Style Nursing with Spatial and Semantic Guidance for Zero-Shot Traffic Scene Style Transfer

Zhen Wang[1,3,4], Zihang Lin[2], Meng Yuan[2,3,4], Yuehu Liu[2,3,4], Chi Zhang[2,3,4]*

[1]School of Software Engineering, Xi'an Jiaotong University
[2]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[3]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence
[4]National Engineering Research Center for Visual Information and Applications
wangzhen_wz@stu.xjtu.edu.cn, chizhang@xjtu.edu.cn

Abstract

Recent advances in text-to-image diffusion models have shown an outstanding ability in zero-shot style transfer. However, existing methods often struggle to balance preserving the semantic content of the input image and faithfully transferring the target style in line with the edit prompt. Especially when applied to complex traffic scenes with diverse objects, layouts, and stylistic variations, current diffusion models tend to exhibit *Style Neglection*, i.e., failing to generate the required style in the prompt. To address this issue, we propose *Style Nursing*, which directs the model to focus on style subject tokens in the text prompt and excites their corresponding visual activations. Moreover, we introduce *Spatial and Semantic Guidance* to guide the preservation of content after editing, which utilizes spatial features from the DDIM sampling process together with attention maps from the semantic reconstruction. To evaluate the performance of zero-shot style transfer methods in traffic scenes, we present *STREET-6K*, a new benchmark dataset comprising 6000 images showcasing diverse traffic scenes and style transfer variations, accompanied by comprehensive annotations and evaluation metrics. Our approach beats state-of-the-art image translation methods in comprehensive quantitative metrics and human evaluations on traffic scene image synthesis while seamlessly generalizing to various other types of images without training or fine-tuning. Further experiments on detection and segmentation tasks show that fine-tuning perception models on our synthesized images improves Recall and mean Intersection over Union (mIoU) by over 10% and 3% respectively in rarely-seen traffic scenes.

Introduction

As visual data of real-world traffic scenes tend to have skewed data with long-tailed characteristics, acquiring image datasets depicting various traffic scene styles (e.g., extreme weather conditions and complex lighting scenarios) is of great significance to autonomous driving testing. Several studies (Jeon et al. 2023; Zheng, Lu, and Narasimhan 2024; Lee et al. 2023; Zhang et al. 2022) used image-to-image translation techniques to generate images containing rarely-seen traffic layouts, styles, or conditions from normal traffic images. While these methods excel in traffic scene style transfer, they often necessitate tailored methods or datasets for specific styles, limiting their generalizability to other traffic scene style transfers.

Recent advancements in text-guided image generation models (Rombach et al. 2022; Ruiz et al. 2023; Gal et al. 2023a; Voynov et al. 2023; Gal et al. 2023b; Kang et al. 2024) have showcased an unparalleled zero-shot style transfer capability to produce high-quality and diverse images. However, the intrinsic challenge still lies in *maintaining the essential content of the source image while ensuring fidelity to the target prompt*. Several prior and concurrent studies (Geng et al. 2024; Mou et al. 2024; Zhang, Rao, and Agrawala 2023; Parmar et al. 2023; Zhang et al. 2024; Wei et al. 2023) have sought to address these issues. P2P (Hertz et al. 2023) keeps the original image content intact by manipulating the cross-attention layers. Plug-and-Play (Tumanyan et al. 2023) presents feature injection to retain the content of the guidance image. InstructPi2Pix (Brooks, Holynski, and Efros 2023) achieves preserving the original content by training a conditional diffusion model.

However, when applied to complex traffic scenes with diverse objects, layouts, and stylistic variations, current diffusion models tend to exhibit *Style Neglection*, failing to generate one or more style subjects from the prompt, as illustrated in Fig. 1. To mitigate this issue and enhance style generation, we introduce the concept of *Style Nursing*, drawing inspiration from (Chefer et al. 2023). Intuitively, for a style to be present in the generated image, the model should assign at least one image patch to the style token. *Style Nursing* embodies this intuition by demanding that each style token be dominant in some image patches, which carefully guides the denoising process and encourages the model to focus on style tokens by enhancing their activations.

Moreover, we propose *Spatial and Semantic Guidance* to preserve the detailed content structure of the input image, which retains both the spatial features of the input image and self-attention maps of the corresponding semantic image throughout the diffusion process. We investigate the internal encoding of content information in a pre-trained text-to-image diffusion model and delve into the spatial features formed during the generation process. Manipulating these features enables fine-grained control over the generated content (Tumanyan et al. 2023). In addition, our empirical analysis reveals that spatial features from different U-Net lay-

