ReSIT: A more Realistic Synthetic Driving Dataset for Multi-Domain Image-to-Image Translation

Anonymous CVPR submission

Paper ID 23

Abstract

Driving dataset is essential for success of autonomous driving system, yet collecting real-world data under diverse domains such as weather, time, and location is challenging and costly. This difficulty results in real-world driving datasets with restricted data domains. Although synthetic driving datasets have been introduced to address this issue, the diversity of domains they can cover remains limited. In this paper, we present ReSIT, a synthetic driving dataset built using a simulation platform that enables precise control over data collection conditions, resulting in more domains and possible combinations than existing datasets. Comparative analyses demonstrate that our dataset is more realistic than previous datasets. Additionally, we present a text-guided diffusion model tailored for multi-domain image-to-image translation, using an adapter for precise source image feature injection and guidance for effective translation. Experimental results show that our model outperforms existing models in preserving the structural content of source images during domain translation even in complex driving scenes. Our code and dataset will be released with the paper.

Figure 1. **Sample images from ReSIT** containing various domains such as weather, time of day, road marking status, road surface, and location.

1. Introduction

The success of autonomous driving depends on the advancement of various computer vision tasks, such as object detection, classification, and segmentation, which necessitate well-curated, large-scale datasets [19, 38]. However, collecting real-world driving data is challenging due to the high costs and laborious data annotation. Moreover, existing real-world datasets [5, 14, 17, 57, 74] have limited diversity in terms of geographical locations, weather conditions, and scene variations, which reduces the generalization performance of the trained model and limits its deployment in diverse environments.

To overcome these challenges, some studies have employed synthetic datasets generated in virtual environments [4, 16, 46, 51, 52, 63]. Those datasets have a great ability to control various domain conditions and provide perfect annotation labels for training machine learning models. However, current synthetic driving datasets are primarily designed for specific sub-tasks such as pedestrian detection [61] and weather classification [42]. Additionally, as shown in Tab. 1, they provide limited domain combinations that cannot reflect complex real-world [46]. As a result, models trained on these datasets often lack robustness when applied to more diverse or unexplored environments [72].

Although synthetic images aim to replicate real-world, there are inherent discrepancies in visual appearance, resulting in suboptimal model performance when directly trans-

	Dataset	Year	Domain	Possible Combination	Image Resolution	Total Frame	Annotations							Graphic Engine
							Sem. Seg.	Ins. Seg.	2D Det.	3D Det.	Depth	Line	Optical Flow	
Real World	KITTI [17]	2012	3	1	1382x512	7K			✓	✓	✓		✓	
	CityScapes [14]	2016	3	27	2048×1024	25K	✓	✓	✓	✓				
	BDD100K [74]	2018	3	18	1280×720	100K	✓	✓	✓			✓		
	INIT [57]	2019	1	4	1920×1208	155K			✓					
	nuScenes [5]	2020	3	12	1600×900	1.4M				✓				
Synthetic	VKITTI [16]	2016	2	6	1242×375	21K	✓	✓	✓	✓	✓		✓	UNITY
	GTA-V [51]	2016	2	4	1914×1052	25K	✓							GTA
	Synthia [54]	2016	3	18	1280×760	9K	✓		✓	✓				UNITY
	SynScapes [70]	2018			1440×720	25K	✓	✓	✓	✓	✓			Procedural Engine
	Apolloscape [27]	2018			1920×1080	273K	✓	✓	✓	✓	✓	✓		UNITY
	VKITTI2 [4]	2020	3	8	1242×375	21K	✓	✓	✓	✓	✓		✓	UNITY
	Shift [63]	2022	4	432	1280×800	2.5M	✓	✓	✓	✓	✓		✓	Carla
	CarlaScenes [33]	2022	2	16	1280×960		✓				✓	✓		Carla
	UrbanSyn [20]	2023			2048×1024	7.5K	✓	✓			✓			UNITY
	ReSIT(Ours)	2024	5	**5040**	1920×1080	300K	✓	✓	✓	✓	✓	✓	✓	UNITY

Table 1. **Comparison of driving datasets in terms of domain, size, and supported tasks.**

ferred to real-world tasks [64]. One reason for this discrepancy is the resolution of available 3D assets and the speed of data generation, which result in non-photorealistic images [20]. Pedestrians and vehicles provide a clear example of the visual gap inherent in this issue, along with background elements such as roads, weather, and time under various domain-specific conditions. While domain adaptation techniques [10, 34, 71] have been developed to bridge this gap, they have yet to offer a fundamental solution, emphasizing the need for more realistic synthetic data.

In this paper, we introduce **ReSIT**, a more realistic synthetic driving dataset for multi-domain image-to-image translation created using Cognata simulation platform [12], to address key limitations including image quality, domain diversity, and rare driving scenarios [18, 20, 38]. Compared to existing synthetic driving datasets [4, 16, 46, 51, 52, 63], ReSIT offers a greater number of domain variations—such as weather, time of day, road marking status, road surface, and location— and their possible combinations, thereby enabling greater data diversity in road scenes. Fig. 1 demonstrates examples of domain diversity in our dataset. Through a comprehensive analysis, our dataset exhibits its superiority in terms of diversity and realism.

To demonstrate the effectiveness of our dataset, we propose a text-guided diffusion model specifically designed for multi-domain image-to-image translation across the extensive range of combinations in our dataset. We incorporated an adapter in our model for accurate input image embedding, along with direct guidance for efficient domain translation. This method preserves dominant contents such as vehicles, pedestrians, and the overall road structure in driving scenes, modifying domain-specific semantics. By leveraging multi-domains in ReSIT, it enables a richer variety of scene generation. It achieves remarkable performance in multi-domain translation metrics, particularly outperforming existing methods in content preservation.

Our work offers the following key contributions:

- **Diverse multi-domain synthetic dataset**: We introduce a new synthetic driving dataset that provides a comprehensive collection of multi-domain combinations, encompassing diverse environmental variations.
- **Novel multi-domain image translation method**: We propose a new unpaired image-to-image translation model capable of handling multiple domain combinations simultaneously, while preserving the structural and contextual content of the source images.

2. Related Work

Real-World and Synthetic Driving Datasets Following the introduction of the KITTI dataset [17], prominent datasets such as Cityscapes [14] and BDD100K [74] have been key contributors to the development of models for tasks like object detection [15], semantic segmentation [62], and depth estimation [3] in self-driving vehicles. However, real-world datasets have shortcomings, making it difficult to encompass the diverse environmental changes in driving scenarios. Furthermore, annotations such as segmentation, depth estimation, lane detection, and optical flow –which require extensive manual labeling– are either unavailable or provided in small quantities [76]. The nuScenes dataset [5], which employs Lidar sensors to provide 1.4 million images with 3D bounding box annotations, has recently been introduced. While nuScenes plays an important role in 3D detection, it offers data from only two cities, and annotations are largely focused on Lidar-based 3D detection.

In response to the challenges faced by real-world datasets, several synthetic datasets have been proposed, starting with VKITTI [4, 16]. The advancement of 3D graphics engines has enabled the creation of various synthetic datasets offering the advantage of providing perfect labels for every frame. However, existing synthetic datasets generate limited objects and biased environments, failing to capture the full diversity of real-world driving scenarios [20]. Datasets like GTA [51] and VIPER [52] face constraints in representing diverse driving domains due to the inherent limitations of the game engine. Although, the re-

cent SHIFT [63] provides 2.5 million images across 8 cities with a variety of domain categories, the domain combinations are not sufficient.

Domain Gap Traditional machine learning approaches assume that the training and test data share the same underlying distribution. When this assumption is violated, the model performance can significantly degrade [48]. In particular, there is a substantial domain gap between synthetic datasets and real-world datasets [56], which hinders the direct application of synthetic data in practical scenarios. Transfer learning techniques, particularly domain adaptation, have been widely explored to bridge this domain gap [2, 7, 34]. Recent studies have investigated advanced strategies such as multi-domain adaptation, domain mixing, and adversarial training to tackle these challenges [1, 10, 71]. While these approaches offer some improvements, they still struggle with the growing complexity of domain combinations, highlighting the need for more robust solutions that can handle diverse driving scenarios.

Multi-domain I2I Translation This enables transformations across multiple visual domains for several applications like style transfer and data augmentation. Early models like StarGAN [11] introduced a unified framework to translate images across multiple domains using a single model, eliminating the need for separate models for each domain pair. Subsequent methods such as MUNIT [28] and DRIT [37] further advanced image translation by enabling multi-modal outputs, allowing for varied translations within each domain and enhancing model flexibility. Recent studies have advanced translation quality using diffusion-based approaches with text-guided methods [32, 35, 55]. Additionally, techniques that invert input images into noise [31, 41, 60] and selectively translate specific areas through attention control [6, 22, 66] have proven highly effective. However, challenges remain in achieving a balance between preserving contextual content and allowing flexible domain translation, especially when dealing with diverse domains.

