ASTAD: Asymmetric Style Transfer for Synthetic-to-Real Adaptation in Autonomous Driving

Dingyi Yao, Xinqi Zhang, Lihui Peng*, Jianming Hu, Danya Yao, and Yi Zhang

Department of Automation, Tsinghua University, Beijing 100084, China
{ydy24,zhang-xq24}@mails.tsinghua.edu.cn; {lihuipeng, hujm, yaody, zhyi}@tsinghua.edu.cn

Abstract. Synthetic data mitigates the data scarcity problem in autonomous driving perception. However, the synthetic-to-real gap leads to performance degradation, hindering real-world model generalization. Although current methods leverage diffusion models for photorealistic style transfer to bridge this gap, they critically ignore a practical asymmetry: while synthetic data possesses perfect pixel-level annotations, real-world style reference images generally lack corresponding labels. Consequently, existing methods relying on symmetric semantic guidance suffer from either prohibitive annotation costs or severe semantic misalignment. To address this dilemma, we formally propose a novel task: Asymmetric Style Transfer for Autonomous Driving (ASTAD), which requires semantically consistent transfer using only labeled synthetic content and unlabeled real-world references. We further introduce the ASTModel, a training-free two-stage framework designed to bridge this domain gap under asymmetric constraints. ASTModel first extracts a coarse semantic prior from the unlabeled target, followed by dynamic prior refinement and class-consistent style injection during the denoising process. Extensive experiments demonstrate that ASTModel significantly outperforms existing methods in downstream perception utility and structural fidelity, while offering a 3.2× inference speedup. This work aligns synthetic-to-real adaptation with practical constraints, holding the potential to accelerate the scalable deployment of robust autonomous driving systems. Code: https://github.com/Dingyi-Yao/ASTAD.

Keywords: Image synthesis · Style transfer · Synthetic-to-real adaptation · Autonomous driving

1 Introduction

Deep learning-based perception systems constitute the backbone of modern autonomous driving, enabling vehicles to interpret complex environments with high

* Corresponding author.

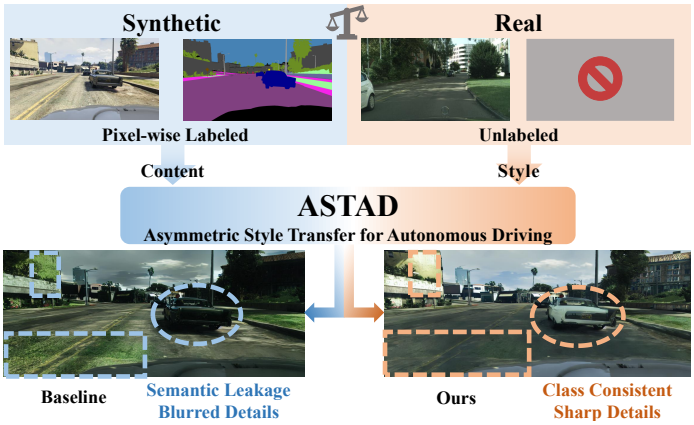

Fig. 1: Illustration of ASTAD and the motivation of ASTModel. **Top:** ASTAD targets asymmetric synthetic-to-real style transfer, where synthetic content images have pixel-wise labels while real style images are unlabeled. **Bottom:** Without style-side semantic guidance, existing baselines suffer from semantic leakage and blurred scene details, whereas ASTModel enables class-consistent style injection, producing target-style images with preserved driving-scene geometry.

precision. However, the success of these systems is predicated on the availability of large-scale, high-density, and fine-grained semantic annotations. Acquiring high-quality annotations for real-world datasets is prohibitively time-consuming. For example, the annotation of a single Cityscapes [9] image requires roughly 1.5 hours [17]. Consequently, synthetic data generated by simulators has emerged as a complementary solution. It enables the rapid and cost-effective generation of large-scale datasets while providing error-free annotations [31]. For instance, the GTA dataset [27] leverages the interception of communication between the game engine and graphics hardware to automatically generate precise pixel-level semantic labels.

Despite these advantages, the practical deployment of synthetic data is bottlenecked by the inherent distribution shift between synthetic and real-world domains. This synthetic-to-real gap primarily stems from the limitations of simulation engines, including approximated physical rendering, idealized material textures [39]. Consequently, models trained purely on synthetic distributions suffer from severe performance degradation when deployed in the real world [32]. Bridging this critical gap is therefore essential to translate synthetic scalability into safe, real-world reliability.

To mitigate this, synthetic-to-real adaptation has become a crucial research direction. A promising direction is generating "target-style synthetic data", images that retain the precise layout of the simulation but exhibit the visual characteristics of the real world. In this context, Diffusion Models have emerged as a powerful paradigm [28,30]. Through an iterative denoising process, they demonstrate superior capabilities in accurately modeling complex data distributions

and generating high-fidelity details [4,37,38]. With their zero-shot generalization capabilities derived from large-scale pre-training, diffusion models offer a powerful mechanism to inject real-world styles into synthetic content [1, 5, 8, 10, 15].

However, a critical contradiction remains overlooked in existing diffusion-based stylization methods: **the inherent information asymmetry between content and style**. Existing methods [5] typically require explicit semantic segmentation maps for both content and style images to guide the transfer process and prevent semantic misalignment (*e.g.*, placing vegetation textures on roads). Yet a practical dilemma arises: while pixel-perfect segmentation labels are effortlessly available for synthetic content images, obtaining corresponding dense annotations for real-world style reference images is difficult.

