6. Appendix

6.1. Implementation Details

We introduce the one-stage detector YOLO-World [8] into the unsupervised domain adaptation (UDA) setting as the baseline detector. In each iteration, one batch of source images with ground truth and one batch of target domain images with pseudo labels are forwarded to calculate the classification, adversarial and regression loss. The hyper-parameter r, λ^I, λ^O is set to $2.0, 1.0, 0.5$. The batch size for each domain is set to 2, using the SGD optimizer with linear warm-up and decay learning rate. The base learning rate is 1e-3, the warm-up period is 5 epochs and the decay starts from the 20th epoch. Mean Average Precision (mAP) with a threshold of 0.5 is taken as the evaluation metric. All experiments are deployed on 1 Tesla V100 GPUs.

6.2. Additional Benchmarks

Table 10. Comparison (%) on benchmarks with various categories.

Benchmark	Category	Baseline (UDA)	DA-Mamba	Gains
C→F	8	52.3	58.1	5.8
C→B	7	41.9	48.7	6.8
P→Clp	20	46.2	52.5	6.3
P→Cmc	6	37.9	43.8	5.9
K→C	1	60.4	62.5	3.1
S→C	1	59.3	62.3	3.0

Evaluation on Multi-Category Benchmarks In the main text, we report experiments on four mainstream multi-category adaptation scenarios. As shown in Table 10 (rows $2 \sim 5$), DA-Mamba consistently achieves substantial improvements of $5.8 \sim 6.8\%$, demonstrating strong cross-domain generalization under diverse domain shifts.

Evaluation on Single-Category Benchmarks Beyond these benchmarks, KITTI [20] and Sim10K [33] are also two important mainstream benchmarks. KITTI includes 7,481 real-world images with different camera Field of View (FoV) settings. The synthetic dataset SIM10k has 10,000 photos from the GTA V video game, designed to evaluate synthetic-to-real transfer. However, both adaptation scenarios KITTI→Cityscapes (K→C) and SIM10k→Cityscapes (S→C) contain a single category "car". These adaptation settings do not include inter-class semantic structures, and thus cannot fully evaluate DA-Mamba's ability to model global semantic relationships across categories. Consequently, compared with the multi-category scenarios (6–20 classes), DA-Mamba obtains smaller but still consistent improvements of $2.1 \sim 3.0\%$ on these single-category tasks. Importantly, since no cross-category semantic differences exist in these cases, the performance gain mainly comes from enhanced modeling of spatial dependencies under domain shift. This result further validates that DA-Mamba remains effective even without cross-category semantics and can robustly model long-range spatial dependencies in single-category scenarios.

Overall, these results show that DA-Mamba provides consistent gains across both semantic-rich and -scarce adaptation settings, highlighting its robustness and broad applicability under diverse domain shifts.