3. Dataset Overview

3.1. Motivation and Background

Most existing driving datasets have limited complexity in domain combinations. This results in a lack of scene diversity and difficulties in reflecting the wide range of real-world environmental conditions, as indicated by the findings of previous studies [38–40]. This insufficiency restricts the capacity of deep learning models to learn from the complex conditions that can be encountered in the real world. Domains such as status of road marking wear and surface conditions are frequently excluded from consideration, making it challenging for models to adapt to real-world environments [20]. To address this limitation, we generated

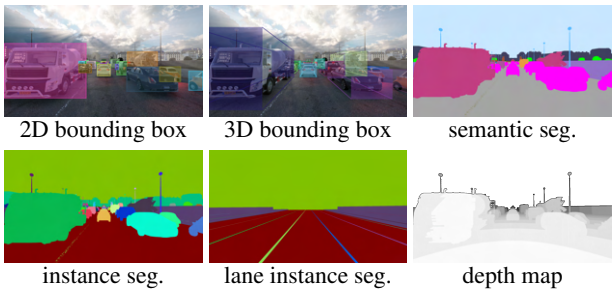
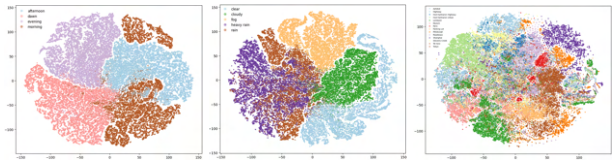


Figure 2. **Annotations in ReSIT.** Comprehensive annotations are provided for all source images.



(a) Time of Day (b) Weather (c) Location

Figure 3. **t-SNE visualization for domain categories in ReSIT.**

the **ReSIT** synthetic driving dataset using the Cognata Simulation Platform [12], which built upon the Unity Graphics Engine. The ReSIT dataset systematically combines multiple domain elements to represent complex real-world driving conditions, allowing for comprehensive model evaluation across diverse scenarios and enhancing the reliability of autonomous driving tasks.

3.2. Dataset Generation Strategy

The primary contribution of our dataset is amplifying domain combinations to closely create and simulate a more diverse real-world driving environment. Unique scenarios can be generated through numerous combinations across the following domain categories:

- **Time of Day**: Various times of day represent sunlight and brightness changes, such as morning, dawn, and evening.
- **Weather**: Combining various climatic variables such as cloudy, rain, and fog enables us to create complex and unpredictable environmental conditions.
- **Location**: Our Dataset reflects geographic characteristics across 14 diverse regions, including various regions of America, Europe, Asia, and the Middle East, ensuring comprehensive representation beyond specific locales.
- **Road Surface**: Diverse road surface conditions are incorporated, including wet roads, puddles, and snow-covered surfaces, to simulate real-world road conditions.
- **Road Marking Status**: Road type is diversified with levels of degradation such as wear, no wear, and faded markings reflecting the variability of road conditions.

As shown in Tab. 1, our dataset encompasses a broad range of domain categories, generating a total of 5,040 unique domain combinations – a level of diversity surpassing that of typical datasets. These domain combinations are

Dataset	Type	Object Detection (mAP)			Semantic Segmentation (mIoU)		
		Total	Car	Person	Total	Car	Person
KITTI [17]	real	0.29	0.35	0.22	-	-	-
INIT [57]	real	0.38	0.54	0.23	-	-	-
BDD100K [74]	real	0.31	0.40	0.23	88.75	97.56	79.93
SHIFT [63]	synthetic	0.29	0.35	0.22	91.86	98.96	84.76
ReSIT (Ours)	synthetic	0.32	0.37	0.27	95.47	99.67	91.26

Table 2. **In-dataset evaluation results.**

Train	Test	Object Detection (mAP)			Semantic Segmentation (mIoU)		
		Total	Car	Person	Total	Car	Person
SHIFT	KITTI	0.20	0.25	0.19	-	-	-
ReSIT (Ours)	KITTI	**0.24**	**0.28**	**0.19**	-	-	-
SHIFT	INIT	0.08	0.13	0.04	-	-	-
ReSIT (Ours)	INIT	**0.14**	**0.16**	**0.12**	-	-	-
SHIFT	BDD100k	0.12	0.14	0.10	59.43	88.19	30.67
ReSIT (Ours)	BDD100k	**0.18**	**0.23**	**0.14**	**69.53**	**92.63**	**46.42**

Table 3. **Cross-dataset evaluation results.**

meticulously considered to capture a wide range of driving conditions, which helps minimize data bias. The dataset is comprised of 300,000 images in total. For further details on the distribution and proportions of each domain category, please refer to the supplementary material.

3.3. Dataset Design

- **Camera Specifications and Frame Rate**: Our dataset uses a front camera with a 100-degree field of view (FOV). Similar to other datasets [63, 74], our dataset was created from continuous scenarios by extracting 1 FPS from 50-second videos, resulting in 50 frames per scenario and balancing continuity with inter-frame variation.
- **Resolution and Compatibility**: The dataset has 1920 $\times$ 1080 resolution, capturing extensive details for object recognition while maintaining compatibility with other datasets.
- **Annotations**: Our dataset provides a richer set of annotations, such as bounding box, segmentation, depth, optical flow, supporting a wide range of computer vision tasks. Figure 2 visualizes the provided annotations.

4. Dataset Analysis

4.1. t-SNE Visualization of domain categories

To validate that the classes within each domain in our dataset are clearly distinguished, we trained a ResNet152 [21] classification model from scratch and extracted feature embeddings for visualization using t-SNE [67]. Figure 3 presents the t-SNE plots for key domains in our dataset: time of day, weather condition, and location. The results show that the images for each class form distinct clusters, even though diverse domain conditions can introduce substantial variations. While a single class may appear as multiple sub-clusters due to different domain factors, these sub-clusters are remain well-separated from other classes within the same domain. This demonstrates that our dataset captures distinguishable image-level features for each class across varying domain conditions. Therefore, our dataset provides reliable and robust representations that enable accurate image classification across multiple domain combinations.

4.2. Dataset Evaluation

To evaluate the applicability and realism of the proposed synthetic dataset, we conducted two key experiments. The first experiment focuses on in-dataset evaluation, where models are trained and tested on the same dataset to assess the dataset's ability to support effective learning for driving scene understanding. The second experiment examines cross-dataset evaluation, where models trained on synthetic datasets are tested on a real-world dataset to assess generalization performance and the realism of the synthetic data.

4.2.1. Evaluation Setup

- **Model**: We employed Faster R-CNN [50] for object detection and DeepLab v3 [9] for semantic segmentation.
- **Dataset Configuration**: We used KITTI [17], BDD100K [74], INIT [57], SHIFT [63], and our dataset for training. Testing was conducted within the same dataset (in-dataset Evaluation) and on KITTI, BDD100K, INIT (cross-dataset Evaluation). For all experiments, the number of training images was limited to 50000, and the number of test images was set to 5000. In the case of segmentation experiments, due to limited labels in BDD10k, we used 7000 training images and 1000 test images. All images were resized to have a 640 width while maintaining the original aspect ratio.
- **Class Selection**: We focused on two common classes: car, and pedestrian, across all experiments.
- **Hyperparameter Settings**: We followed the default hyperparameter settings of MMDetection [8] and MMSegmentation [13] to ensure fair baseline comparisons.

4.2.2. In-Dataset Evaluation

The in-dataset evaluation aims to compare the performance of models trained and tested on the same dataset across five datasets: KITTI, BDD100K, INIT, SHIFT, and Ours. This experiment is intended to validate that our dataset can effectively serve as training data for vision tasks in autonomous driving (e.g., Object Detection, Semantic Segmentation), similar to both existing real and synthetic datasets. As shown in Tab. 2, models trained and tested on each dataset achieved comparable performance. These results demonstrate that our synthetic dataset provides sufficient learning capability for vision tasks, similar to well-established real (KITTI, BDD100K, INIT) and synthetic (SHIFT) datasets. This finding emphasizes that our dataset effectively captures key features of driving scenes, making it a viable alternative or complement to existing datasets.

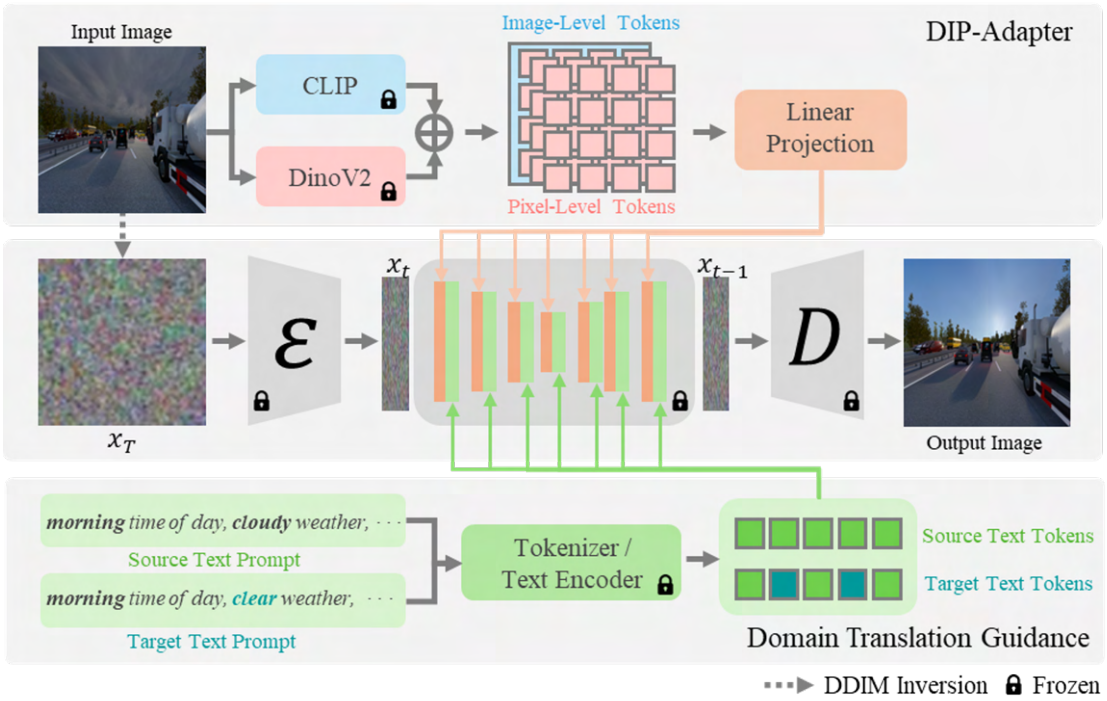


Figure 4. **Pipeline of our method for multi-domain I2I translation.** DIP-Adapter projects both image-level and pixel-level tokens from the input image into the U-Net to preserve important structures in complex driving scenes. Domain Translation Guidance utilizes the difference between the latent vectors obtained from the source and target text prompts to enable efficient and direct domain translation.