We formally term this fundamental imbalance **"Asymmetry"**: synthetic content images are paired with pixel-level semantic masks, whereas real-world style images are available only as raw images without dense semantic labels. Because of this, existing methods often fail in practical autonomous driving scenarios, either by relying on requiring expensive manual annotation for style images or by suffering from severe cross-class pollution when such labels are absent, as shown in Fig. 1.

To address this dilemma, we formally propose a novel task: **Asymmetric Style Transfer for Autonomous Driving (ASTAD)**. ASTAD aligns synthetic-to-real adaptation with practical constraints by requiring semantically consistent transfer using only labeled synthetic content and a raw, unlabeled real-world style image. To tackle this, we introduce **ASTModel**, a training-free, two-stage framework. Our main contributions are summarized as follows:

- We formalize the ASTAD task, identifying the critical label asymmetry in synthetic-to-real adaptation and establishing a realistic problem setting where target style labels are unavailable.
- We introduce the ASTModel to resolve the asymmetry constraints. It incorporates Prototype-Guided Style Semantic Prior Extraction to address the semantic cold-start problem via foundation models, and integrates Multi-Layer Semantic Voting, Semantically Constrained Adaptive Attention Filtering and Pixel-Proportion Modulated Hybrid AdaIN, to robustly enforce semantic consistency and correct mismatches during the denoising process.
- Extensive experiments demonstrate that ASTModel generates images with enhanced semantic consistency, with its inference speed substantially improved.

2 Related Work

2.1 General Neural Style Transfer

Early Neural Style Transfer (NST) optimized feature Gram matrices [14], later accelerated by feed-forward networks utilizing Adaptive Instance Normalization (AdaIN) [18]. To address AdaIN's loss of local details and structural "content

leak", subsequent works introduced point-wise attention [23], Transformer architectures [11, 33], and reversible flows [2]. Concurrently, GANs formulated stylization as image-to-image translation across paired [20], unpaired [25, 44], multimodal [19], and multi-domain [6, 7] settings, or via explicit structure-texture latent swapping [26], though often bounded by training instability [13]. Recently, Diffusion Models [28] dominated generative tasks [4, 37, 38]. Strategies adapting diffusion models for style transfer can generally be categorized into training-based and training-free methods. Training methods [3, 12, 35, 41] extract style concepts from reference images into conditional embeddings to guide diffusion-based style transfer. Conversely, training-free methods modify the denoising process directly. Early approaches guided reverse diffusion using disentangled ViT representations [22]. Later methods swapped self-attention keys and values to inject style [1, 8, 10], while advanced approaches further enforce semantic consistency by rearranging features based on dense correspondences [15].

Nevertheless, significant differences still exist between these generic artistic stylization methods and the specific demands of autonomous driving. First, artistic transfer typically focuses on object-centric images with a clear foreground-background separation. In contrast, driving scenes utilize high semantic density, characterized by multi-instance occurrences and diverse categories within a single view. Second, the primary metric for artistic transfer is perceptual aesthetics, often tolerating or even encouraging geometric deformations. Conversely, autonomous driving mandates the rigorous preservation of structural and semantic boundaries, as any distortion directly compromises the utility of the data for downstream perception tasks. Third, the autonomous driving field possesses a unique advantage: the inherent availability of segmentation maps, particularly for synthetic data. This facilitates stylization guided by explicit semantic information, a critical capability that general artistic methods typically overlook. These limitations necessitate specialized adaptation strategies tailored for autonomous driving perception tasks.

2.2 Style-Based Synthetic-to-Real Adaptation

To bridge the gap between general stylization and the rigorous demands of autonomous driving, researchers have developed specialized style-based adaptation methods. The primary goal is to mitigate the distribution shift between synthetic simulation and the real world by injecting realistic textures into annotated synthetic data. Pioneering works like CyCADA [16] and SHADE [42] advanced synthetic-to-real adaptation by enforcing semantic consistency and enhancing style diversity. However, these methods often struggle with training instability or limited generative fidelity compared to modern standards. Diffusion Models have significantly elevated generation quality. Methods such as DGInStyle [21], SimGen [43], and Weather-Diff [34] achieve photorealism by integrating structural guidance (*e.g.*, semantic masks, simulator layouts) and adapting to specific domains through fine-tuning. However, these training-based approaches are computationally intensive and inflexible, requiring a dedicated model training for each specific target domain. To bypass training costs, recent research employs

training-free attention injection. Notably, CACTIF [5] mitigates semantic inconsistencies by utilizing class-aware mechanisms, employing explicit segmentation masks to strictly confine style transfer to corresponding semantic regions.

However, these methods rely on a "Symmetric Information" assumption, requiring dense annotations for both source content and target style. This is impractical for autonomous driving, where real-world style labels are often unavailable. In this asymmetric setting, symmetric methods fail to distinguish semantic regions, leading to severe cross-class semantic leakage. Addressing this unresolved dilemma motivates our proposed work.

3 Methodology

3.1 Task Formulation

Let $\mathcal{D}_{src} = \{(I_c^i, M_c^i)\}_{i=1}^N$ denote the synthetic source domain, containing N content images I_c with pixel-wise semantic masks M_c. Conversely, the real-world target domain is represented by a style reference image I_s.