*Corresponding Author

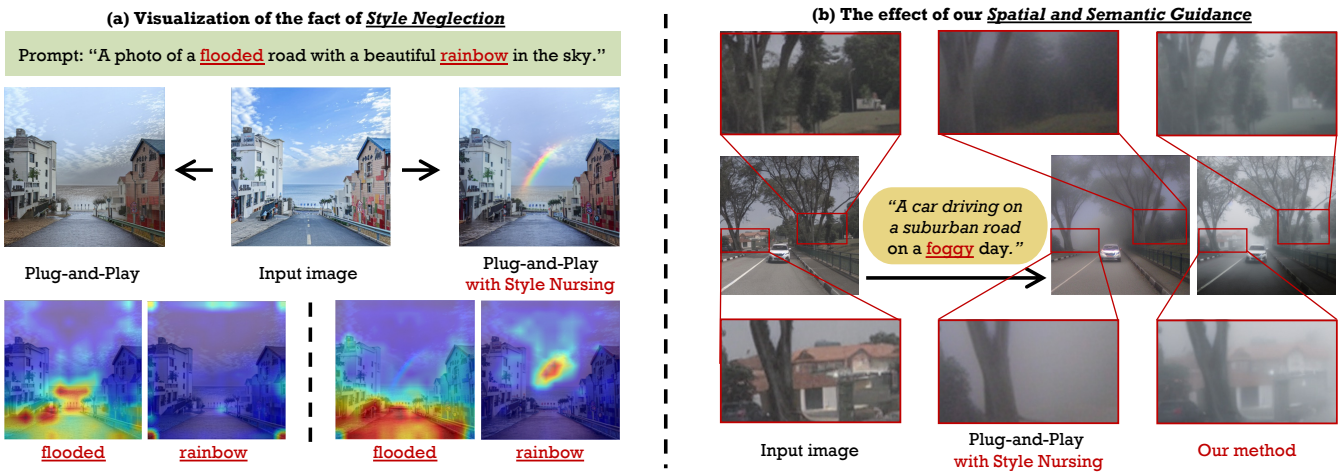

Figure 1: Illustration on the *Style Neglection* phenomenon and the superiority of *Style Nursing with Spatial and Semantic Guidance*. (a) illustrates the existence of *Style Neglection* in current diffusion-based image translation method, e.g., Plug-and-Play (Tumanyan et al. 2023). Visualization of the cross-attention maps for each style token demonstrates the effect of *Style Nursing*. (b) illustrates the effectiveness of the proposed *Spatial and Semantic Guidance*. Our method preserves the detailed content of traffic scenes compared with others, ensuring that the trees and houses in the background do not deform or disappear.

ers have varying degrees of control over the attributes of the synthesized image, i.e., the coarse layers primarily affect its content structure while the fine layers predominantly influence its appearance. Therefore, we propose *Spatial Guidance* to keep the spatial features consistent before and after translation. Specifically, spatial features from coarse layers of U-Net during the image reconstruction process are extracted as guidance features, which guarantees the image content does not deviate by spatial gradient optimization in the process of text-guided generation. However, *Spatial Guidance* is insufficient to better preserve the object structure in complex traffic scenes, e.g., Fig. 5(a). To address this issue, we employ *Semantic Guidance* to ensure precise preservation of semantic structure. The self-attention maps of the semantic image reconstruction process are utilized to guide the latent sample during the style transfer process by gradient optimization. To address the lack of standardized benchmarks for image translation in traffic scene style transfer, we introduce *STREET-6K* (**S**tylized **TR**affic Sc**E**ne **E**diting **T**est Benchmark with **6K** images) from diverse traffic scenes spanning 10 style categories. Each entry includes a source image, source prompt, target prompt, editing instruction and editing type. As a training-free approach, our method outperforms existing techniques in traffic scene style transfer on 7 metrics and can generalize to other forms of style transfer robustly. To validate the effectiveness of data generated by our method, we further fine-tune autonomous driving perception models Yolop (Wu et al. 2022) and HybridNets (Vu, Ngo, and Phan 2022) with *STREET-6K*, mitigating the performance degradation problem due to lack for data under rarely-seen traffic scenes. Our key contributions are summarized as follows:

(i) We identify the *Style Neglection* phenomenon in current diffusion models when applied to traffic scene style transfer. To address this, we propose *Style Nursing*, which ensures each style token is dominant in specific image patches.
(ii) We propose the *Spatial and Semantic Guidance* for content preserving, utilizing spatial features and self-attention maps to direct the gradient optimization without extra training or fine-tuning.
(iii) We introduce *STREET-6K* which fills the gap in standardized benchmarks for image editing in the traffic domain and we validate the significance of our approach to autonomous driving with it.

Preliminary

Latent diffusion model. Our method leverages a pre-trained text-conditioned Latent Diffusion Model (Rombach et al. 2022), the Stable Diffusion (SD), which operates in the latent space of an autoencoder. SD aims to map a random noise vector z_t and textual condition P to an output image z_0. To perform sequential denoising, the network ϵ_θ is trained to predict artificial noise, following the objective:

$$\min_\theta \mathbb{E}_{z_0,\epsilon\sim N(0,1),t\sim Uniform(1,T)}||\epsilon - \epsilon_\theta(z_t, t, C)||_2^2 \quad (1)$$

Note that C is the embedding of the text condition and z_t is a noised sample, where noise is added to the sampled data z_0 according to timestamp t. At inference, given a noise vector z_t, The noise is gradually removed by sequentially predicting it using our trained network for T steps. Here, ϵ_θ is a U-Net network, and each layer of the U-Net comprises a residual block, a self-attention block, and a cross-attention block. The residual block will compute spatial features f_t^l by convolv latent features from the previous layer $l-1$.