4.2.3. Cross-Dataset Evaluation

The cross-dataset evaluation assesses the generalization capability of models trained on synthetic datasets (SHIFT and Ours) by applying them to real-world datasets (KITTI, BDD100K, INIT). The primary objective is to validate whether the proposed synthetic dataset offers greater realism and better captures diverse driving scenarios compared to SHIFT, thereby leading to improved performance in real-world tasks. According to Tab. 3, models trained on our synthetic dataset outperforms models trained on SHIFT when tested on real-world datasets. This indicates that our proposed dataset can cover a broader range of driving scenarios and provides more realistic representations of real-world driving conditions. The superior cross-dataset performance of our dataset suggests that it is better suited for training models with enhanced generalization capabilities, making it a more effective choice for real-world applications.

5. Proposed Method

5.1. Preliminaries

Diffusion models [26, 53, 59] are probabilistic generative models that iteratively denoise random noise to approximate a target image. By employing a specialized loss function that minimizes the difference between target noise ϵ and predicted noise, these models accurately learn the transformation from noise to image:

$$\mathcal{L} = \mathbb{E}_{x_0,\epsilon\sim\mathcal{N}(0,I),c,t}\|\epsilon - \epsilon_\theta(x_t, c, t)\|^2, \quad (1)$$

where $\epsilon_\theta(\cdot)$ represents the noise predicted by the model with image data x, additional condition c, and t denotes the current time step in the diffusion process.

In multi-domain unpaired image-to-image translation, it is essential to preserve the structural and contextual content of the source image while modifying only the necessary domain-specific attributes [35]. Particularly for driving scenes, which contain both large objects such as pedestrians and vehicles and semantically important smaller elements like traffic lights and signs, it is important to maintain the structural content of these elements while translating domain-specific attributes, such as weather or location.

Existing models [6, 22, 31, 41, 43, 60, 66] often invert the input image into noise to retain as much image content as possible and then apply attention map control for selective editing of domain-specific attributes. While effective across general datasets with few primary objects, these approaches face limitations in preserving critical structures within complex driving scenes. Therefore, we propose the first text-guided diffusion model, tailored for multi-domain image-to-image translation using an adapter. Our model uses the adapter to embed the input image more precisely and offers direct guidance for effective domain translation.

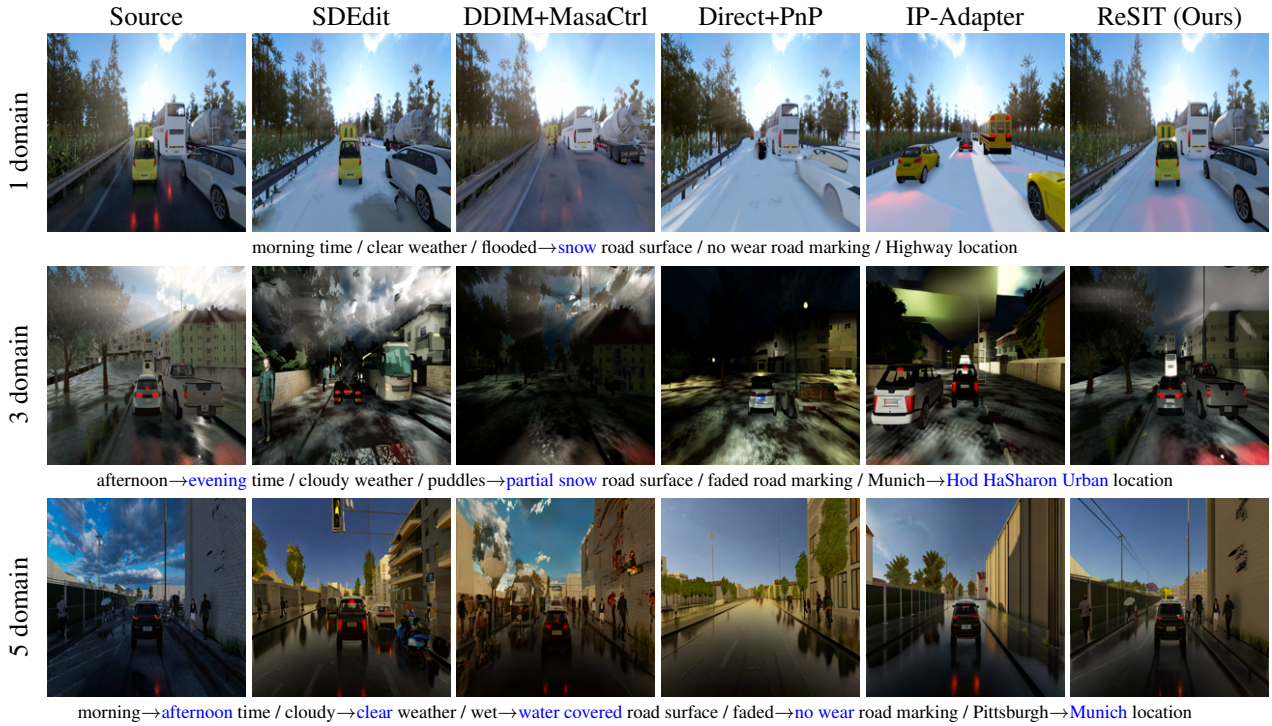


Figure 5. **Qualitative evaluation for multi-domain image-to-image translation methods.**

5.2. Dense Image Prompt Adapter

Adapters [44, 73] have frequently been employed in diffusion models to incorporate image prompts alongside text prompts. By training only the projection network of the adapter that injects extracted features, while keeping the image encoder and diffusion model frozen, the model can efficiently integrate visual information without additional training. The loss function with the text prompt c_t, additional image prompt c_i is represented as follows:

$$\mathcal{L} = \mathbb{E}_{x_0,\epsilon\sim\mathcal{N}(0,I),c_t,c_i,t}\|\epsilon - \epsilon_\theta(x_t, c_t, c_i, t)\|^2. \quad (2)$$

To embed visual information, CLIP [49] image encoder enables efficient semantic image representation by embedding image and text features into a shared latent space. However, due to global supervision based on image captions, this image encoder has limitations in learning detailed pixel-level information such as color and position, making it less effective in capturing fine-grained pixel details [30]. DINOv2 [47], trained using self-supervised learning on image data alone, can encode fine-grained pixel-level representations. Therefore, as shown in Fig. 4, we designed Dense Image Prompt (DIP) Adapter, which concatenates image-level tokens from CLIP and pixel-level tokens from DINOv2 to produce semantically rich and accurate image embeddings. This method demonstrates improved retention of detailed image features, outperforming the IP-Adapter, which relies solely on image-level token embeddings.

5.3. Domain Translation Guidance

Classifier-free guidance (CFG) [25] is a technique in diffusion models that enables conditional control in image generation without a separate classifier model. The diffusion model is trained simultaneously on both conditional and unconditional setting, and during sampling steps, it utilizes a guidance scale w to adjust the strength of conditioning by combining the conditional prediction $\epsilon_\theta(x_t, c, t)$ and unconditional prediction $\epsilon_\theta(x_t, t)$. This approach allows for flexible control over the generated image's characteristics, as expressed in the following equation:

$$\hat{\epsilon}_\theta(x_t, c, t) = w\epsilon_\theta(x_t, c, t) + (1 - w)\epsilon_\theta(x_t, t). \quad (3)$$

CFG has proven successful in conditional diffusion models [45, 53], particularly in high-quality image sampling. However, this guidance is designed for tasks with only a single prompt type, making it inefficient for translation tasks where both a source image and text prompt are provided. Although the text prompt shifts from source to target, the injected source image retains source domain attributes. Consequently, adjusting the guidance scale affects both the target prompt and the source attributes, limiting efficient control over the domain transition. We introduce domain translation guidance (DTG), which directly guides the gap between the source and target domain based on the source image. The degree of translation can also be appropriately