Existing diffusion-based style transfer methods, such as CACTIF [5], typically operate under a Symmetric Information Assumption. Formally, the symmetric transfer function Φ_{sym} is defined as $I_{out} = \Phi_{sym}(I_c, M_c, I_s, M_s)$.

To align with practical deployment constraints, we introduce the Asymmetric Style Transfer for Autonomous Driving (ASTAD) task. In this paper, "asymmetric" specifically refers to the annotation-level information imbalance between the labeled synthetic content image and the unlabeled real-world style image. That is, the synthetic source provides (I_c, M_c), whereas the style image is only available as a raw image I_s, implying $M_s = \emptyset$.

Under this annotation-asymmetric setting, the objective of ASTAD is to generate a stylized image I_{out} that inherits the visual characteristics of the real-world style image I_s while preserving the semantic layout specified by the synthetic mask M_c. Formally, the asymmetric transfer function Φ_{asym} is defined as $I_{out} = \Phi_{asym}(I_c, M_c, I_s, \emptyset)$.

3.2 Framework Overview

As illustrated in Fig. 2, we propose ASTModel, a training-free framework for synthetic-to-real adaptation. Our pipeline operates in two stages: First, Stage I (Sec. 3.3) extracts a coarse prior from the unlabeled target via foundation model prototype matching. Next, Stage II (Sec. 3.4–Sec. 3.6) refines this prior during the diffusion reverse process using a Multi-Layer Semantic Voting (Sec. 3.4). This dynamically sharpened guidance then strictly governs two novel injection modules: Semantically Constrained Adaptive Attention Filtering (Sec. 3.5) to prevent semantic leakage, and Pixel-Proportion Modulated Hybrid AdaIN (Sec. 3.6) to robustly align class-aware feature statistics.

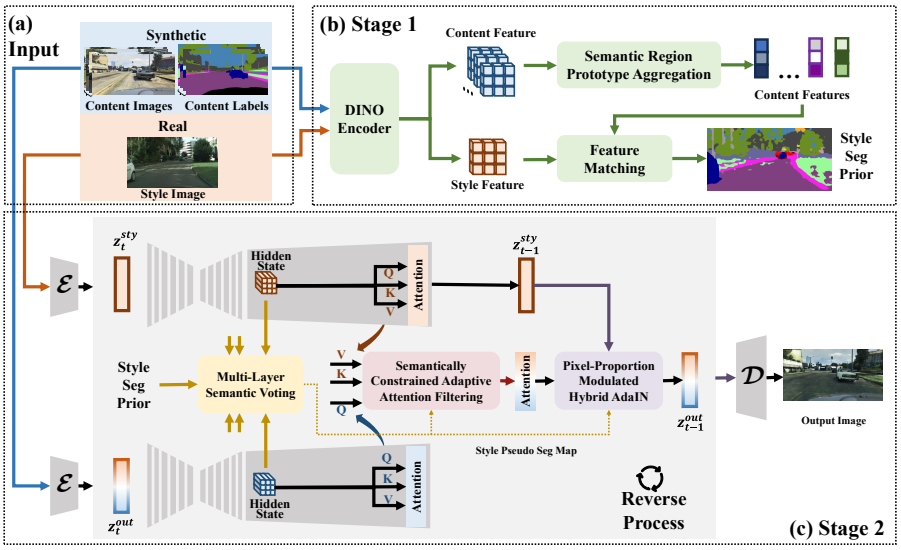

Fig. 2: Overview of the proposed ASTModel. The pipeline operates as a training-free, two-stage framework to address asymmetric constraints. **(a) Inputs:** Abundant labeled synthetic data and a single unlabeled real-world style image. **(b) Stage I (Implicit Semantic Discovery):** Extracts a coarse Style Segmentation Prior by leveraging the semantic correspondence between synthetic prototypes and style features in the DINO space. **(c) Stage II (Asymmetric Style Injection):** Refines this prior via the Multi-Layer Semantic Voting during the diffusion reverse process, and performs style injection using Semantically Constrained Adaptive Attention Filtering and Pixel-Proportion Modulated Hybrid AdaIN.

3.3 Prototype-Guided Style Semantic Prior Extraction

To extract a coarse semantic prior for the unlabeled style image, we leverage the robust feature space of a frozen pre-trained Vision Transformer, DINO [24], as our feature extractor $\mathcal{F}$. Given a synthetic content image I_c and a real-world style image I_s, we extract their L_2-normalized patch-level feature representations $\mathbf{F}_c = \mathcal{F}(I_c)$ and $\mathbf{F}_s = \mathcal{F}(I_s)$.

We leverage the available synthetic annotations M_c to construct robust semantic prototypes. For each semantic category $k \in \{0, \ldots, C-1\}$, we define the semantic region $\mathcal{R}_k = \{(h, w) \mid M_c(h, w) = k\}$. To compute a class-specific prototype $\mathbf{p}_k$, we aggregate the feature vectors within each region

$$\mathbf{p}_k = \text{Normalize}\left(\frac{1}{|\mathcal{R}_k|} \sum_{(h,w)\in\mathcal{R}_k} \mathbf{F}_c(h, w)\right) \tag{1}$$

Treating these synthetic prototypes as semantic anchors, we formulate style parsing as a dense nearest-neighbor matching in the shared DINO space. The

coarse semantic prior M_{prior} is obtained by assigning each spatial location to the category with the highest cosine similarity

$$M_{prior}(h, w) = \arg\max_k \langle \mathbf{F}_s(h, w), \mathbf{p}_k \rangle \tag{2}$$

While effective for global structural layout, M_{prior} is inherently bounded by the patch-level resolution of the architecture, resulting in jagged boundaries. This coarse blueprint is thus forwarded to Stage II for dynamic refinement.