Self-attention block. In the self-attention block, it propagates the feature at each spatial location to a similar region in the feature map. The spatial features f_t^l of the network at layer l are first projected into queries $Q_t^l = f_Q(f_t^l)$, keys

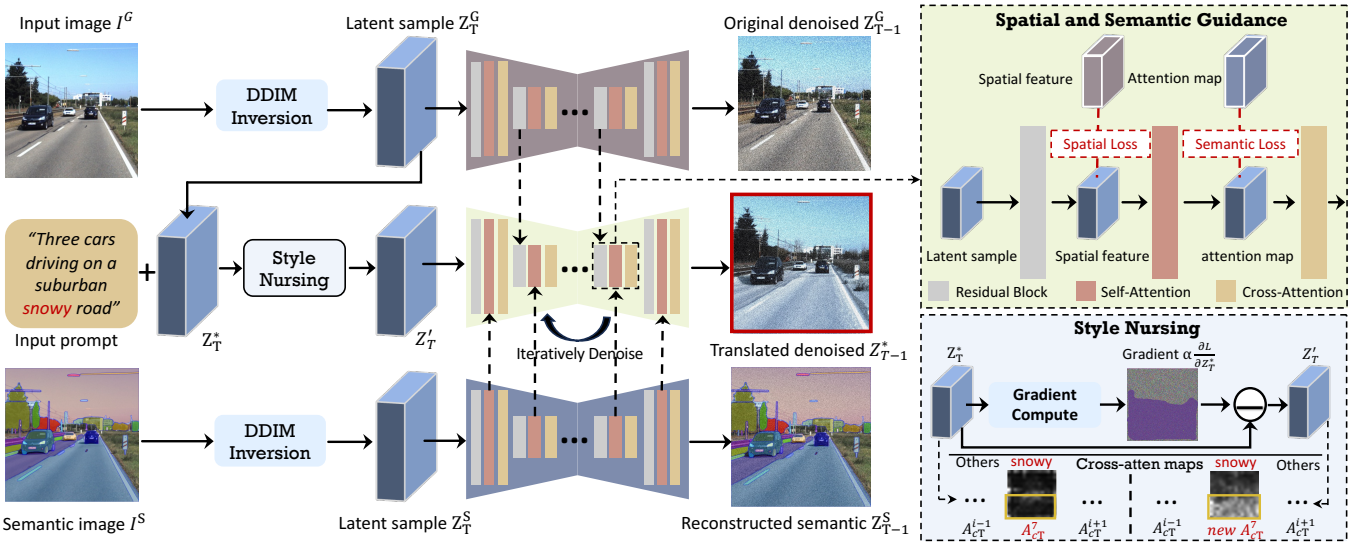

Figure 2: *Style Nursing with Spatial and Semantic Guidance*. Our framework takes a source image I^G and a text prompt that describes the desired style as input. Firstly, I^G is inverted to the latent sample Z_T^G and then updated by *Style Nursing* shown in the bottom right, which enhances the maximal activation for the style token from the cross-attention map. After that, *Spatial and Semantic Guidance* is introduced to prevent the content and structure of output images from changing dramatically. As shown in the top right, spatial features from the coarse layers of the U-Net during the reconstruction of I^G together with attention maps from the reconstruction of the semantic image are used as guidance through gradient optimization.

$K_t^l = f_K(f_t^l)$, and values $V_t^l = f_V(f_t^l)$ through learned linear projections f_Q, f_K,f_V, and the output of the block and self-attention map A_t^l are given by:

$$f_t^l = A_t^l \cdot V_t^l,$$
$$where\ A_t^l = Softmax\left(\frac{Q_t^l \cdot {K_t^l}^T}{\sqrt{d}}\right). \quad (2)$$

This process is applied independently for all queries, enabling the model to capture correspondences across the entire image.

Cross-attention block. Finally, text guidance in SD is performed using the cross-attention mechanism. Specifically, a pretrained CLIP encoder (Radford et al. 2021) is usually used to encode the guided prompt P and obtain the text embeddings C. The keys and values are obtained from text embedding C with a linear mapping. Denote by D the spatial dimension of the intermediate feature map ($i.e., D \in 64, 32, 16, 8$), and by N the number of text tokens in the guided prompt P. Then, we can get a cross-attention map $A_{ct} \in \mathbb{R}^{D \times D \times N}$ by queries computed from the output of self-attention block f_t^l with the keys computed from text embedding C at timestep t. A_{ct} defines a distribution over the text tokens for each spatial patch (i, j). Specifically, A_{ct}^n[i,j] denotes the probability assigned to token n for the (i, j)-th spatial patch of the intermediate feature map. Intuitively, this probability indicates the amount of information that will be passed from token n to patch (i, j). Note that the maximum value of each of the $D \times D$ cells is 1.

Method

We propose *Style Nursing with Spatial and Semantic Guidance* as illustrated in Fig. 2, given an input image I^G and a target style prompt P, our goal is to generate a new image I^* with different style that complies with P and preserves the content and spatial semantic layout of I^G.

Style Nursing

The core of *Style Nursing* is gradually shifting the noised latent at each timestep t toward a more semantically faithful generation. At each denoising step, we consider the cross-attention maps of the style subject tokens in the prompt P. Intuitively, visual information in the synthesized image should strongly influence some patches. As such, we define a loss objective that attempts to maximize the cross-attention values for each style subject token. We then update the noised latent at time t according to the gradient of the computed loss. This encourages the latent at the next timestep to better generate the forgotten style described in the prompt.