Number of translations	Methods	FID↓	FID_{clip}↓	Structure Distance↓	Background Preservation				CLIP Similarity↑
					PSNR↑	LPIPS↓	MES↓	SSIM↑	
1 domain	SDEdit [41]	47.60	3.41	0.0642	17.71	0.30	0.0202	0.51	29.48
	DDIM [60]+PnP [66]	54.19	5.02	0.0889	16.67	0.31	0.0296	0.51	29.62
	DDIM [60]+MasaCtrl [6]	53.66	5.81	0.0801	17.40	0.28	0.0232	0.52	27.63
	Direct [31]+PnP [66]	54.22	4.93	0.0881	16.66	0.31	0.0297	0.51	**29.70**
	Direct [31]+MasaCtrl [6]	49.74	4.88	0.0774	17.56	0.27	0.0227	0.52	28.04
	IP-Adapter [73]	64.23	6.61	0.1074	14.81	0.41	0.0422	0.44	28.46
	ReSIT (Ours)	**36.06**	**2.29**	**0.0326**	**21.18**	**0.14**	**0.0129**	**0.67**	28.93
3 domains	SDEdit [41]	56.05	5.11	0.0959	15.54	0.41	**0.0321**	0.43	**29.79**
	DDIM [60]+PnP [66]	60.05	6.20	0.1501	12.95	0.43	0.0667	0.41	29.59
	DDIM [60]+MasaCtrl [6]	58.53	7.43	0.1121	14.76	0.36	0.0443	0.44	26.28
	Direct [31]+PnP [66]	60.07	6.02	0.1483	12.92	0.43	0.0672	0.41	29.63
	Direct [31]+MasaCtrl [6]	55.46	6.30	0.1105	14.84	0.36	0.0438	0.45	26.66
	IP-Adapter [73]	67.50	7.06	0.1392	12.90	0.47	0.0643	0.39	28.53
	ReSIT (Ours)	**42.87**	**2.87**	**0.0737**	**16.31**	**0.23**	0.0345	**0.56**	28.28
5 domains	SDEdit [41]	61.16	6.27	0.1071	**15.12**	0.45	**0.0344**	0.40	28.52
	DDIM [60]+PnP [66]	60.52	6.35	0.1835	11.60	0.49	0.0914	0.36	30.04
	DDIM [60]+MasaCtrl [6]	60.26	8.24	0.1254	13.80	0.40	0.0554	0.41	25.85
	Direct [31]+PnP [66]	59.78	6.27	0.1829	11.58	0.49	0.0919	0.36	**30.08**
	Direct [31]+MasaCtrl [6]	56.67	7.47	0.1234	13.86	0.40	0.0547	0.41	26.22
	IP-Adapter [73]	67.37	7.39	0.1612	12.01	0.50	0.0793	0.35	28.71
	ReSIT (Ours)	**45.39**	**3.17**	**0.1012**	14.41	**0.28**	0.0516	**0.50**	28.21

Table 4. **Quantitative evaluation for multi-domain image-to-image translation methods.** The best results are highlighted in bold, the second best results are marked with an underline.

controlled through the translation scale s, as shown below:

$$\hat{\epsilon}_\theta(x_t, \hat{r}, c_i, t) = \epsilon_\theta(x_t, r, c_i, t) + s\left\{\epsilon_\theta(x_t, \hat{r}, c_i, t) - \epsilon_\theta(x_t, r, c_i, t)\right\}, \quad (4)$$

where r refers to the source text prompt (*e.g.*cloudy weather), and $\hat{r}$ indicates the target text prompt (*e.g.*clear weather). Additionally, DTG technique allows for the use of different scale values $s_1, s_2, ..., s_D$ across D domains, enabling differential scaling for each domain:

$$\hat{\epsilon}_\theta(x_t, \hat{r}, c_i, t) = \epsilon_\theta(x_t, r, c_i, t) + \sum_{d=1}^{D} s_d\left\{\epsilon_\theta(x_t, \hat{r}_d, c_i, t) - \epsilon_\theta(x_t, r, c_i, t)\right\}. \quad (5)$$

The application of differential DTG enhances the flexibility and practicality of multi-domain translation, making it more intuitive. We demonstrate this effect through additional experiments in the supplementary material.

6. Experiments

In our experiments, we fine-tuned Stable Diffusion v2.1 base model on our dataset using two NVIDIA A40 GPUs with a batch size of 16 per GPU over 300,000 steps. Input images were resized to 512 × 512 to match the pre-trained model's size, then encoded into latents with a (4, 64, 64) shape via a VAE. The learning rate was fixed at 1e-05 throughout training. In addition, the adapters also were trained with a batch size of 24 per GPU over 300,000 steps, with OpenCLIP ViT-H/14 [29] and DINOv2-large [47] used as image encoders. For compatibility with baseline models, we utilized the HuggingFace diffusers [68] library for diffusion models in our experiments.

We also performed additional ablation studies for our method, detailed in the supplementary material.

6.1. Comparisons with Existing Methods

We conducted comparative experiments with existing multi-domain I2I translation models to evaluate the effectiveness of our method. For baseline models, we included SDEdit [41], DDIM, Direct Inversion [31, 60] with editing methods MasaCtrl [6], Plug-and-Play [66], and IP-Adapter [73] to compare performance across a diverse range of models. Leveraging the multiple domain characteristics of our dataset, we conducted model performance evaluations by adjusting the number of translated domains. From 600 validation scenarios, we randomly selected one image and applied translations across a randomly chosen set of 1, 3, or 5 domains.

Qualitative evaluation was performed on four popular criteria: image quality (FID [24], FID_{clip} [36]), structure distance [65], background preservation (PSNR, LPIPS [75], MES, SSIM [69]), translation quality (CLIP Similarity [23]). Table 4 shows the results, indicating that our method achieved state-of-the-art (SOTA) performance in most metrics, though not all. This suggests that our model effectively preserves the structural and contextual content of

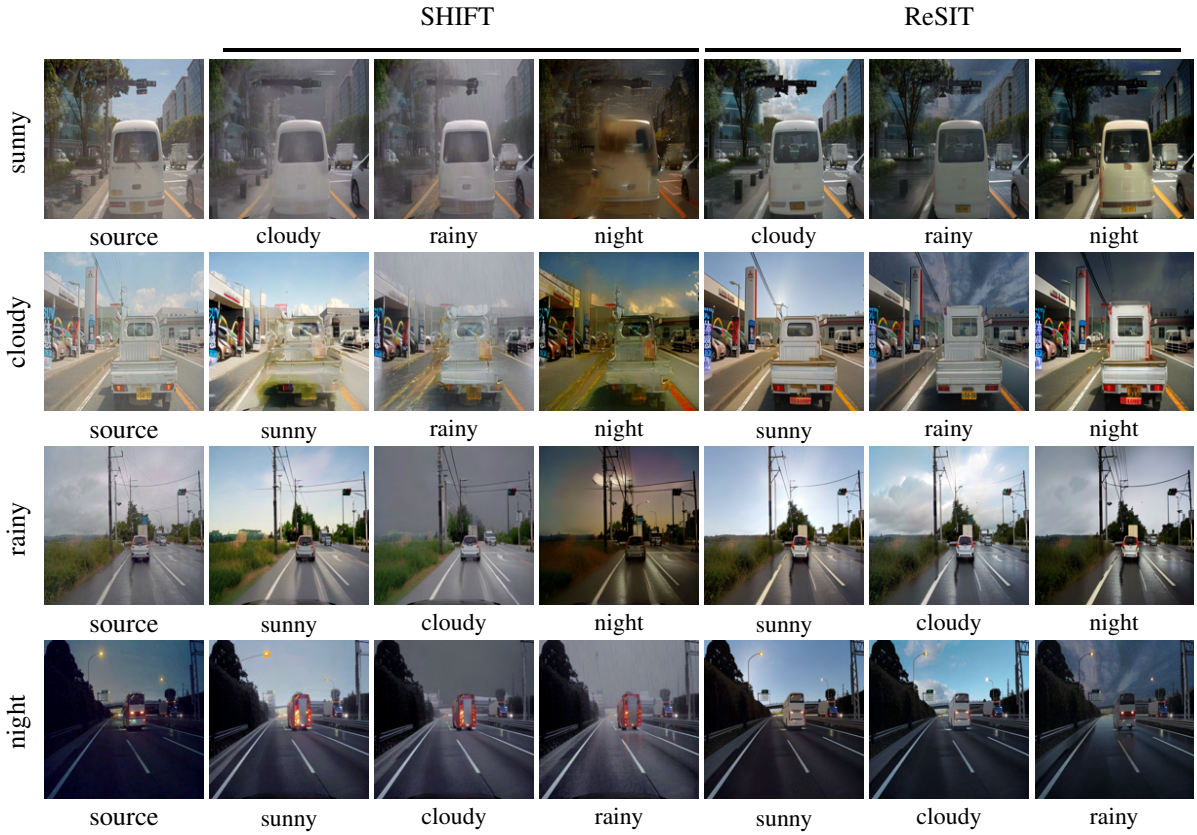


Figure 6. **Image translation results of adaptation to real-world dataset using our method trained synthetic datasets.**

the source image during translation. Although our method did not achieve SOTA, especially in CLIP Similarity, due to its emphasis on structural preservation, the results remained competitive, as depicted in Fig. 5

6.2. Adaptation to Real-World Scenarios

In the previous section, we demonstrated that our dataset contains diverse domains and that our translation method effectively preserves the structural and background information of the source image. As a next step, we tested the applicability of our method on real-world data to determine its practical viability. Applying models trained solely on synthetic data to real-world scene presents a significant challenge due to the domain gap, often resulting in performance degradation. However, our translation model trained on realistic ReSIT dataset, we achieved promising results.

As shown in Tab. 5, model trained on ReSIT achieved better FID scores than trained on SHIFT [63] for weather transformation tasks on INIT [57] real-world data, demonstrating enhanced alignment with real-world conditions. Figure 6 provides visual examples supporting these findings. Overall, this experiment demonstrates that realistic synthetic data, when paired with high-quality vision task models, has the potential to be effectively applied to real-world.