3.4 Multi-Layer Semantic Voting

Stage II operates within the reverse process of a pre-trained Latent Diffusion Model (LDM) [28]. While the Stage I prior M_{prior} provides a semantic blueprint, it lacks spatial precision. Conversely, the high-resolution decoder layers of the U-Net [29] capture rich fine-grained structures. To synergize these complementary strengths, we introduce a conservative Multi-Layer Semantic Voting to dynamically refine the semantic guidance.

At each timestep t, we extract feature maps from a set of high-resolution decoder layers $\mathcal{L}$. To interpret these latent features semantically, we inherit the prototype-based paradigm established in Stage I. For a specific layer l, we compute layer-specific prototypes p_k^l from the content features $f_{content}^l$ guided by M_c, reusing the aggregation logic defined in Eq. (1) and Eq. (2). We then perform nearest-neighbor matching in this shared feature space to generate an instantaneous pseudo-segmentation map $\hat{M}_l^t$

$$\hat{M}_l^t = \arg\max_k \langle \mathbf{f}_{style}^l, \mathbf{p}_k^l \rangle \tag{3}$$

where $\mathbf{p}_k^l$ represents the prototype for class k at layer l.

Due to the stochastic nature of the diffusion process, a single layer's prediction $\hat{M}_l^t$ may be unstable. However, semantic structures that are intrinsic to the image should manifest consistently across different layers. We define a Consensus Mask $\mathcal{C}^t$, which activates only when all layers unanimously agree on the classification result, indicating high confidence

$$\mathcal{C}^t(u, v) = \mathbb{I}\left[\forall i, j \in \mathcal{L}, \hat{M}_i^t(u, v) = \hat{M}_j^t(u, v)\right] \tag{4}$$

The refined Style Pseudo-Segmentation Map $\hat{M}_s^t$ is then synthesized by dynamically updating the static Stage I prior only where the diffusion features exhibit high-confidence consensus

$$\hat{M}_s^t(u, v) = \begin{cases} \hat{M}_l^t(u, v) & \text{if } \mathcal{C}^t(u, v) = 1 \\ M_{prior}(u, v) & \text{otherwise} \end{cases} \tag{5}$$

This mechanism ensures that the system benefits from the fine-grained details captured by the diffusion model while maintaining the semantic stability provided by the foundation model. This refined map $\hat{M}_s^t$ serves as the ground truth for the subsequent filtering and alignment modules, ensuring that style injection is guided by sharp, dynamically corrected semantic boundaries.

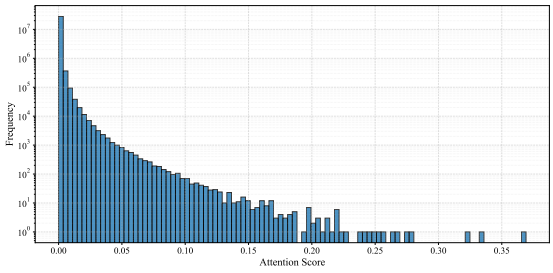

Fig. 3: Long-tailed distribution of attention scores. Valid correspondences appear as sparse, high-value outliers against dominant background noise.

3.5 Semantically Constrained Adaptive Attention Filtering

Standard cross-attention mechanisms are prone to semantic leakage in asymmetric settings. While recent methods like CACTIF [5] filter correspondences using a fixed percentile of feature similarity, this imposes a rigid assumption of a constant signal-to-noise ratio across all images, layers, and timesteps. To dynamically isolate valid style features, an adaptive thresholding strategy is essential.

A straightforward strategy is to define this boundary using standard parametric statistics, such as the mean and standard deviation. However, visualizing the cross-attention affinity matrix (Fig. 3) reveals a severe heavy-tailed distribution: valid structural correspondences manifest as extremely sparse, high-value outliers against dominant background noise. In such highly skewed data, these outliers severely bias the mean and inflate the variance, rendering standard statistics unreliable. To rigorously address this, we propose a dual-constrained filtering mechanism, prioritizing a robust statistical metric followed by an explicit semantic barrier.