6.3. Structure Variants

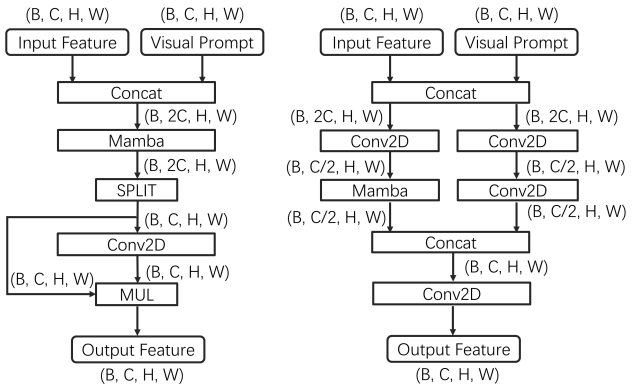
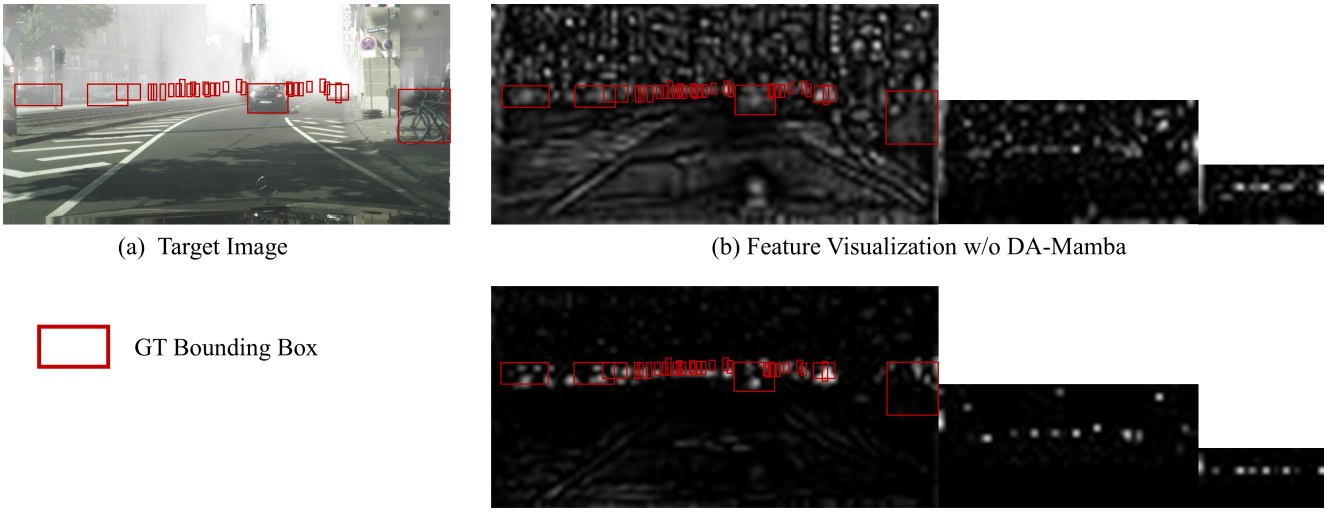


Figure 4. Different structure for IA-SSM and OA-SSM.

Table 11. Comparison (%) of different SSM structure.

Structure	Param(M)	FPS	mAP
Serial	8.610	12.6	58.1
Parallel	1.732	14.1	58.1

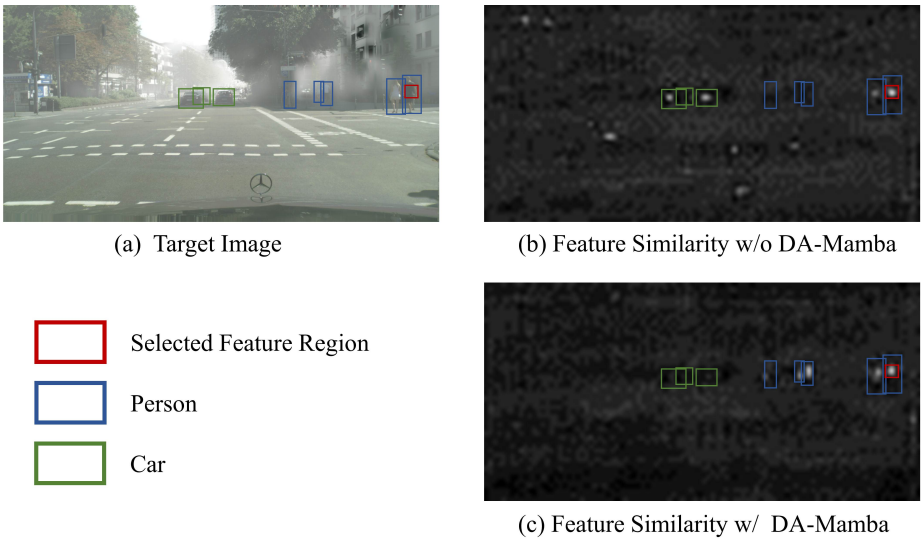
Serial or Parallel? *Parallel network achieves comparable accuracy to the serial design with higher computational efficiency.* Depending on the ordering of global and local alignment, we explore two network variants: (a) serial network and (b) parallel network. As shown in Fig. 4(a), the serial network first feeds the concatenated input-prompt features into the Mamba layer to learn global visual dependencies, and subsequently applies gated convolution to extract local domain-invariant features. In contrast, the parallel network in Fig 4(b) exploits a dual-pipeline design, where local and global information are learned independently through convolution pipeline and Mamba pipeline, and then fused to form the output feature. As reported in Table 11, the parallel network requires only 20%(1.732M) of the parameters of the serial network(8.610M) while achieving the same mAP. It also reaches a higher inference speed (14.1 FPS vs. 12.6 FPS), demonstrating better computational efficiency. This efficiency gain comes from avoiding the processing of expanded feature channels (2C) through both Mamba and



(a) Target Image (b) Feature Visualization w/o DA-Mamba

(c) Feature Visualization w/ DA-Mamba

Figure 5. Visualizations of the extracted feature map.