Cross-Attention Maps. Given the input prompt P, we consider the set of style tokens $S = s_1, ..., s_k$ present in P. Then, we extract cross-attention maps A_{ct}^s for each token $s \in S$, indicating the influence of s on each image patch. After the Softmax operation, the (i, j)-th entry of the resulting matrix A_{ct}^s indicates the probability of each textual token being present in the corresponding image patch. We then extract the 16×16 normalized attention maps A_{ct}^s for each style subject token s. We apply a Gaussian filter over A_{ct}^s to obtain smooth attention maps. After that,

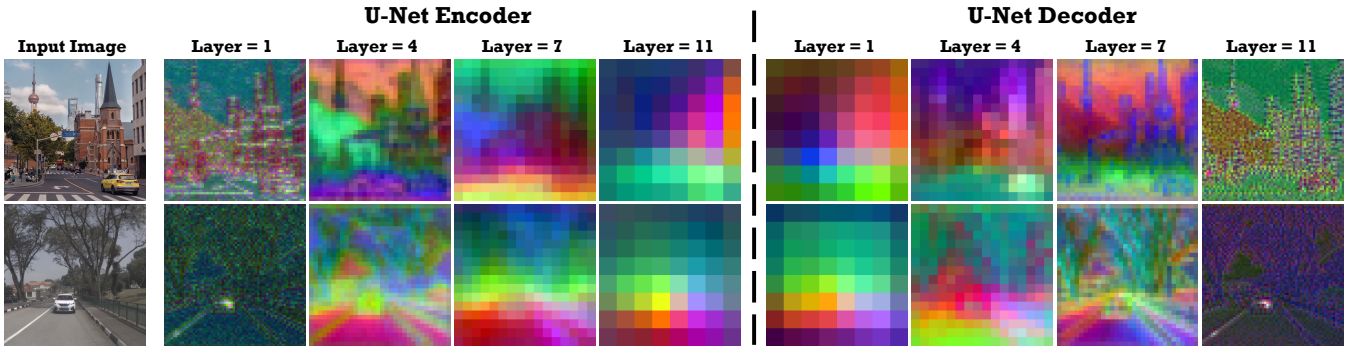

Figure 3: Visualizing spatial features. We collect different types of real traffic scene images and extract spatial features from different U-net layers at roughly 80% of the generation process (t = 200). For each block, we applied PCA on the extracted features across all images and visualized the top three leading components.

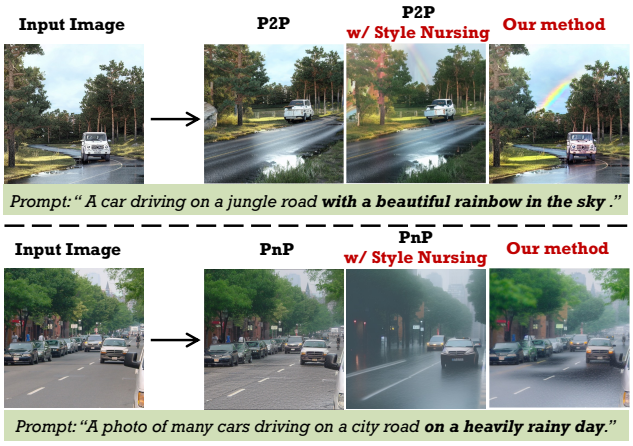

Figure 4: The effect of our proposed method. Although *Style Nursing* can address the challenge of *Style Neglection*, current diffusion-based image translation methods (e.g. P2P (Hertz et al. 2023), PnP (Tumanyan et al. 2023)) cannot well preserve the content structure of the input image.

the attention value of the maximally-activated patch is dependent on its neighboring patches since each patch is a linear combination of its neighboring patches.

Gradient Adjustment. Successfully generated styles should have an image patch that significantly attends to their corresponding token. For each style token in S, we promote the presence of at least one patch in A^s_{ct} exhibiting a high activation value. The corresponding loss function is as follows:

$$\mathcal{L} = \max_{s \in S} \mathcal{L}_s \quad where \quad \mathcal{L}_s = 1 - max(A^s_{ct}). \tag{3}$$

Specifically, the loss enhances the activations of the most neglected style token at timestep t. After computing the loss $\mathcal{L}$, the latentz_t is updated as:

$$z'_t \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} \mathcal{L}. \tag{4}$$

where α_t is a scalar that determines the step size for the gradient update. Subsequently, a forward pass through ϵ_θ is performed using z'_t to compute z_{t-1} for the next denoising step. This update process is repeated over a subset of timesteps, based on the observation that the final timesteps have minimal impact on the spatial arrangement of objects in the generated image.

Spatial and Semantic Guidance

Although *Style Nursing* can address the challenge of *Style Neglection*, current diffusion-based image translation methods can not better preserve the content structure of the input image, as shown in Fig. 4. Therefore, to preserve the detailed content structure of the input image after ensuring fidelity to the target prompt, we propose *Spatial and Semantic Guidance*, which aims to retain the spatial features of the input image and self-attention maps of the semantic image throughout the diffusion process.

