Seed-to-Seed: Unpaired Image Translation in Diffusion Seed Space

Or Greenberg[1,2]

Eran Kishon[1]

Dani Lischinski[2]

[1] General Motors R&D

[2] The Hebrew University of Jerusalem
Jerusalem, Israel

Abstract

We introduce Seed-to-Seed Translation (StS), a novel approach that combines GANs and diffusion models (DMs) for unpaired Image-to-Image Translation. Our approach is aimed at global translations of complex automotive scenes, where close adherence to the structure and semantics of the source image is essential. We demonstrate that the semantic information encoded in the space of inverted latents (seeds) of a pretrained DM, dubbed as the *seed-space*, can be used for discriminative tasks, and leverage this information to perform image-to-image translation. Our method involves training an *sts-GAN*, an unpaired seed-to-seed translation model, based on CycleGAN. The translated seeds are used as the starting point for the DM's sampling process, while structure preservation is ensured using a ControlNet. We demonstrate the effectiveness of our approach for structure-preserving translation of complex automotive scenes, showcasing superior performance compared to existing GAN-based and diffusion-based methods. In addition to advancing the SoTA in automotive scene translations, our approach offers a fresh perspective on leveraging the semantic information encoded within the seed-space of pretrained DMs for effective image editing and manipulation.

1 Introduction

Diffusion Models (DMs) have emerged as powerful generative tools, synthesizing images by iteratively transforming noise samples, commonly referred as *"seeds"*, into images [17, 40, 41]. State-of-the-art text-to-image diffusion models, *e.g.*, [34, 35, 37] generate diverse and photo-realistic content, prompting efforts to repurpose DMs for a wide range of image editing tasks.

One such task is Image-to-Image Translation (I2IT), where an image is converted from one domain to another while preserving various aspects. In fact, I2IT encompasses a variety of tasks that differ in their required adherence to the source image. For instance, day-to-night translation demands perfect adherence to the structure of the original image, altering only appearance. In contrast, cat-to-dog translations might only preserve the pose and/or fur colors, while allowing other details to change.

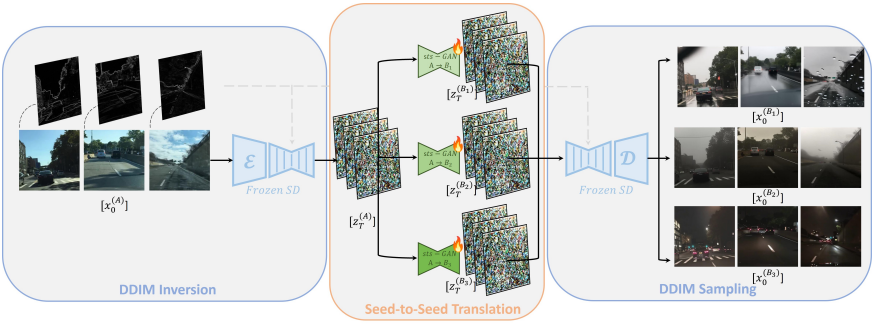

Figure 1: **Seed-to-Seed Translation** addresses the unpaired Image-to-Image Translation task from the source domain A to some target domain B_i by performing the translation in the seed-space of a pretrained diffusion model. The source image $x_0^{(A)}$ is first inverted to a corresponding seed $z_T^{(A)}$. Then the initial seed is translated to a target domain referred seed $z_T^{(B)}$, which is finally sampled to yield the target domain output $x_0^{(B)}$. Here we demonstrate a translation from the source domain "clear day" to 3 different target domains: "rainy day", "foggy day" and "clear night" denoted B_1, B_2 and B_3, respectively.

Despite their generative power, DM-based editing methods often struggle with I2IT tasks requiring strict preservation of complex structures, such as those found in automotive scenes. Consequently, most current approaches in this domain rely on GANs [11], which offer stronger structural guarantees but lack the generative richness of DMs.