(a) Target Image (b) Feature Similarity w/o DA-Mamba

(c) Feature Similarity w/ DA-Mamba

Figure 6. Visualizations of the similarity between the features of a certain point and other regions on the feature map.

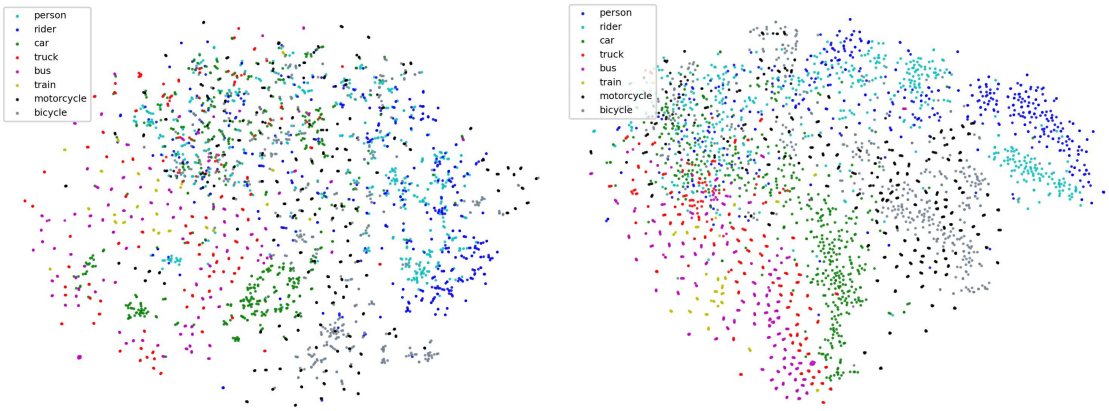
convolution branches and instead keeping them lightweight (C/2) before fusion.

Given its superior efficiency and its natural alignment with the dual-pipeline modeling objective of DA-Mamba (*i.e.*, global-local alignment), we adopt the parallel network structure in the design of IA-SSM and OA-SSM.

Table 12. Comparison (%) on prototypes usage in OA-SSM.

Method	C→F	C→B	P→Clp	P→Cmc
Learnable prototypes	57.5	47.3	51.7	42.6
CLIP embedding	58.1	48.7	52.5	43.8

CLIP embedding or Learnable prototypes? *CLIP embedding provides more stable semantic priors for OA-SSM.* In Table 12, we compare the performance of using learnable prototypes and CLIP embeddings as category prototypes in OA-SSM. The results show that using CLIP embedding yields consistently and moderately better performance than the learned per-class prototypes. This indicates that using CLIP embedding as the category prototypes can effectively anchor instance-level object semantics, providing efficient semantic priors and facilitating subsequent modeling of region-level semantics.



(a) T-SNE w/o DA-Mamba (b) T-SNE w/ DA-Mamba

Figure 7. T-SNE visualization of the features extracted from the target domain.

6.4. Feature Visualization

Feature Visualization *DA-Mamba accurately extracts the features of foreground objects in the target domain while suppressing redundant, semantically-irrelevant background information.* We visualize the multi-scale features before and after inserting the IA-SSM and OA-SSM modules. As shown in Fig. 5 (b), without DA-Mamba, the extracted features respond strongly to background regions while failing to highlight foreground objects, indicating insufficient domain alignment. In contrast, Fig. 5 (c) shows that DA-Mamba suppresses irrelevant textures (such as lane markings and background building) and emphasizes true foreground areas, suggesting more compact and domain-invariant representations. This improved localization and background suppression reduce feature noise transferred across domains, thereby enhancing cross-domain generalization.