Spatial Guidance. Each U-Net layer contains a residual, self-attention, and cross-attention block. Spatial features from the residual block can fine-grained control spatial content information during the generation process (Tumanyan et al. 2023). Inspired by (Voynov et al. 2023), which analyzes different U-net layers with varying degrees of control over the attributes of the synthesized image. But these experiments happen in cross-attention blocks. Therefore, we delve into how spatial features from different U-Net layers affect the preservation of content information of the input image. To this end, we perform a simple PCA analysis shown in Fig.3, which allows us to reason that spatial features from U-Net coarse layers most reveal semantic structure and layout. Deeper features from U-Net fine layers capture more high-frequency information, such as appearance color.

Based on these observations, we propose *Spatial Guidance*. First, let Z^G_T be their initial noise, obtained by inverting I^G using DDIM. Given the target prompt P, the generation of the translated image I^* is carried with the same initial noise, i.e., $Z^*_T = Z^G_T$; At each step t of the backward process, we extract the guidance features f^l_t from the denoising step: $Z^G_{\text{t-1}} = \epsilon_\theta(Z^G_t, \varnothing, t)$. Then, in the denoising step of Z^*_t, we take a gradient step with f^{*l}_t from Z^*_t towards matching the

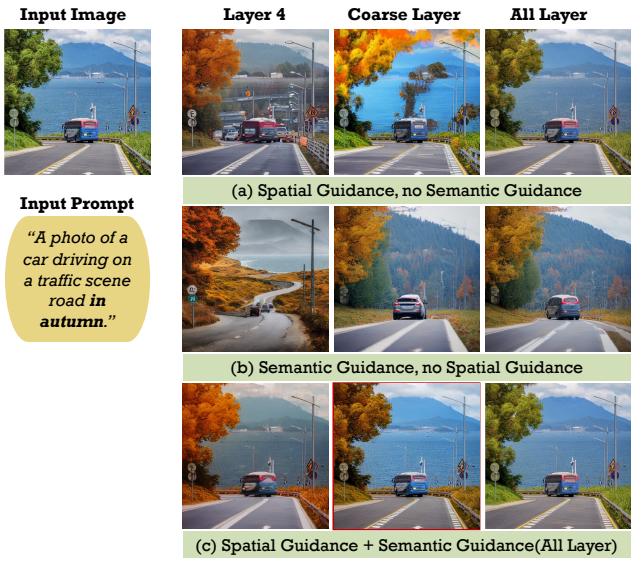

Figure 5: Ablating *Spatial and Semantic Guidance*. (a) Spatial features extracted from the input image guide the generation process of the translated image. While features at layer 4 exhibit localized semantic information (Fig. 3), sole use of these features is insufficient for retaining the structure of the input traffic scene. Incorporating more features from U-Net layers leads to better content preservation, but style transfer results will be affected by the appearance information of the input image (All layers). (b) Generation using semantic self-attention maps as guidance without spatial features. (c) Guiding self-attention maps at all layers and gradually adding spatial features from different layers.

reference f_t^l, reducing the spatial feature loss $\mathcal{L}_{\text{fea}}$ below.

$$\mathcal{L}_{\text{fea}} = ||f_t^{*l} - f_t^l||_2. \tag{5}$$

This loss encourages f_t^{*l} to not deviate from f_t^l after applying the edit. As seen in Fig. 5, guidance features only at layer 4 are insufficient for preserving the structure of the input. The structure is better preserved as we guide features in all U-Net layers. Yet spatial features from input contain much appearance information, which leaks into the generated image. As a result, the generated image is very close to the input, and the effect of text guidance is not obvious. To better balance preserving the content of I^G and style transfer by prompt, we do not modify spatial features at deep layers but rather leverage the *Semantic Guidance* as discussed below.

Semantic Guidance. Self-attention modules compute the affinities A_t^l between the spatial features after linearly projecting them into queries and keys. These affinities are tightly connected to the established concept of self-similarity, which has been used to design structure descriptors by classical and modern works (Tumanyan et al. 2023, 2022; Shechtman and Irani 2007). This motivates us to consider self-attention maps A_t^l to fine-grained control of the structure of the generated content. To allow the diffusion

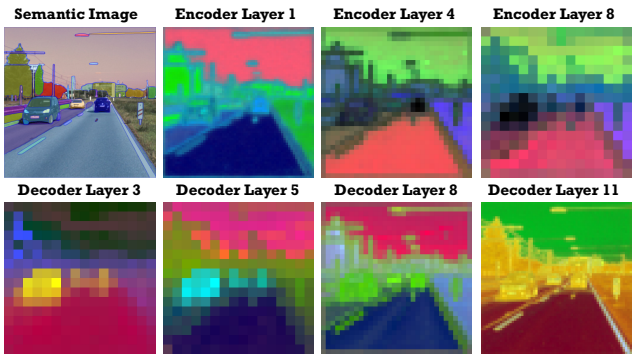

Figure 6: Self-attention visualization. Showing leading components of the self-attention map A_t^l computed for the semantic image for different U-Net layers.

model to better identify semantic structure in the image, we use SAM(Kirillov et al. 2023) to obtain the semantic image and reconstruct it in the denoising process to get reference self-attention maps A_t^l. As seen in Fig. 6, self-attention captures the detailed semantic structure.