Let $\mathbf{A}$ denote the raw cross-attention map. Instead of the conventional mean and standard deviation, we utilize the robust median and Median Absolute Deviation (MAD). This robust thresholding strategy is designed to suppress the influence of rare extreme attention responses while retaining reliable style correspondences. Specifically, the MAD is defined as

$$\mathrm{MAD}(\mathbf{A}) = \mathrm{Median}(|\mathbf{A} - \mathrm{Median}(\mathbf{A})|) \tag{6}$$

Then, the decision boundary τ and the resulting statistical mask $\mathcal{M}_{stat}$ are computed as follows

$$\tau = \mathrm{Median}(\mathbf{A}) + \lambda \cdot \mathrm{MAD}(\mathbf{A}) \tag{7}$$

$$\mathcal{M}_{stat}(i,j) = \mathbb{I}[\mathbf{A}_{i,j} > \tau] \tag{8}$$

While prior works like CACTIF [5] rely entirely on feature similarities and lack explicit semantic boundaries during filtering, this inevitably leads to cross-class texture pollution. To prohibit this, we construct a binary semantic mask $\mathcal{M}_{sem}$ leveraging the refined pseudo-label $\hat{M}_s^t$ from Eq. (5)

$$\mathcal{M}_{sem}(i,j) = \mathbb{I}[M_c(i) = \hat{M}_s(j)] \cdot \mathbb{I}[\hat{M}_s(j) \neq \emptyset] \tag{9}$$

The final attention filter is the intersection

$$\mathcal{M} = \mathcal{M}_{sem} \odot \mathcal{M}_{stat} \tag{10}$$

To prevent feature collapse and preserve structural integrity, we implement a soft-gated fallback mechanism. We compute the filtered style features $\mathbf{h}_i^{style}$, where $\mathbf{h}$ denotes the hidden state of the attention layer, and a spatial preservation score $\alpha_i \in [0,1]$, which quantifies the reliability of the injected style

$$\mathbf{h}_i^{style} = \sum_j \left(\mathcal{M}(i,j) \cdot \mathbf{A}(i,j)\right) \mathbf{v}_j^{style} \tag{11}$$

$$\alpha_i = \sum_j \mathcal{M}(i,j) \cdot \mathbf{A}(i,j) \tag{12}$$

The final hidden state $\mathbf{h}_i$ is then synthesized via adaptive mixing

$$\mathbf{h}_i = \mathbf{h}_i^{style} + (1 - \alpha_i) \cdot \mathbf{h}_i^{content} \tag{13}$$

where $\mathbf{h}_i^{content}$ denotes the original output of the content-branch attention.

This dynamic formulation ensures the model prioritizes stylized textures $\mathbf{h}^{style}$ in high-confidence regions ($\alpha_i \to 1$), while reverting to the original content structure $\mathbf{h}^{content}$ where correspondences are ambiguous or absent ($\alpha_i \to 0$).

3.6 Pixel-Proportion Modulated Hybrid AdaIN

Standard Class-wise AdaIN [5] relies on the ideal assumption that pixel-perfect semantic annotations are available for both domains. However, this assumption collapses in our asymmetric setting, where the style segmentation $\hat{M}_s$ is inferred and inevitably noisy. Consequently, a specific semantic class might occupy a significant portion of the synthetic content image but appear only as a few fragmented pixels or be entirely absent in the single real-world style reference. Relying on such noisy statistics to normalize a large content region can lead to catastrophic feature distortion, where the style of a few outliers dominates the appearance of the entire object.

To mitigate this, we propose a robust normalization scheme that dynamically calibrates the strength of style injection. By explicitly tying the reliability of feature statistics to their relative pixel abundance, it prevents blind statistical alignment onto unreliable regions.

For each semantic class $k \in \{0, \ldots, C-1\}$, we first compute the channel-wise mean and standard deviation for both the content features $\mathbf{F}_c$ (denoted as $\mu_{content}^{(k)}, \sigma_{content}^{(k)}$) and the style features $\mathbf{F}_s$ (denoted as $\mu_{style}^{(k)}, \sigma_{style}^{(k)}$) within their respective masks, $M_c^{(k)}$ and $\hat{M}_s^{(k)}$. Let $N_c^k = |M_c^{(k)}|$ and $N_s^k = |\hat{M}_s^{(k)}|$ denote

the number of pixels for class k in the content and style masks, respectively. We define the relative difference ratio

$$\delta_k = \frac{N_c^k - N_s^k}{N_c^k + N_s^k} \tag{14}$$

The modulation factor is computed via a sigmoid-based confidence function

$$\beta_k = Sigmoid\left(\gamma \cdot \delta_k\right) \tag{15}$$

where γ is a scaling hyperparameter.

The final stylized feature $\mathbf{F}_{out}$ is synthesized by linearly interpolating between original content features and AdaIN-transformed features, weighted by the class-specific modulation map $\mathbf{B}$, where $\mathbf{B}(i,j) = \beta_k$ for pixels belonging to class k.

$$\mathbf{F}_{aligned} = \sigma_{style}^k \left(\frac{\mathbf{F}_c - \mu_{content}^k}{\sigma_{content}^k} \right) + \mu_{style}^k \tag{16}$$

$$\mathbf{F}_{out} = \mathbf{B} \odot \mathbf{F}_c + (\mathbf{1} - \mathbf{B}) \odot \mathbf{F}_{aligned} \tag{17}$$

This pixel-proportion modulation ensures that style transfer is aggressive only where the semantic evidence is abundant and trustworthy, while gracefully degrading to content preservation in ambiguous or sparse regions.

4 Experiments

4.1 Experimental Setup

Following established protocols in prior literature [5], we utilize 5,000 annotated images from the GTA dataset [27] as the synthetic source. For the target domain, we evaluate our approach under an asymmetric setting using a single unlabeled reference image. As a representative case study, the main text focuses on the "Vegetation Dominant" scene (Fig. 1, upper-right). The robust thresholding parameter λ is set to 2.0, and the fusion scale γ is set to 8.0. Evaluation relies on SegFormer [36] for segmentation and LPIPS [40] for structural fidelity. All experiments are conducted on a single NVIDIA RTX 4090 GPU. We benchmark ASTModel against training-free diffusion-based method: Cross-Image Attention [1] and CACTIF [5]. Comprehensive experimental details and results across diverse target scenarios are provided in the Supplementary Material.