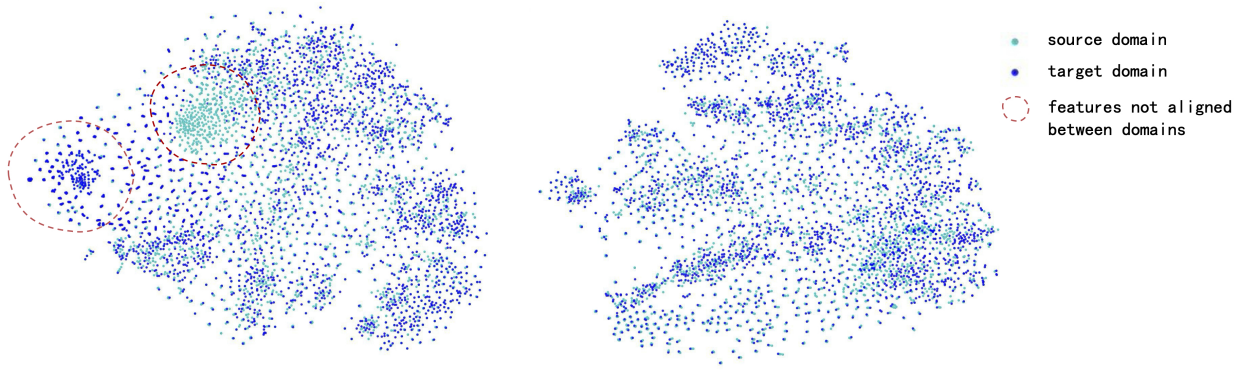
Feature Similarity *DA-Mamba exhibits high intra-class similarity and low inter-class similarity across spatially distant regions, reflecting its capability in capturing global domain-invariant features and long-range semantic relationships.* To intuitively verify DA-Mamba's ability to capture global domain-invariant features and capture long-distance semantic dependencies, we select a "Person" object from the target image (red boxes), and compute its similarity with all other regions in the feature map. Fig. 6(b) and (c) show the visualization results of baseline and DA-Mamba, respectively, where brighter regions indicate higher similarity and darker regions indicate lower similarity. As shown in Fig. 6(b), the baseline exhibits low similarity even within the same class (blue boxes) and undesired high similarity with other categories (green boxes), indicating weak global semantic discrimination. Conversely, Fig. 6(c) shows that DA-Mamba consistently highlights same-class regions (blue boxes) while suppressing cross-category correlations (green boxes), regardless of spatial distance. This demonstrates that the global-local alignment in DA-Mamba effectively propagates semantic dependencies over long ranges, showing a good ability to extract global domain-invariant representations that are robust to domain shifts.

Together, these qualitative results confirm that DA-Mamba achieves more stable and consistent domain alignment by simultaneously enhancing local domain-invariant features and global semantic-spatial dependencies, thereby providing robust cross-domain representations.

6.5. T-SNE Visualization

T-SNE on the Target Domain. *The features extracted by DA-Mamba exhibit tighter intra-class clusters and larger inter-class margins, reflecting better semantic structuring in the target domain.* To intuitively demonstrate the effect of DA-Mamba, we provide a T-SNE visualization of features extracted from the target domain. As shown in Fig. 7(a), before introducing DA-Mamba, features from different categories are heavily mixed and instances of the same category do not form compact clusters. In contrast, Fig. 7(b) shows that DA-Mamba produces clear boundaries between categories. At the same time, categories with strong semantic relevance lie closer in the embedding space, such as person–rider, car–truck, and bicycle–motorcycle. Moreover, categories that are spatially related in the scene, such as rider, bicycle, and motorcycle, are also grouped nearby. This indicates that benefiting from global–local alignment, DA-Mamba effectively learns both semantic and spatial dependencies, leading to more structured and discriminative representations in the target domain.

An interesting observation is that the rider category is split into two sub-clusters, which we attribute to the semantic difference between riders on bicycles and riders on mo-



(a) T-SNE w/o DA-Mamba (b) T-SNE w/ DA-Mamba

Figure 8. T-SNE visualization of the features extracted from both domains.

torcycles.

Finally, we note that both (a) and (b) contain a region where different categories are mixed, mainly due to heavy occlusions and visually inseparable objects in those cases.