Similar to *Spatial Guidance*, we extract the guidance self-attention maps A_t^l from the denoising step: $Z_{\text{t-1}}^S = \epsilon_\theta(Z_t^S, \varnothing, t)$. Then in the denoising step of Z_t^*, we take a gradient step with A_t^{*l} from Z_t^* towards matching the guidance A_t^l , reducing the self-attention loss $\mathcal{L}_{\text{self}}$ below.

$$\mathcal{L}_{\text{self}} = ||A_t^{*l} - A_t^l||_2. \tag{6}$$

This loss encourages A_t^{*l} to not deviate from A_t^l after applying the edit.

Results

Benchmark. Since there is no existing benchmark for image translation in diverse traffic scene, to systematically validate our proposed method for traffic scene style transfer and compare our method with existing image translation methods, as well as compensate for the absence of standardized performance criteria for traffic scene style transfer, we construct a benchmark dataset, named *STREET-6K*. It comprises 6K images featuring 10 distinct editing types, accompanied by versatile annotations and comprehensive evaluation metrics. All the images in *STREET-6K* from large-scale autonomous driving open-source datasets, such as KITTI (Liao, Xie, and Geiger 2022) and nuScenes (Caesar et al. 2020). Additionally, to prove our method has better generalization on synthetic images, we created a synthetic dataset for style transfer and baseline comparisons.

Qualitative Comparisons. We focus qualitative comparisons on some previous and concurrent diffusion-based image editing methods as illustrated in Fig. 7. As can be seen, our method successfully translates diverse traffic scene images and works well for both real and synthetic images. Contrasts SDEdit, which suffers from an inherent tradeoff between the two. With a low noise level, the input content structure is well preserved, but in the expanse of hardly changing appearance, larger appearance deviations can be

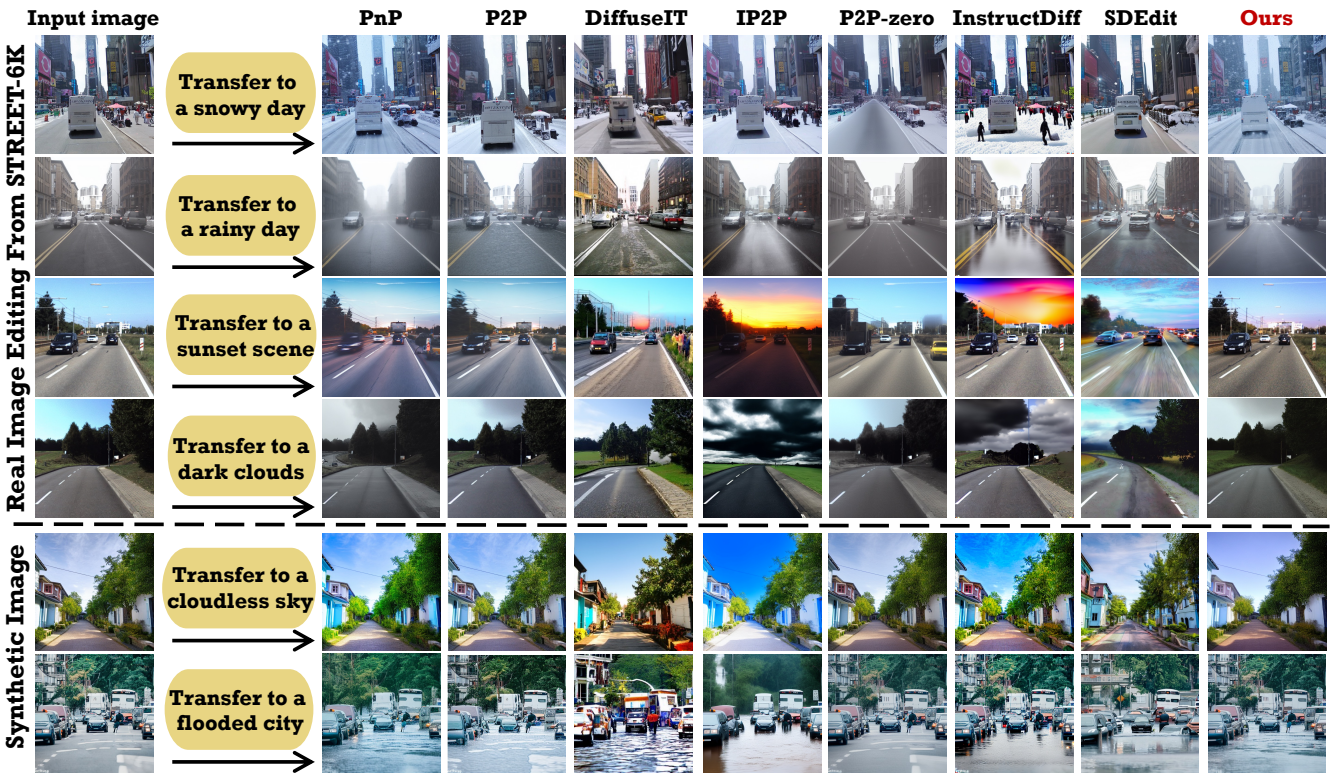

Figure 7: Comparison with different baselines for real and synthetic images. Our method successfully applies the style transfer while fine-grained preserving the content structure of the input image.