4.2 Qualitative Evaluation

Accurate Semantic Alignment As shown in the first row of Fig. 4, baselines erroneously map "mountain" regions to "building" textures, as they rely on superficial appearance similarity while overlooking semantic distinction. Conversely, ASTModel preserves semantic identity through Prototype-Guided Style Semantic Prior Extraction and Multi-Layer Semantic Voting. By anchoring semantic prototypes with foundation model priors and exploiting cross-layer diffusion consistency, our framework ensures content queries attend purely to semantically congruent style keys.

Fig. 4: Qualitative comparison of ASTModel and baseline methods.

Mitigating Semantic Leakage In Fig. 4, baselines suffer severe semantic leakage by projecting dominant greenery onto roads and vehicles. ASTModel mitigates this via two mechanisms: Semantically Constrained Adaptive Attention Filtering enforces strict semantic barriers for texture injection, while Pixel-Proportion Modulated Hybrid AdaIN calibrates stylization intensity based on the statistical reliability of each class. It prevents overfitting to dominant textures, preserving the semantic purity in asymmetric settings.

High-Fidelity Structural Preservation As shown in the third row of Fig. 4, baselines suffer from significant detail loss, often reducing vehicles to blurred shapes. Conversely, ASTModel preserves high-fidelity structures, maintaining crisp contours and fine-grained textures for intricate objects like vehicle out-

Table 1: Quantitative Evaluation on Downstream Segmentation

Method	Pixel Acc ↑	mIoU ↑
Source Only [27]	0.844	0.275
Cross-Image Attn. [1]	0.653	0.224
CACTIF [5]	0.782	0.289
ASTModel (Ours)	**0.847**	**0.309**

Table 2: Quantitative Evaluation on Structural Fidelity

Method	LPIPS ↓
Cross-Image Attn. [1]	0.5205
CACTIF [5]	0.4184
ASTModel (Ours)	**0.3588**

Table 3: Computational Efficiency Comparison. Total refers to the entire 5,000 images.

Method	Time ↓	Total ↓
Cross-Image Attn. [1]	18s	28h
CACTIF [5]	80s	114h
ASTModel (Ours)	25s	37h

Table 4: Quantitative Ablation Study

Method	LPIPS ↓
w/o Attention Filtering	0.5114
w/o Hybrid AdaIN	0.3737
w/o Semantic Voting	0.3641
Ours (Full Model)	**0.3588**

lines and architectural elements. This superior clarity ensures that the generated datasets serve as high-quality training data for downstream perception tasks.

Robustness to Reference-Unseen Scenarios In challenging scenarios like "Tunnel" and "Sunset" (bottom two rows of Fig. 4), baselines struggle with absent style references, generating noise by forcing invalid correspondences. ASTModel maintains structural fidelity by detecting these weak signals: Attention Filtering reverts to content features when correspondences drop, and Pixel-Proportion Modulated Hybrid AdaIN suppresses injection for statistically sparse features. This dynamic adaptability ensures that ASTModel can gracefully handle a wide range of target styles, even those with minimal semantic overlap with the source domain.

4.3 Quantitative Evaluation

Downstream Perception Utility The primary metric for synthetic-to-real adaptation is its impact on downstream perception tasks. As shown in Tab. 1, our ASTModel consistently achieves the highest Pixel Accuracy and mIoU, demonstrating its superior ability to effectively transfer target styles while strictly preserving original semantic contents compared to all baselines. Further details on the segmentation model's training and validation are provided in the Supplementary Material.

Structural Fidelity Beyond semantic correctness, autonomous driving applications mandate the strict preservation of geometric layouts and structural

Fig. 5: Ablation study on key components of ASTModel.

boundaries. Tab. 2 demonstrates that ASTModel achieves the lowest LPIPS score, indicating minimal structural distortion. The baselines exhibit higher LPIPS scores due to severe spatial warping and hallucination artifacts in the asymmetric setting.

Computational Efficiency Tab. 3 details the inference latency for synthesizing the 5,000-image dataset. CACTIF suffers from a severe computational bottleneck (80s/image, 114h total) due to exhaustive pixel-wise feature similarity calculations in its attention filtering. In contrast, ASTModel achieves a 3.2× speedup (25s/image). This high efficiency is achieved by replacing computationally heavy dense matching with robust statistical metrics to establish adaptive decision boundaries.

4.4 Ablation Study

Effectiveness of Semantically Constrained Adaptive Attention Filtering Without filtering, the model incorrectly maps texture from the vegetation-heavy style reference onto shadowed road regions in the content image (Fig. 5). This results in cross-class pollution, manifested as unnatural greenish, moss-like textures on the road, spiking LPIPS to 0.5114 (Tab. 4). Our semantic barrier successfully restricts style injection exclusively to congruent categories.

Effectiveness of Pixel-Proportion Modulated Hybrid AdaIN Reverting to global AdaIN blindly forces dominant style statistics onto the entire image, causing severe color casts and identity loss (*e.g.*, the darkened SUV in Fig. 5, Row 2). Additionally, the overall scene suffers from an unnatural global color cast. Our hybrid approach aligns statistics within corresponding categories and scales injection strength based on pixel abundance.