T-SNE across Source and Target Domains. *DA-Mamba achieves better feature alignment between the source and target domains.* We further visualize the T-SNE of features from both domains in Fig. 8. As highlighted by the red dashed circles in Fig. 8(a), features from the source domain (cyan) are not well aligned with those from the target domain (blue), indicating that the baseline model fails to achieve sufficient cross-domain alignment. In contrast, with the proposed DA-Mamba, features from both domains in Fig. 8(b) are tightly mixed and share similar distributions. This enhanced alignment results from DA-Mamba's global–local alignment mechanism, where global SSM pipeline establishes domain-invariant long-range structure while local convolutional pipeline anchors the alignment to stable low-level patterns. Such complementary alignment enables the features of both domains to be aligned into a shared, well-organized distribution.

Overall, these T-SNE visualizations demonstrate that DA-Mamba fundamentally improves cross-domain feature alignment, that is, better semantic structure in the target domain and better alignment across source and target distributions, confirming that its performance gains primarily arise from fine-grained global–local domain-invariant representation learning rather than merely improving the detector capacity.

6.6. Detection Results

We visualize the detection results of baseline and DA-Mamba in Fig. 9 While baseline misclassifies the rider as person, our DA-Mamba correctly classifies the rider. Meanwhile, DA-Mamba shows higher prediction confidence, and detects more objects in the fog that are missed in the baseline. This indicates that DA-Mamba effectively learns local and global domain properties, achieving fine-grained domain alignment.

6.7. Pseudo Label

Table 13. Analysis (%) on Pseudo Labels.

Benchmark	Pseudo Labels	C→F	C→B	P→Clp	P→Cmc
Baseline(UDA)		50.6	40.0	43.8	35.9
Baseline(UDA)	✓	52.3	41.9	46.2	37.9
DA-Mamba		57.1	48.0	51.0	42.5
DA-Mamba	✓	58.1	48.7	52.5	43.8

Effect of Pseudo Labels. *DA-Mamba does not rely on pseudo labels for performance gains.* Pseudo labels are selected by classification max softmax logits ≥ 0.8. Table 13 reports the results of applying pseudo label in the target domain (*i.e.*, the loss term $\mathcal{L}_{cls}^{T}$) during training. For the baseline UDA detector, introducing pseudo labels yields noticeable improvements ($1.7 \sim 2.4\%$), reflecting its strong dependence on target domain supervision to compensate for insufficient domain alignment. In contrast, DA-Mamba already achieves significant improvements without pseudo labels, and adding pseudo labels provides only marginal additional gains ($0.7 \sim 1.5\%$). This smaller improvement margin indicates that DA-Mamba substantially reduces the domain discrepancy itself, leaving less room for pseudo labels to correct target domain errors. Therefore, the performance gains of DA-Mamba do not stem from pseudo labels but from its stronger global–local domain alignment capability.

6.8. Failure Case

We observe that the missed detections of DA-Mamba mainly occur in regions with strong ambiguity or heavy occlusion, as shown in the middle area of Fig. 9. Since such