Source domain	Train Dataset	Translated domain			
		sunny	cloudy	rainy	night
sunny	SHIFT	-	103.67	117.56	112.95
	ReSIT		**92.55**	**102.62**	**94.52**
cloudy	SHIFT	97.82	-	105.68	105.68
	ReSIT	**94.21**		**98.82**	**93.68**
rainy	SHIFT	93.70	91.28	-	104.07
	ReSIT	**89.85**	**88.00**		**91.81**
night	SHIFT	114.32	110.75	115.58	-
	ReSIT	**110.34**	**108.31**	**111.17**	

Table 5. **FID score of image translation results.**

7. Conclusion

In this paper, we present **ReSIT**, a more realistic synthetic driving dataset specifically designed to support multi-domain image-to-image translation. Our dataset offers a significantly broader range of domain combinations than existing datasets. We also propose a novel text-guided diffusion model tailored for multi-domain I2I translation trained on our new dataset, achieving outstanding performance by preserving the structural content of source images in complex driving scenes. For future work, we will expand the applicability of ReSIT across various vision tasks, including joint training with real-world datasets.

References

[1] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021. 3

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010. 3

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4009–4018, 2021. 2

[4] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 1, 2

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2

[6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023. 3, 5, 7

[7] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 627–636, 2019. 3

[8] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 4

[9] Liang-Chieh Chen. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 4

[10] Shuaijun Chen, Xu Jia, Jianzhong He, Yongjie Shi, and Jianzhuang Liu. Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11018–11027, 2021. 2, 3

[11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 3

[12] Cognata Ltd. Cognata - autonomous vehicle simulation platform, 2024.11. https://www.cognata.com/. 2, 3

[13] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 4

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2

[15] Di Feng, Ali Harakeh, Steven L Waslander, and Klaus Dietmayer. A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):9961–9980, 2021. 2

[16] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016. 1, 2

[17] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 2, 4

[18] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3): 362–386, 2019. 2

[19] Junyao Guo, Unmesh Kurup, and Mohak Shah. Is it safe to drive? an overview of factors, metrics, and datasets for driveability assessment in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 21(8):3135–3151, 2019. 1

[20] Jose L. Gómez, Manuel Silva, Antonio Seoane, Agnès Borrás, Mario Noriega, Germán Ros, Jose A. Iglesias-Guitian, and Antonio M. López. All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes. *arXiv preprint arXiv:2312.12176*, 2023. 2, 3

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[22] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3, 5

[23] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 7

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7, 1

[25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6, 1

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[27] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 954–960, 2018. 2

[28] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 3

[29] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 7

[30] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023. 6

[31] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023. 3, 5, 7

[32] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 3

[33] Andreas Kloukiniotis, Andreas Papandreou, Christos Anagnostopoulos, Aris Lalos, Petros Kapsalas, Duong-Van Nguyen, and Konstantinos Moustakas. Carlascenes: A synthetic dataset for odometry in autonomous driving. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4520–4528, 2022. 2

[34] Wouter M Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018. 2, 3

[35] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022. 3, 5, 2

[36] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. *arXiv preprint arXiv:2203.06026*, 2022. 7, 1

[37] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 3

[38] Mingyu Liu, Ekim Yurtsever, Jonathan Fossaert, Xingcheng Zhou, Walter Zimmer, Yuning Cui, Bare Luka Zagar, and Alois C. Knoll. A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook. *arXiv preprint arXiv:2401.01454*, 2024. 1, 2, 3

[39] Aboli Marathe, Deva Ramanan, Rahee Walambe, and Ketan Kotecha. Wedge: A multi-weather autonomous driving dataset built from generative vision-language models. *arXiv preprint arXiv:2305.07528*, 2023.

[40] Aboli Marathe, Rahee Walambe, and Ketan Kotecha. In rain or shine: Understanding and overcoming dataset bias for improving robustness against weather corruptions for autonomous vehicles. *arXiv preprint arXiv:2204.01062*, 2023. 3

[41] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3, 5, 7

[42] Saad Minhas, Zeba Khanam, Shoaib Ehsan, Klaus McDonald-Maier, and Aura Hernández-Sabaté. Weather classification by utilizing synthetic data. *Sensors*, 22(9): 3193, 2022. 1

[43] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 5

[44] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 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*, pages 4296–4304, 2024. 6

[45] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 6

[46] Joshua Niemeijer, Sudhanshu Mittal, and Thomas Brox. Synthetic dataset acquisition for a specific target domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4055–4064, 2023. 1, 2

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6, 7, 1

[48] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015. 3

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6, 1

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 4

[51] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 1, 2

[52] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Proceedings of the IEEE international conference on computer vision*, pages 2213–2222, 2017. 1, 2

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 5, 6

[54] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 2

[55] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 3

[56] Manuel Schwonberg, Joshua Niemeijer, Jan-Aike Termöhlen, Nico M Schmidt, Hanno Gottschalk, Tim Fingscheidt, et al. Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving. *IEEE Access*, 11:54296–54336, 2023. 3

[57] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas S Huang. Towards instance-level image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3683–3692, 2019. 1, 2, 4, 8

[58] Zhangzhang Si and Song-Chun Zhu. Learning hybrid image templates (hit) by information projection. *IEEE Transactions on pattern analysis and machine intelligence*, 34(7):1354–1367, 2011. 2

[59] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 5

[60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 5, 7, 1

[61] Thomas Stauner, Frederik Blank, Michael Fürst, Johannes Günther, Korbinian Hagn, Philipp Heidenreich, Markus Huber, Bastian Knerr, Thomas Schulik, and Karl-Ferdinand Leiß. Synpeds: A synthetic dataset for pedestrian detection in urban traffic scenes. In *Proceedings of the 6th ACM Computer Science in Cars Symposium*, pages 1–10, 2022. 1

[62] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 2

[63] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382, 2022. 1, 2, 3, 4, 8

[64] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 969–977, 2018. 2

[65] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 7

[66] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 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*, pages 1921–1930, 2023. 3, 5, 7

[67] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 4

[68] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022. 7

[69] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[70] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018. 2

[71] Rongchang Xie, Fei Yu, Jiachao Wang, Yizhou Wang, and Li Zhang. Multi-level domain adaptive learning for cross-domain detection. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 2, 3

[72] Xingxu Yao, Sicheng Zhao, Pengfei Xu, and Jufeng Yang. Multi-source domain adaptation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3273–3282, 2021. 1

[73] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 6, 7, 1

[74] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1, 2, 4

[75] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[76] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021. 2

ReSIT: A more Realistic Synthetic Driving Dataset for Multi-Domain Image-to-Image Translation

Supplementary Material

A. More Details about ReSIT Dataset

In this section, we present more details about our dataset.

A.1. Domain Category and Example

Figure 7 illustrates the list of categories provided for the five domains in our dataset. By combining these categories, the dataset can generate a total of 5,040 unique domain combinations. Figure 9 present sample images corresponding to the domain categories. In addition to the t-SNE visualization results presented in Sec. 4.1, the sample images also demonstrate that the dataset exhibits distinguishable features for each category.

A.2. Domain-wise Category Distribution

To demonstrate that our dataset is consistently generated across all domain categories, we measured the category-wise image distribution ratios of the generated dataset for each domain and visualized the results in a graph. As shown in Fig. 10, the images for all categories within each location are well-distributed.

A.3. Object Statistics

Our dataset contains a sufficient number of key objects commonly encountered in driving environments. We counted the occurrence frequency of various objects, such as cars, buses, pedestrians, and trucks, across all images in the dataset. As shown in Fig. 8, the dataset includes a diverse range of objects in quantities adequate for various vision tasks.

B. Ablation Study

We conducted detailed ablation studies to empirically validate the effectiveness of each component in our proposed method. The IP-Adapter [73], which supports both text and image prompts as input, was used as the baseline for these evaluations. Performance changes were measured as components were sequentially incorporated. The evaluations followed the protocol described in Sec. 6.1, with results averaged over 1, 3, and 5 multi-domain translation tasks.

Initially, we enhanced the baseline adapter by fully incorporating pixel-level tokens, in addition to the original image-level tokens, through the CLIP [49] image encoder. This modification significantly enhanced both FID [24] and FID_{clip} [36] scores. Replacing the CLIP with DINOv2 [47] and subsequently adopting the DIP-Adapter led to further quantitative improvements, particularly in the retention of pixel-level details such as color, as illustrated in Fig. 11. Furthermore, replacing Classifier-free Guidance (CFG) [25] with Domain Translation Guidance (DTG) resulted in substantial performance gains across various metrics, including FID. Finally, incorporating DDIM Inversion [60] enabled precise computation of initial noise, which significantly enhanced the preservation of structural and contextual content in the source images. A comprehensive overview of the performance variations introduced by each component is presented in Tab. 6

C. Variations in Image Translation

In this section, we present a detailed exploration of variations in image translation enabled by our method. Using the BDD100k [74] dataset, we trained our method and demonstrated translation experiments across the weather and time of day domains.

C.1. Translation Scale

In Sec. 5.3, we demonstrated that the degree of translation can be effectively controlled by adjusting the translation scale s. Figure 12 showcases the visual variations in multi-domain image-to-image translation applied to the source image as guided by the target, based on different values of s. The results indicate that higher translation scale values lead to stronger guidance to the target, and the appropriate scale value can be depending based on the dataset and the domain.

C.2. Differential Translation Scale

As previously shown, the translation scale provides an effective mechanism for controlling the degree of translation, but there are situations where finer granularity is required. In multi-domain i2i translation, where translations across multiple domains are performed simultaneously, the inability to adjust the translation intensity for each domain can significantly limit the method's usability. To overcome this limitation, we propose the use of differential translation scales, which allow flexible and domain-specific adjustments. This method is both straightforward and effective. Figure 13 depicts the visual outcomes of image translation under varying s_1 and s_2. While this example divides the scale into two, as demonstrated in Sec. 5.3, more detailed divisions can be applied to achieve significantly diverse variations in translation.