Effectiveness of Multi-Layer Semantic Voting Without this voting mechanism, the coarse DINO prior fails to precisely isolate small, irregular regions in the style reference, such as fragmented skies or thin utility poles. Consequently,

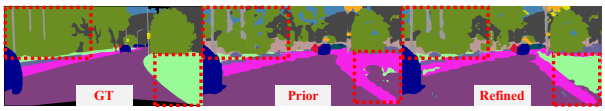

Fig. 6: Validation of pseudo-label refinement.

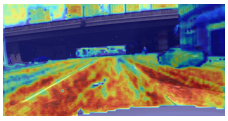

Fig. 7: Visualization of α.

mismatched style features are transferred: the sky region is corrupted by unrelated textures and artifacts (Fig. 5, Row 1), and poles suffer from feature loss, resulting in disconnected or blurred structures (Fig. 5, Row 2). The quantitative benefit of correcting these artifacts is evident in Tab. 4. By aggregating consensus from the layers of the diffusion decoder, our mechanism refines these coarse priors and preserves intricate details.

4.5 Pseudo-Label Refinement Validation

We further evaluate the quality of the style-side pseudo labels. As shown in Fig. 6, the Stage-I prior captures coarse semantics but contains inaccurate boundaries and local mismatches. Stage-II refinement improves boundary alignment and corrects ambiguous regions. Quantitatively, Pixel Accuracy increases from 0.688 to 0.790, and mIoU improves from 0.281 to 0.339, verifying that multi-layer semantic voting provides more reliable guidance for class-consistent style injection. Notably, the refinement is adaptive to content-style semantic compatibility: stronger semantic alignment between the content image and style reference leads to more consistent multi-layer voting, more accurate pseudo labels, and consequently higher-quality style transfer.

4.6 Visualization of Preservation Score

We visualize the preservation score α in Eq. (12) to analyze how ASTModel balances style injection and content preservation. As shown in Fig. 7, regions with reliable semantic correspondences obtain higher α and receive stronger style transfer. In contrast, regions that are absent or weakly represented in the style image obtain lower α, causing the model to fall back to content preservation. For example, the bridge region has low α because no corresponding bridge appears in the style image, which prevents invalid texture injection and reduces semantic leakage.

4.7 Robustness of Median/MAD Filtering

We further validate the proposed Median/MAD threshold against a Mean/Std-based alternative. As shown in Tab. 5, removing the top-20 largest attention scores changes the Std-based scale from 173.3 to 171.9, while the Median/MAD statistics remain unchanged. Mean/Std filtering preserves only 1.8% of attention entries, leading to over-filtering and insufficient style transfer. In contrast, Median/MAD preserves 33.2% of entries and retains more reliable correspondences

Table 5: Robustness comparison. Values are scaled by 10^{-5}.

Stat.	Orig.	Top-rm.
Mean/Std	41.8/173.3	41.8/171.9
Med./MAD	5.1/4.8	5.1/4.8

Fig. 8: Visual comparison between Mean/Std and Median/MAD filtering.

under the heavy-tailed attention distribution. This difference is also reflected in Fig. 8: Mean/Std under-transfers the road region and leaves source-domain textures insufficiently adapted, whereas Median/MAD produces more complete and natural road-style transfer.

4.8 Discussion

Scalability with Foundation Models As a modular, training-free framework, ASTModel is inherently scalable. Future upgrades to vision encoder or diffusion backbone will translate to enhanced semantic alignment and photorealism.

Robustness Across Domains Conventional adaptation methods require per-domain retraining. ASTModel derives its robust generalizability fundamentally from its reference-guided style transfer paradigm. This enables our framework to seamlessly tackle other critical distribution shifts, such as severe weather and day-to-night translation. Results are provided in the Supplementary Material.

5 Conclusion

In this paper, we introduced Asymmetric Style Transfer for Autonomous Driving (ASTAD), a task formulation that addresses the practical setting where dense semantic labels are available for synthetic source data but absent for real-world style references. To tackle this challenge, we proposed ASTModel, a training-free framework designed to bridge the synthetic-to-real gap under asymmetric conditions. Experiments demonstrate that ASTModel improves semantic consistency and inference efficiency over relevant training-free baselines under the evaluated protocol.

A limitation of this work is that the current protocol relies on a single reference image, whose scene coverage is inherently limited. As shown in Fig. 7, for example, the rear part of the car receives a stronger response than the front because car-related cues in the reference are spatially imbalanced and dominated by rear-side regions. Future work will extend ASTAD to multiple reference images, which can provide more balanced category-level style statistics, cover sparse local cues more comprehensively, and reduce the conflict between semantic safety and style richness in challenging scenes. By transforming label-rich synthetic data into semantically reliable target-style samples, ASTAD offers a practical path toward scalable and robust perception model development for autonomous driving.

