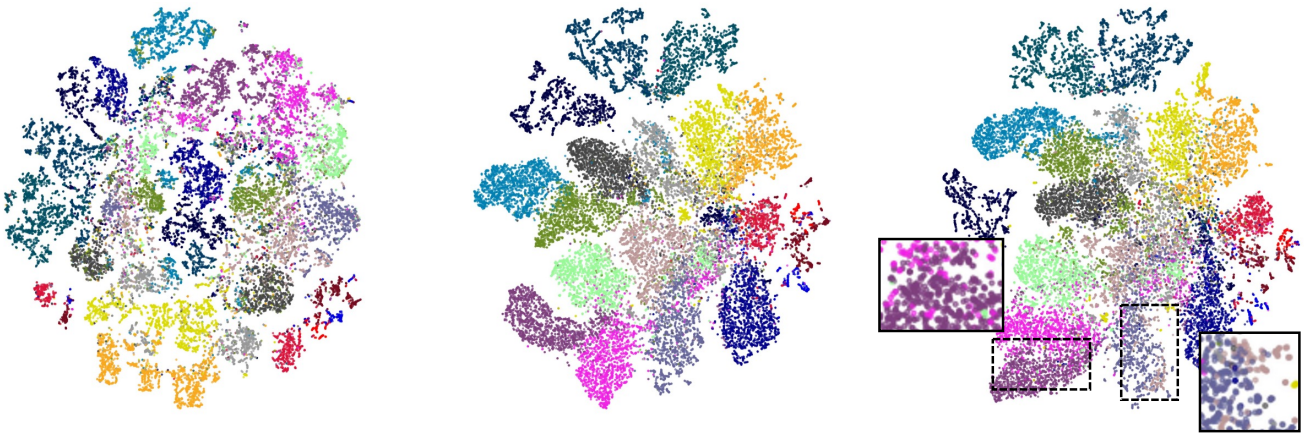


Figure 6. t-SNE visualization of pixel-level token embeddings on the ACDC dataset for the pre-trained DINOv2-L backbone (left), GeCo (middle), and the RP+MSE (right). Each color denotes a semantic class.