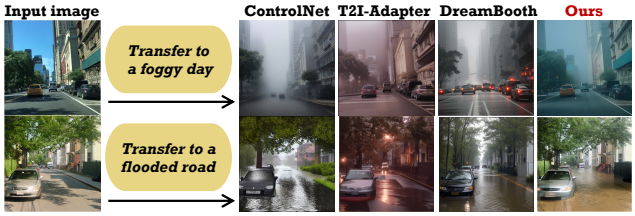

Figure 8: Qualitative comparisons to additional baselines.

achieved with a higher noise level, yet the content structure is damaged. Pix2Pix-zero exhibits the same behavior, with overall lower edit effects. Similarly, DiffuseIT has trouble preserving the content of the input. InstructPix2Pix and InstructDiff are greatly affected by the editing prompt and cannot better preserve the content information of the input, e.g., the trees beside the road in the input have changed in Fig. 7 (row 4). P2P cannot preserve the content of input well in some style transfer tasks due to the interference of text on the DDIM image reconstruction process. In contrast, our method performs DDIM inversion with an empty prompt, allowing us to use arbitrary guidance scales or prompts at generation. Additionally, To prove the effectiveness of our method in traffic scene style transfer, we expand our qualitative comparison to ControlNet (Zhang, Rao, and Agrawala 2023), T2I-Adapter (Mou et al. 2024), and DreamBooth (Ruiz et al. 2023). As shown in Fig. 8. These methods severely deviate from the content of the input image or result in noticeable visual artifacts.

Quantitative Analysis. Since ground truth is not available for text-based image editing, quantitative evaluation remains an open challenge. We present a user study in Tab. 2. 70 participants have rated a total of 50 images from the style fidelity, prompt fidelity, content fidelity, and whole quality. It can be observed that our method performs very well among these metrics. Besides, we numerically evaluate these results using seven metrics from three criteria: (i) whether the structure of the input is retained in the edited image. (ii) if the background content of the image stays unchanged. (iii) whether the edit is applied successfully. The structural consistency is measured using Structure Dist (Tumanyan et al. 2022). To ensure that we retain the content after edits, we calculate PSNR, LPIPS (Zhang et al. 2018), MSE, SSIM (Wang et al. 2004), and the CLIP image similarity. We measure the extent of the edit applied with the text-image CLIP Score, which calculates the percentage of instances where the edited image has a higher similarity to the target text, as measured by CLIP, than to the source text. Tab. 1 shows quantitative comparisons. Results show that our methods perform better in retaining content and structure while improving or maintaining editability compared with other methods.

Image Translation Method	Structure Distance ↓	Background Preservation PSNR ↑	LPIPS ↓	$MSE_{\times 10^4}$ ↓	$SSIM_{\times 10^2}$ ↑	CLIP Similarity↑ Image	Text-Image
Plug-and-Play	0.018	22.29	0.142	84.64	76.77	87.4%	88.9%
Prompt-to-Prompt	0.022	18.97	0.164	91.25	77.23	84.1%	88.1%
InstructPix2Pix	0.026	20.12	0.166	88.79	77.65	87.7%	90.5%
DiffuseIT	0.052	17.56	0.261	182.44	70.98	80.1%	86.7%
Pix2Pix-zero	0.016	20.22	0.155	144.24	77.89	90.6%	79.9%
SDEdit	0.032	20.12	0.182	98.32	76.56	87.9%	87.7%
InstructDiffusion	0.044	18.67	0.229	94.08	71.14	84.5%	86.1%
Ours (w/o ***Spatial Guidance***)	0.111	16.18	0.314	189.88	62.59	67.2%	87.8%
Ours (w/o ***Semantic Guidance***)	0.056	20.11	0.206	147.28	73.12	84.2%	89.9%
Ours (w/o ***Style Nursing***)	**0.011**	**26.56**	**0.114**	80.89	79.21	**91.7%**	88.3%
Ours (full)	0.014	25.28	0.126	**80.24**	**79.46**	89.7%	**92.6%**

Table 1: Quantitative comparison. We compare our method with other diffusion-based image translation methods in various metrics. We also conduct an ablation study where we remove different components of our method and observe the effects.

Metric	Plug-and-Play	DiffuseIT	Ours
Style Fidelity↑	0.149	0.092	**0.759**
Prompt Fidelity↑	0.189	0.117	**0.694**
Content Fidelity↑	0.235	0.108	**0.657**
Whole Quality↑	0.192	0.127	**0.681**

Table 2: Human Evaluations. The participants were asked to select the best editing result.

Perception Model	Testing Data	Object Detection Recall ↑ w/o ft	Recall ↑ w ft	mAP50 ↑ w/o ft	mAP50 ↑ w ft
Yolop	**Snowy**	75.8	**82.0**	60.8	**64.7**
	Rainy	60.5	**76.5**	44.1	**53.5**
	Foggy	52.4	**74.7**	36.5	**52.3**
HybridNets	**Snowy**	**88.7**	88.5	64.1	**64.7**
	Rainy	84.3	**85.3**	49.1	**50.8**
	Foggy	80.4	**83.5**	43.6	**49.9**

Table 3: Comparison of performance of perception models in extreme weather conditions before and after fine-tuning.