(a) Ground Truth

(b) Baseline

(c) DA-Mamba

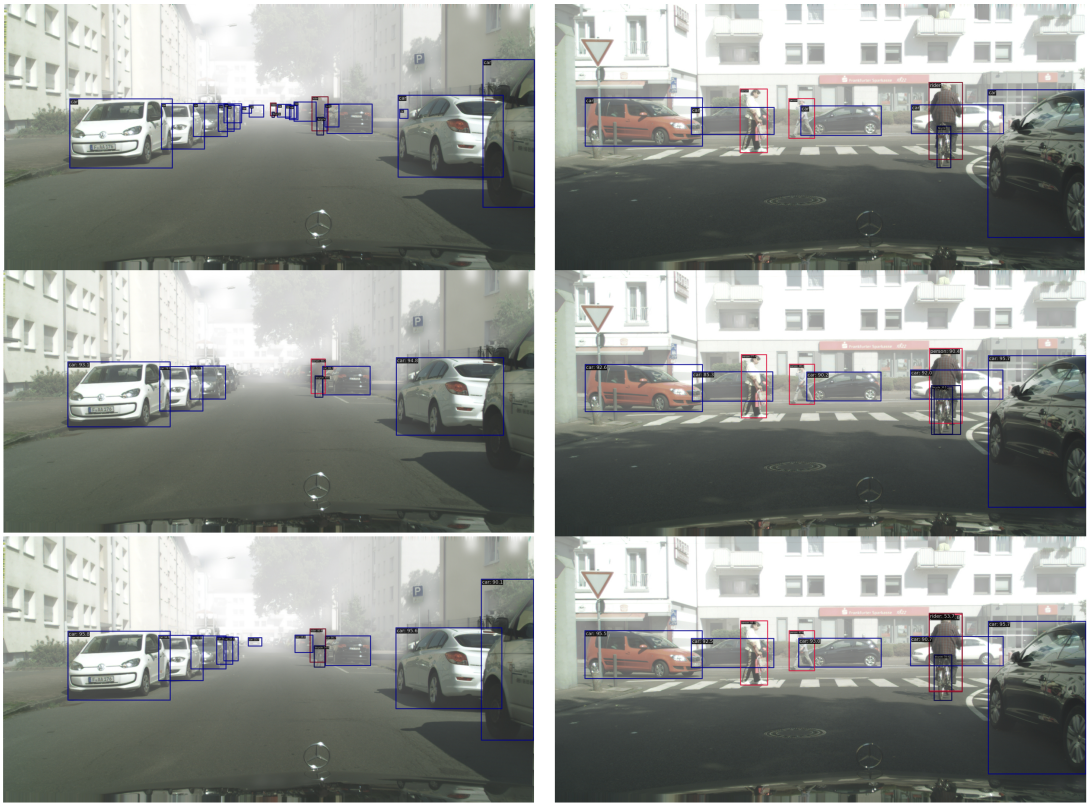


Figure 9. Detection results visualization.

cases suffer from severely degraded visual semantics, similar failures are commonly observed in most object detection frameworks and are not specific to our method.

6.9. Error Bars

Table 14. mAP(%) on four benchmarks.

C→F	C→B	P→Clp	P→Cmc
58.1(± 0.2)	48.7(± 0.4)	52.5(± 0.3)	43.8(± 0.4)

We provide error bars in Table 14. The error bars are captured by multiple running with given experimental conditions.

6.10. Extended Discussion on Global-Local Alignment Design

Why DA-Mamba Outperforms Transformer-Based DAOD methods and Convolutional Attention Variants. Table 6 in the main text shows that DA-Mamba not only surpasses the SOTA transformer-based DAOD method DATR [5] with significantly lower computational overhead, but also outperforms convolutional attention variants under similar cost. This advantage stems from DA-Mamba's novel synergistic integration of convolutional locality and SSM-based global perception, enabling efficient and fine-grained global–local alignment.

Transformer-based DAOD methods introduce self-attention to capture global dependencies across domains. However, their quadratic complexity forces the use of large patch sizes to reduce token counts. Since the same patch is treated as a whole in attention computation, larger patches make the granularity of attention computation coarser. Moreover, Transformer attention is domain-agnostic, treating all positions uniformly and failing to emphasize locally stable domain-invariant features, which are crucial for local alignment. In contrast, DA-Mamba employs a dedicated convolutional pipeline that preserves fine-grained spatial detail and efficiently extracts local domain-invariant features.

Convolutional attention variants attempt to approximate global interactions with lower cost but rely on globally shared parameters. For example, GCNet [2] discards Q(query) and only use K(key) and V(value) to simplify attention calculation. However, eliminating the query branch prevents pixel-level semantic analysis and makes the model sensitive to domain bias. And HAM [37] decomposes attention into spatial attention and channel attention based on convolution and pooling, which is restricted by the sliding window receptive fields and inherently cannot capture se-

mantic dependencies between distant objects. Such shared global descriptors suppress domain-specific variations and are easily biased by style differences between source and target domains, causing insufficient global alignment. In contrast, the SSM pipeline in DA-Mamba performs a directed scan across the entire feature map, allowing each position to aggregate information from all previous positions. This gives DA-Mamba the ability to model long-range semantic and spatial dependencies across domains.

Overall, DA-Mamba adopts a global–local hybrid design: (1) the convolution pipeline efficiently extracts stable local domain-invariant features through local connectivity and translation invariance, and (2) the SSM pipeline provides linear-time global modeling of long-range dependencies. This decomposition allows DA-Mamba to decouple the domain adaptation process into local alignment and global alignment, avoiding both the domain-agnostic behavior of Transformers and the rigidity of convolutional attention, leading to more effective cross-domain representation learning.