References

1. Alaluf, Y., Garibi, D., Patashnik, O., Averbuch-Elor, H., Cohen-Or, D.: Cross-image attention for zero-shot appearance transfer. In: ACM SIGGRAPH 2024 conference papers. pp. 1–12 (2024)
2. An, J., Huang, S., Song, Y., Dou, D., Liu, W., Luo, J.: Artflow: Unbiased image style transfer via reversible neural flows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 862–871 (2021)
3. Cheng, B., Liu, Z., Peng, Y., Lin, Y.: General image-to-image translation with one-shot image guidance. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22736–22746 (2023)
4. Cheng, B., Li, J., Shi, J., Fang, Y., Zhang, G., Chen, Y., Zeng, T., Li, Z.: Weafu: Weather-informed image blind restoration via multi-weather distribution diffusion. IEEE Transactions on Circuits and Systems for Video Technology (2024)
5. Chigot, E., Wilson, D.G., Ghrib, M., Oberlin, T.: Style transfer with diffusion models for synthetic-to-real domain adaptation. Computer Vision and Image Understanding p. 104445 (2025)
6. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018)
7. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8188–8197 (2020)
8. Chung, J., Hyun, S., Heo, J.P.: Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8795–8805 (2024)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
10. Deng, Y., He, X., Tang, F., Dong, W.: Z^*: Zero-shot style transfer via attention rearrangement. arXiv preprint arXiv:2311.16491 (2023)
11. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: Stytr2: Image style transfer with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11326–11336 (2022)
12. Frenkel, Y., Vinker, Y., Shamir, A., Cohen-Or, D.: Implicit style-content separation using b-lora. In: European Conference on Computer Vision. pp. 181–198. Springer (2024)
13. Gao, M., Dong, Q.: Adaptive conditional denoising diffusion model with hybrid affinity regularizer for generalized zero-shot learning. IEEE Transactions on Circuits and Systems for Video Technology **34**(7), 5641–5652 (2024)
14. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
15. Go, S., Choi, K., Shin, M., Uh, Y.: Eye-for-an-eye: Appearance transfer with dense semantic correspondence in diffusion models. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision pp. 4641–4650 (2026)

16. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International conference on machine learning. pp. 1989–1998. Pmlr (2018)
17. Hoyer, L., Dai, D., Van Gool, L.: Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9924–9935 (2022)
18. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
19. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV). pp. 172–189 (2018)
20. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
21. Jia, Y., Hoyer, L., Huang, S., Wang, T., Van Gool, L., Schindler, K., Obukhov, A.: Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In: European Conference on Computer Vision. pp. 91–109. Springer (2024)
22. Kwon, G., Ye, J.C.: Diffusion-based image translation using disentangled style and content representation. arXiv preprint arXiv:2209.15264 (2022)
23. Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding, E.: Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6649–6658 (2021)
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research Journal (2024)
25. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European conference on computer vision. pp. 319–345. Springer (2020)
26. Park, T., Zhu, J.Y., Wang, O., Lu, J., Shechtman, E., Efros, A., Zhang, R.: Swapping autoencoder for deep image manipulation. Advances in Neural Information Processing Systems **33**, 7198–7211 (2020)
27. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: European conference on computer vision. pp. 102–118. Springer (2016)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 10674–10685. IEEE (2022)
29. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
30. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
31. Song, Z., He, Z., Li, X., Ma, Q., Ming, R., Mao, Z., Pei, H., Peng, L., Hu, J., Yao, D., et al.: Synthetic datasets for autonomous driving: A survey. IEEE Transactions on Intelligent Vehicles **9**(1), 1847–1864 (2023)

32. Song, Z., Yao, D., Ming, R., Peng, L., Yao, D., Zhang, Y.: Synthetic dataset evaluation based on generalized cross validation. arXiv preprint arXiv:2509.11273 (2025)
33. Tumanyan, N., Bar-Tal, O., Bagon, S., Dekel, T.: Splicing vit features for semantic appearance transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10748–10757 (2022)
34. Wang, Z., Gao, H.a., Zhang, G., Zhao, H.: Weather-diff: Towards arbitrary adversarial weather generation with diffusion models. In: International Conference on Intelligent Computing. pp. 134–145. Springer (2025)
35. Wang, Z., Zhao, L., Xing, W.: Stylediffusion: Controllable disentangled style transfer via diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7677–7689 (2023)
36. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)
37. Yan, Q., Hu, T., Sun, Y., Tang, H., Zhu, Y., Dong, W., Van Gool, L., Zhang, Y.: Toward high-quality hdr deghosting with conditional diffusion models. IEEE Transactions on Circuits and Systems for Video Technology **34**(5), 4011–4026 (2023)
38. Yan, Q., Hu, T., Wu, P., Dai, D., Gu, S., Dong, W., Zhang, Y.: Efficient image enhancement with a diffusion-based frequency prior. IEEE Transactions on Circuits and Systems for Video Technology (2025)
39. Yao, D., Han, X., Ming, R., Song, Z., Peng, L., Hu, J., Yao, D., Zhang, Y.: A style-based profiling framework for quantifying the synthetic-to-real gap in autonomous driving datasets. arXiv preprint arXiv:2510.10203 (2025)
40. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
41. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10146–10156 (2023)
42. Zhao, Y., Zhong, Z., Zhao, N., Sebe, N., Lee, G.H.: Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In: European conference on computer vision. pp. 535–552. Springer (2022)
43. Zhou, Y., Simon, M., Peng, Z., Mo, S., Zhu, H., Guo, M., Zhou, B.: Simgen: Simulator-conditioned driving scene generation. Advances in Neural Information Processing Systems **37**, 48838–48874 (2024)
44. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)