Ablation Study. We ablate our key design choices by evaluating our performance for the following cases: (i) w/o *Spatial Guidance*, (ii) w/o *Semantic Guidance*, (iii) w/o *Style Nursing*. The metrics are reported in Tab. 1 and a representative example is shown in Fig. 5. The results demonstrate that spatial features and attention maps are critical for preserving the input content, and *Style Nursing* effectively addresses the issue of *Style Neglection*.

Autonomous Driving Testing. To demonstrate the significance of our approach in autonomous driving testing, we conduct further study on downstream tasks with pre-trained SOTA perception models Yolop (Wu et al. 2022) and HybridNets (Vu, Ngo, and Phan 2022), which support basic perception tasks including object detection, lane segmentation and driving area segmentation. Specifically, we report

Figure 9: The performance of our method on general image translation. Input image from dataset *ImageNet*.

that poor visibility under extreme weather conditions (e.g., heavily snowy) indeed leads to a drastic decrease in the performance of perception models. For comparison shown in Tab. 3, we fine-tune the pre-trained models on images in extreme weather from *STREET-6K*, and the performance of all three tasks is significantly improved, which validates the effectiveness of our approach as a tailed weather data augmentation method in autonomous driving tests.

Discussion and Conclusion

We propose an image translation method for text-guided traffic scene style transfer, which can be generalized to general image translation tasks as Fig. 9 shows. We introduce the method of *Style Nursing* to address the phenomenon of *Style Neglection* and propose *Spatial and Semantic Guidance* to preserve the content structure of the input. Moreover, we present *STREET-6K* to address the lack of standardized benchmarks for image editing in traffic scenes. Our work demonstrates the balance of preserving the content of the input and achieving edit fidelity. We hope it will provide new insights for real-world content creation tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62202370. We would also like to extend our gratitude to Tingting Long, Jinghang Chen, Yuchen Li, and other students for their valuable discussions and insights during the submission process.

References

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *ACM Trans. Graph.*, 42(4).

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-or, D. 2023a. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

Gal, R.; Arar, M.; Atzmon, Y.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023b. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–13.

Geng, Z.; Yang, B.; Hang, T.; Li, C.; Gu, S.; Zhang, T.; Bao, J.; Zhang, Z.; Li, H.; Hu, H.; et al. 2024. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12709–12720.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-or, D. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations*.

Jeon, H.; Seo, J.; Kim, T.; Son, S.; Lee, J.; Choi, G.; and Lim, Y. 2023. RainSD: Rain Style Diversification Module for Image Synthesis Enhancement using Feature-Level Style Distribution. arXiv:2401.00460.

Kang, M.; Zhang, R.; Barnes, C.; Paris, S.; Kwak, S.; Park, J.; Shechtman, E.; Zhu, J.-Y.; and Park, T. 2024. Distilling Diffusion Models into Conditional GANs. In *European Conference on Computer Vision (ECCV)*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lee, J.; Shiotsuka, D.; Bang, G.; Endo, Y.; Nishimori, T.; Nakao, K.; and Kamijo, S. 2023. Day-to-night image translation via transfer learning to keep semantic information for driving simulator. *IATSS Research*, 47(2): 251–262.

Liao, Y.; Xie, J.; and Geiger, A. 2022. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3292–3310.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4296–4304.

Parmar, G.; Kumar Singh, K.; Zhang, R.; Li, Y.; Lu, J.; and Zhu, J.-Y. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Shechtman, E.; and Irani, M. 2007. Matching local self-similarities across images and videos. In *2007 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.

Tumanyan, N.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2022. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10748–10757.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. P+: Extended Textual Conditioning in Text-to-Image Generation. arXiv:2303.09522.

Vu, D.; Ngo, B.; and Phan, H. 2022. Hybridnets: End-to-end perception network. *arXiv preprint arXiv:2203.09035*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wei, X.-S.; Xu, Y.-Y.; Zhang, C.-L.; Xia, G.-S.; and Peng, Y. 2023. CAT: a coarse-to-fine attention tree for semantic change detection. *Visual Intelligence*, 1(1).

Wu, D.; Liao, M.-W.; Zhang, W.-T.; Wang, X.-G.; Bai, X.; Cheng, W.-Q.; and Liu, W.-Y. 2022. Yolop: You only look once for panoptic driving perception. *Machine Intelligence Research*, 19(6): 550–562.

Zhang, C.; Lin, Z.; Xu, L.; Li, Z.; Tang, W.; Liu, Y.; Meng, G.; Wang, L.; and Li, L. 2022. Density-Aware Haze Image Synthesis by Self-Supervised Content-Style Disentanglement. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4552–4572.

Zhang, C.; Ma, X.; Liu, Y.; Wang, L.; Su, Y.; and Liu, Y. 2024. Unified Regularity Measures for Sample-wise Learning and Generalization.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zheng, S.; Lu, C.; and Narasimhan, S. G. 2024. TPSeNCE: Towards Artifact-Free Realistic Rain Generation for Deraining and Object Detection in Rain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5394–5403.