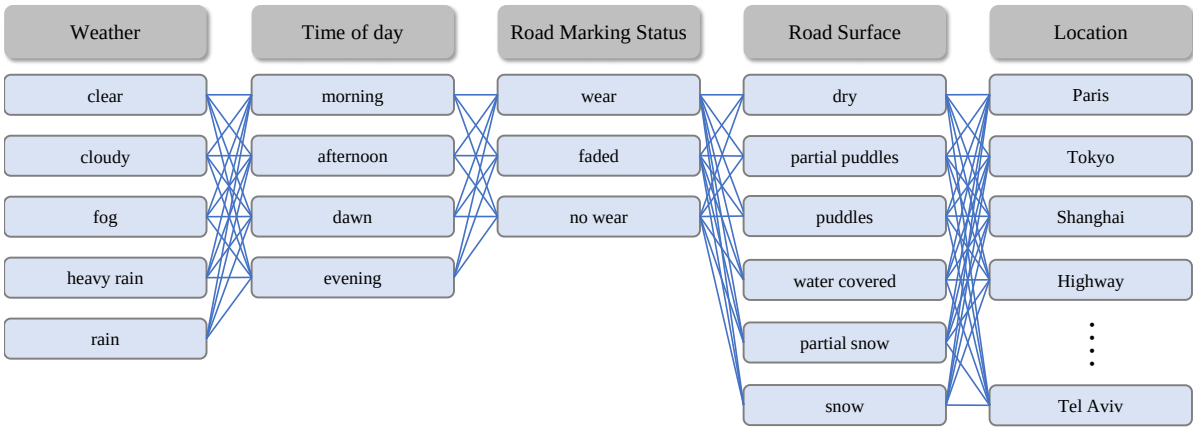


Figure 7. **Possible combinations of categories for each domain:** For location, the following categories are excluded to maintain the visual clarity of the figure: Roadways, Munich, Pittsburgh, Parking Lot, Ashdod, Hod HaSharon Highway, Hod HaSharon Urban, Lombard, and Stevenes Creek.

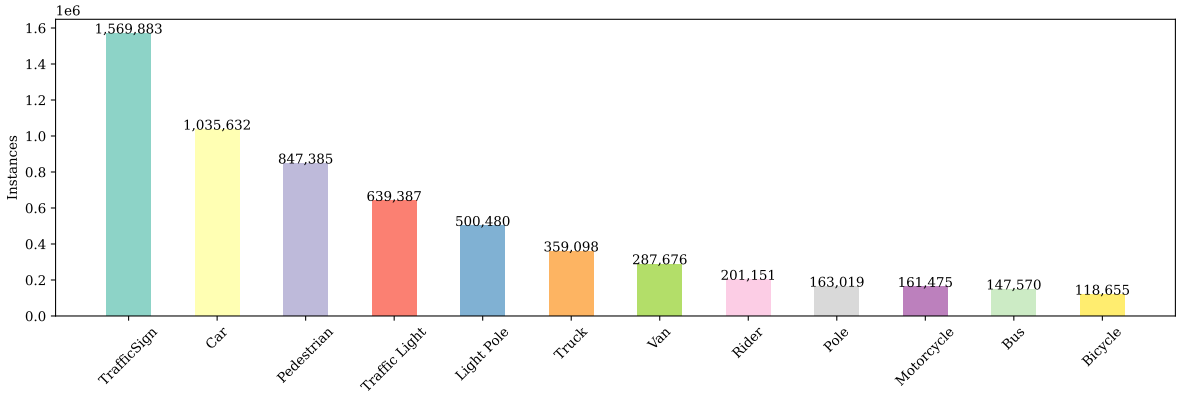


Figure 8. **Number of instances of each object category:** Our dataset contains more diverse objects, such as trailers and gantries, but we excluded objects with less than 100,000 records for clarity of the graph.

D. More Results of Adaptation to Real-World

This section elaborates on the experiments described in Sec. 6.2 and presents supplementary visual results. We trained our method on two synthetic driving datasets and applied it to the real-world driving dataset, INIT [57], for weather domain translation. Although INIT has only one translatable domain, it was chosen for its similarity to the synthetic datasets in terms of camera settings, resulting in relatively minimal performance degradation. In contrast, while BDD100k is a multi-domain real-world driving dataset, its domain gap is larger due to the camera being mounted inside the vehicle windshield, causing issues such as light reflections. Through this experiment, we demonstrate that sufficiently realistic synthetic data, combined with a vision model that preserves source image characteristics, can be effectively applied to real-world scenarios. Additional visual results are presented in Fig. 14.

E. Adaptation to other dataset

In the previous experiments, we demonstrated that our method achieves excellent performance in image translation tasks. We further conducted a qualtitative comparison with text-guided translation models, such as DiffuseIT [35]. Given limitations in computational resources, we leveraged the experimental setup used in DiffuseIT, training our model under identical conditions to compare results with previously evaluated models. We selected Animals Faces [58] dataset for comparison to showcase the effectiveness of our method on non-driving datasets. Figure 15 presents the comparison results, highlighting that our model effectively preserves the structural content from the source image while performing accurate translations. We provided two translation results per case to emphasize that reducing the image prompt scale allows the translation to shift focus from source content to better align with the target.

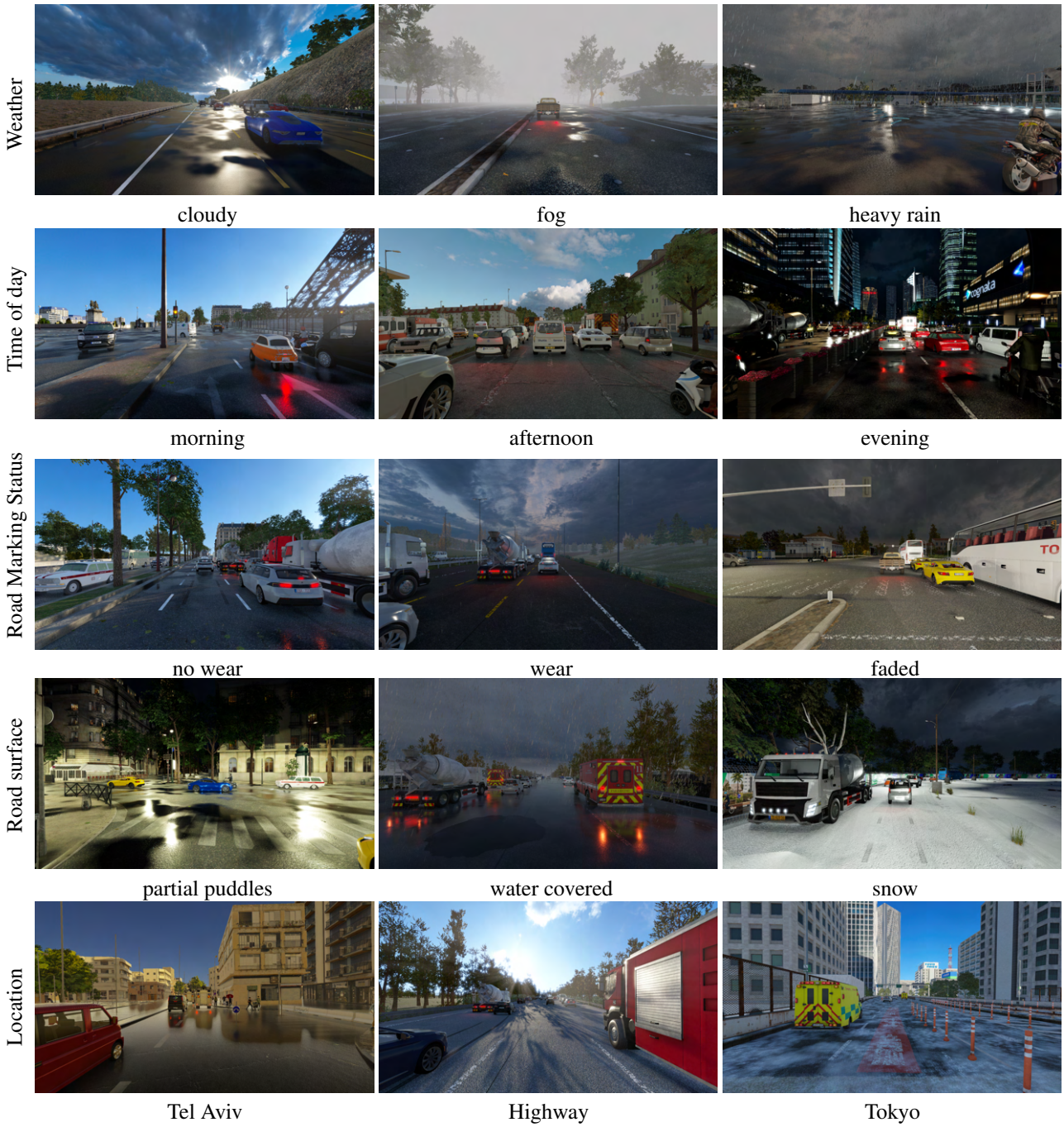


Figure 9. **Sample images for categories in 5 domains:** Weather, Time of day, Road Marking Status, Road Surface, Location.