Why Enlarging the Receptive Field Promotes Domain Alignment. Domain adaptive object detection typically optimizes an adversarial domain classifier using an MSE loss:

$$\mathcal{L}_{\text{adv}} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(\mathcal{D}(\mathbf{f}_{ij}) - y\right)^2, \tag{17}$$

where $\mathbf{f}_{ij}$ denotes the feature at location (i, j) on the feature map $\mathbf{f}$, the domain classifer $D(\cdot)$ predicts the probability that a feature comes from the source domain, and $y \in \{0, 1\}$ is the domain label. This objective enforces the feature distributions of the two domains to be indistinguishable at *every* pixel position.

When the receptive field is small, $\mathbf{f}_{ij}$ contains only local information, *i.e*. the content around (i, j) on feature map $\mathbf{f}$. Formally, if $\mathbf{f}_{ij}^d$ depends only on a local region $\mathbf{N}_{ij}^d$, we have

$$\mathbf{f}_{ij}^d = \mathcal{F}(\mathbf{N}_{ij}^d), \tag{18}$$

where $d \in \{S, T\}$ denotes the source or target domain, and $\mathcal{F}$ denotes the visual encoder. In this case, the domain classifier processes each spatial location independently, without explicitly modeling relationships between different regions. Therefore, the adversarial loss in Eq. (17) encourages

$$p\big(\mathcal{F}(\mathbf{N}_{i_s j_s}^S)\big) = p\big(\mathcal{F}(\mathbf{N}_{i_t j_t}^T)\big), \tag{19}$$

where $p(\cdot)$ denotes the feature distribution. However, Eq. (19) only enforces domain invariance of *local* features within each region, and does not constrain the global relationships among regions. Therefore, such alignment remains local and is insufficient to capture global domain shifts (*e.g*., object co-occurrence, layout, or scene style).

When the receptive field covers the entire image, each $\mathbf{f}_{ij}^d$ additionally aggregates information from a broader semantic context:

$$\mathbf{f}_{ij}^d = \left[\mathcal{F}(\mathbf{N}_{ij}^d);\, \mathbf{c}_{ij}^d\right], \tag{20}$$

where

$$\mathbf{c}_{ij}^d = \mathcal{S}\Big(\{\mathbf{N}_{kl}^d\}_{k=1,l=1}^{H,W} \setminus \{\mathbf{N}_{ij}^d\}\Big) \tag{21}$$

contains features from all other regions via a global modeling network $\mathcal{S}$. In this setting, each $\mathbf{f}_{ij}^d$ simultaneously encodes its own local feature and its relationships to all other regions, including object structure, inter-object relations, and scene layout. After optimizing Eq. (17), we then expect

$$\begin{aligned} p\big(\mathcal{F}(\mathbf{N}_{i_s j_s}^S)\big) &= p\big(\mathcal{F}(\mathbf{N}_{i_t j_t}^T)\big), \\ p\big(\mathbf{c}_{i_s j_s}^S\big) &= p\big(\mathbf{c}_{i_t j_t}^T\big). \end{aligned} \tag{22}$$

This means that, for each local region, both its own local features and its global relationships with other regions are required to be domain-invariant, thereby achieving *simultaneous local and global alignment*.

In summary, enlarging the receptive field allows the adversarial loss to act on both local semantics and long-range dependencies at each spatial location, effectively promoting alignment between the two domains. In DA-Mamba, this principle is instantiated by the dual-pipeline design: the convolution pipeline corresponds to $\mathcal{F}(\mathbf{N}_{ij}^d)$ and focuses on local domain-invariant information, while the SSM pipeline corresponds to $\mathbf{c}_{ij}^d$ and models global semantic and spatial dependencies. Their fusion enables DA-Mamba to implement global–local domain alignment in a theoretically grounded manner, which explains its superior cross-domain performance in our experiments.

Future Work. While DA-Mamba demonstrates that global–local alignment can significantly improve cross-domain generalization, several directions remain promising for future exploration. First, instead of inserting SSM blocks into FPN as in the current design, enabling SSM with native multi-scale global modeling capability may further strengthen global semantic propagation across resolutions. Second, integrating uncertainty estimation schemes could allow the model to dynamically balance global and local alignment according to the degree of domain shift. Finally, applying the proposed global–local alignment principle to other adaptation tasks such as instance segmentation, open-vocabulary detection, or video-domain adaptation may broaden the applicability of DA-Mamba. We leave these directions for future investigation.