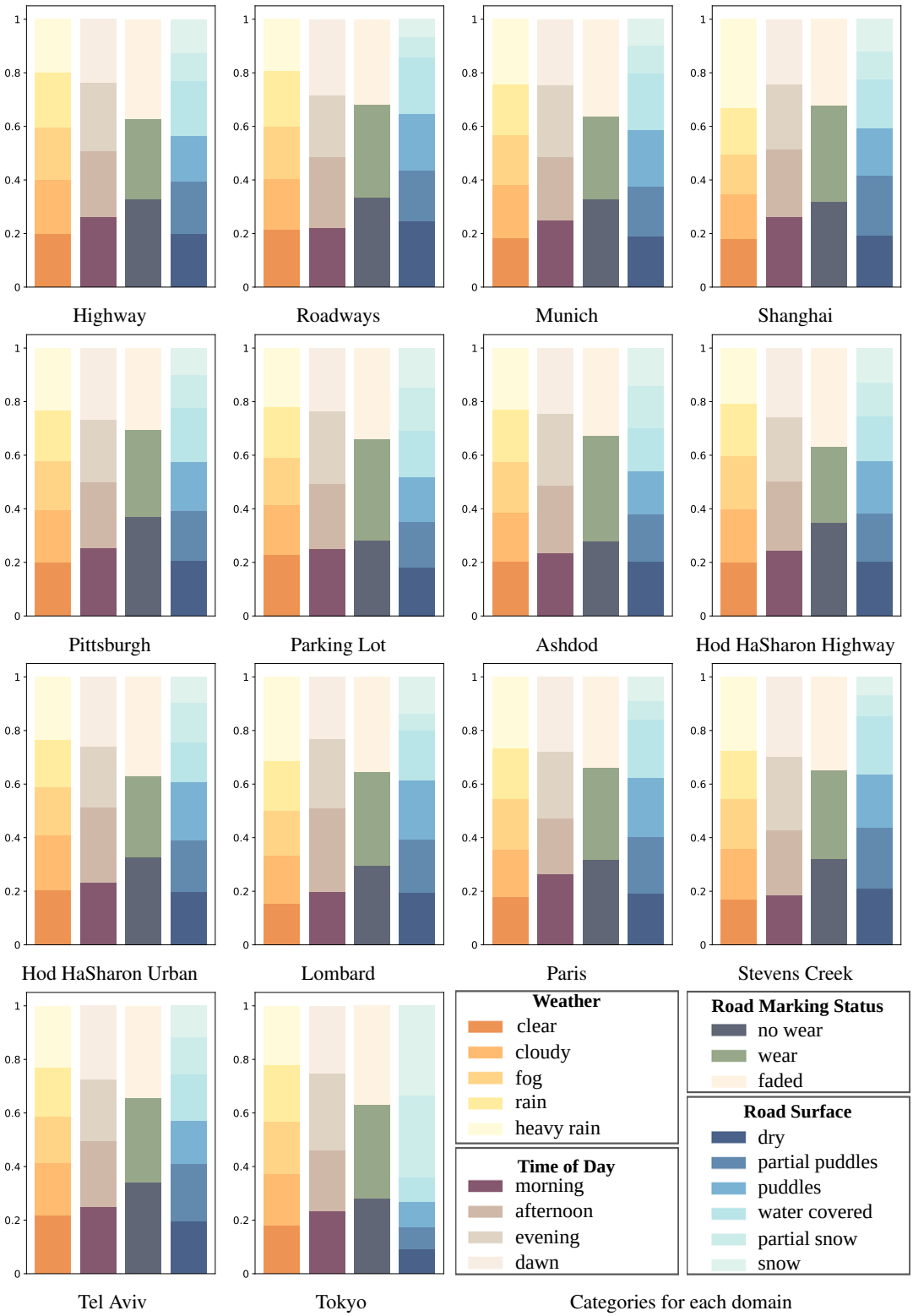


Figure 10. **Visualization of the dataset distribution across domains, grouped by location.**

Method	FID↓	FID_{clip}↓	Structure Distance↓	Background Preservation PSNR↑	LPIPS↓	MES↓	SSIM↑
Baseline	66.37	7.02	0.14	13.24	0.46	0.06	0.39
+CLIP-Full	- 16.46	- 3.16		- 0.51	- 0.03	+ 0.01	- 0.03
+DinoV2	- 0.19	+ 0.20	- 0.01	+ 0.36	- 0.04		+ 0.01
+DIP-Adapter	- 0.26	- 0.14		+ 0.02			
+DTG	- 7.03	- 0.57	- 0.04	+ 1.88	- 0.09	- 0.02	+ 0.06
+DDIM	- 0.98	- 0.58	- 0.02	+ 2.30	- 0.08	- 0.01	+ 0.14
ReSIT (Ours)	41.44	2.78	0.07	17.30	0.21	0.03	0.58

Table 6. **Ablation study results showing performance metric variations in the multi-domain i2i translation methods.**

Figure 11. **The ablation study of multi-domain image-to-image translation methods.** The addition of each component leads to improved preservation of the structural details in the source image, while maintaining effective translation results.

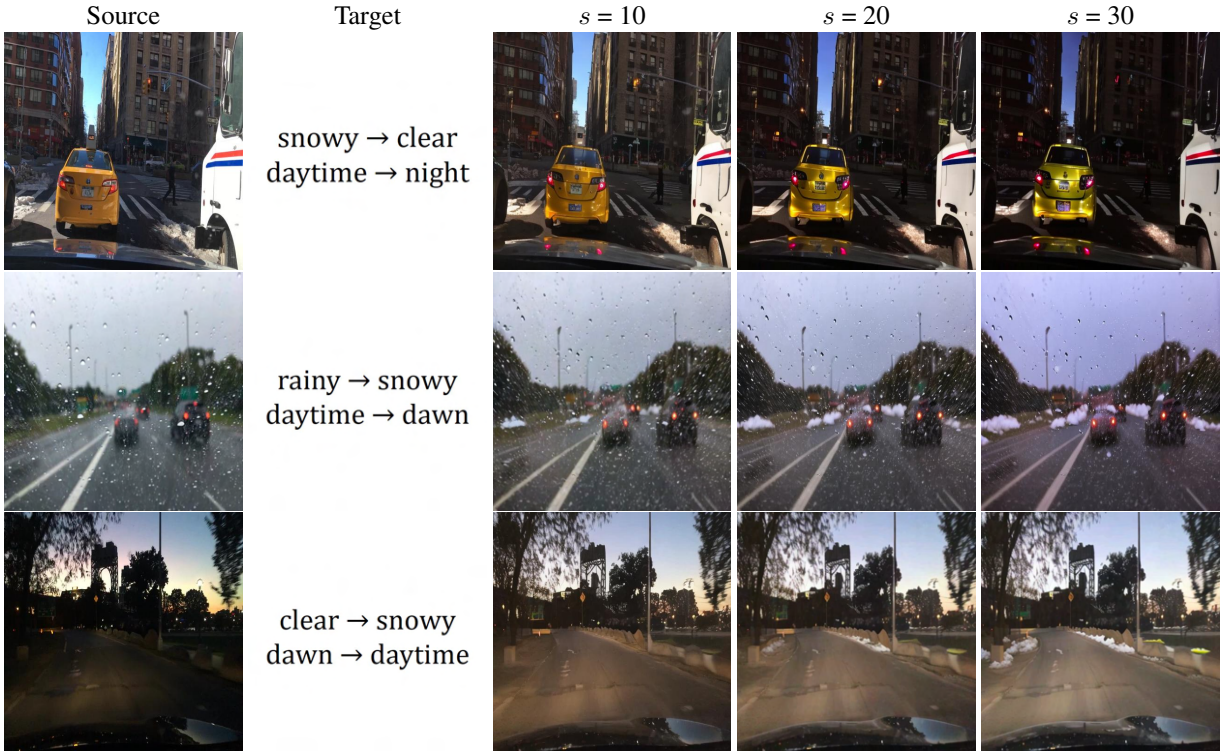


Figure 12. **Variations in image translation results across the translation scale s.** As the translation scale increases, the visual results show stronger translations in both domains guided by the target.

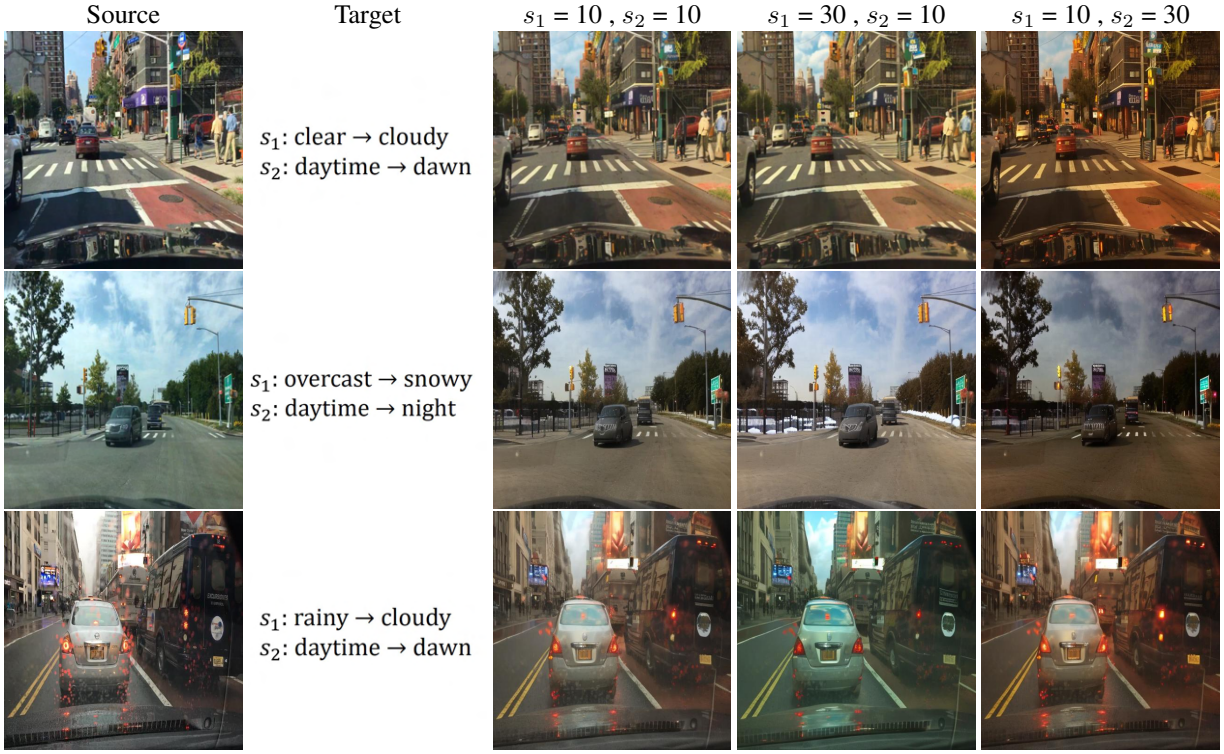


Figure 13. **Variations in image translation results across the differential translation scales s_1, s_2.** The two differential translation scales, s_1 and s_2, control the degree of translation for the weather and time of day domains, respectively.

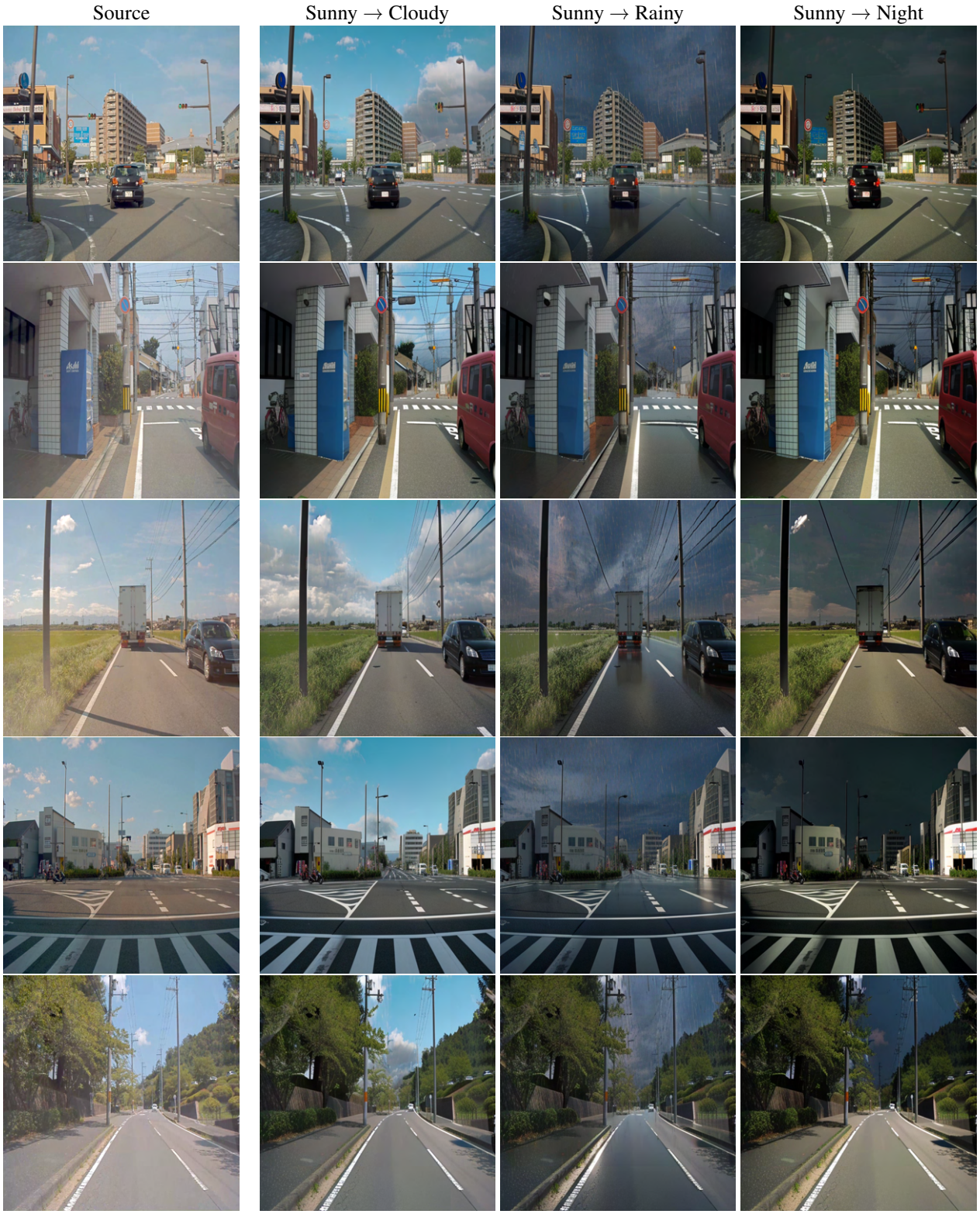


Figure 14. **Image translation results of adaptation to real-world using our method trained ReSIT dataset.**

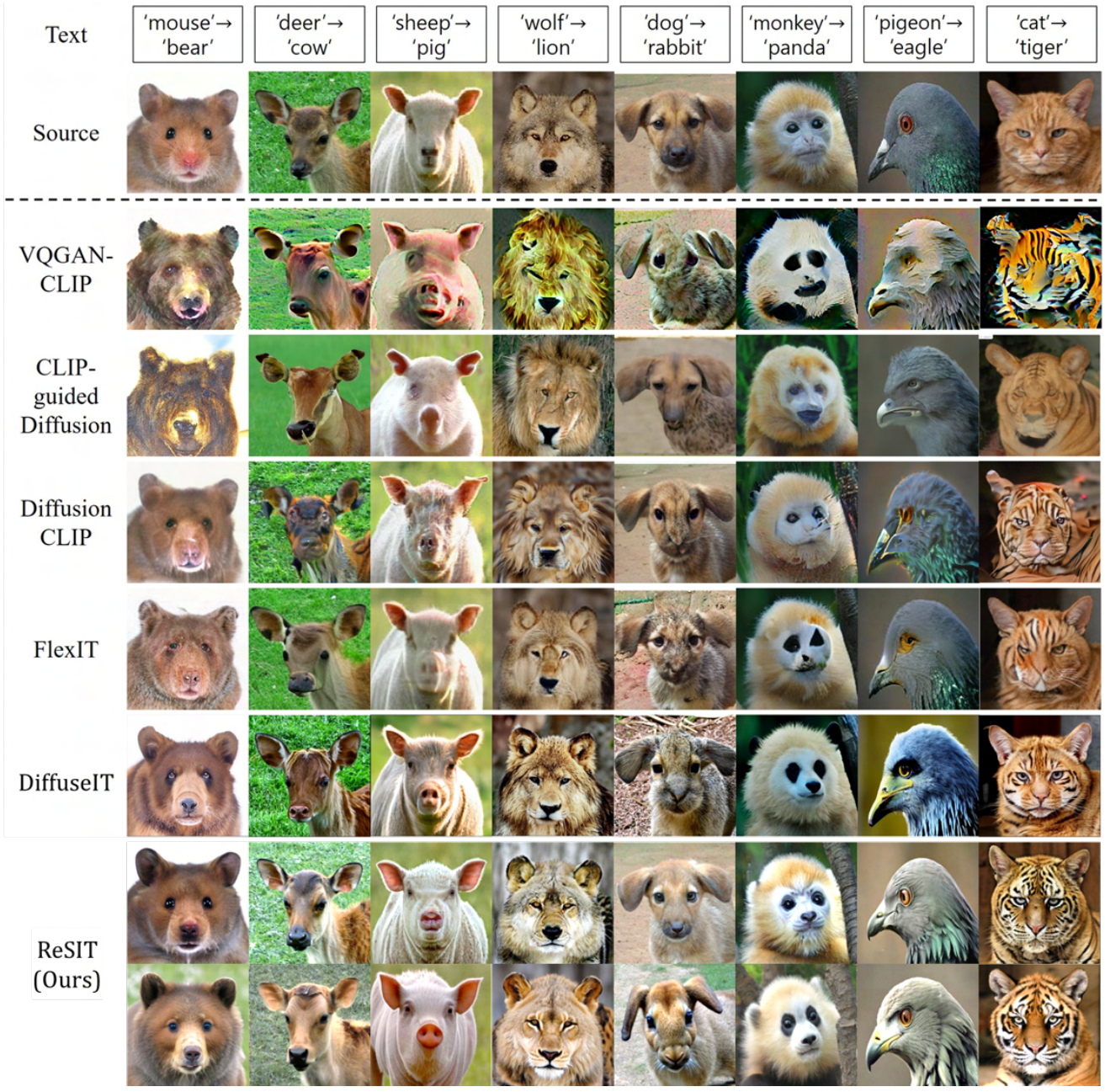


Figure 15. **Qualitative comparison of text-guided translation on *Animal Faces* dataset.** We presented two results for each case in our method by applying image prompt scales of 1.0 and 0.7, respectively. This demonstrates that decreasing the impact of the source image facilitates translations that align more closely with the target.