In DM-based editing, manipulations often begin from inverted latents—e.g., via DDIM inversion [41]. In this work, we follow the common convention and denote the initial latent as a *seed*, even when it is obtained by inverting an image rather than by random sampling. These seeds encode both attributes to be retained and those to be changed, making it difficult to isolate domain-specific elements. Moreover, editing typically occurs along the DM sampling trajectory, spanning multiple latent spaces. This further complicates the separation of domain-specific and agnostic features at each step.

To address these challenges, we introduce a novel approach to unpaired I2IT that operates directly in the *seed-space* of a pretrained DM, prior to sampling. We show that inverted seeds encode rich semantic information and allow meaningful manipulations within this space. Our method enables controlled translations that preserve structural content while modifying domain-specific features.

We refer to this process as *Seed-to-Seed Translation* (StS), implemented using a GAN-based translator reffered as *sts-GAN*, trained on pairs of inverted seeds from the source and target domains. At inference, an input image is inverted to a seed, translated using *sts-GAN*, and sampled via the DM to produce the target image (Figure 1). A ControlNet [50] ensures the structural integrity of the source image is preserved during sampling.

Our focus is on translations that require high structural fidelity—particularly in complex automotive scenes—while enabling substantial changes in appearance. For example, in day-to-night translation, the geometry and layout must be preserved while lighting and atmospheric effects change. In photo restoration, structure and identity must remain intact even if appearance varies significantly. We demonstrate our method on automotive scene translations, including day-to-night and weather changes. Unlike GANs, which often fail to

produce realistic target domain content, or DMs, which may distort structure, our approach maintains geometric integrity while achieving high realism. In the supplementary material, we also explore the applicability of our approach for several nonautomotive unpaired image translation tasks.

We summarize our main contributions as follows:

1. A fresh perspective on the semantic content of the seeds inverted by a pre-trained DMs, and the structure of the seed-space defined by them, highlighting the potential of image manipulations within the seed space, rather than along the sampling trajectory.

2. We propose StS: a novel method for unpaired, structure-preserving domain translation in complex scenes.

3. A hybrid framework combining GAN-based translation with diffusion-based generation, leveraging their complementary strengths.

2 Related Work

Unpaired I2IT. Unpaired Image-to-Image Translation (I2IT) focuses on translating images across domains without paired training examples. It has gained traction for applications like style transfer [9, 19, 20, 27, 51], semantic segmentation [13, 21, 48], and image enhancement [1, 5]. Many unpaired I2IT methods are GAN-based [11], using cycle consistency [18, 51] to preserve structure during translation. This regularization mitigates mode collapse and encourages content preservation.

Diffusion-based Editing. Diffusion models (DMs) [8, 17, 35] have enabled various image editing tasks. Local editing approaches, such as using masks or attention maps [2, 15, 32], yield strong results but lack the global consistency needed for full-scene translations. In contrast, methods like Imagic [23] allow non-rigid edits but often sacrifice structural fidelity.

Globally-constrained methods like SDEdit [29] and PnP [44] inject source image information early in the generation process, achieving better structure preservation but facing trade-offs between realism and faithfulness. Layout-to-image approaches [25, 31, 50] offer diverse outputs by conditioning on low-dimensional spatial cues (e.g., depth, edges), though they lack the detail required for precise I2IT.

Seed Manipulations. Most methods initiate sampling from a fixed or inverted seed and modify images during denoising. SeedSelect [38] instead optimizes the seed itself to produce rare objects via backpropagation through the entire sampling process. In contrast, our method performs seed-level optimization for unpaired I2IT, operating directly in seed-space rather than along the sampling trajectory.

3 Method

In this section we introduce *StS*, an image translation model that operates directly in the seed-space of a pretrained diffusion model. We begin by exploring the meaningfulness of the seed-space and the ability to access the information encoded within the seeds (Section 3.1). Next, in Section 3.2 we show how seed meaningfulness may be leveraged to perform unpaired image translation within the seed-space using our proposed *StS* model.

Figure 2: **Inverted seed interpolation.** Spherical interpolation between DDIM-inverted seeds from img_A and img_B yields a semantically coherent transformation between the images.

3.1 Meaningful Seed Space

Diffusion models [17, 40, 42] generate images by mapping Gaussian noise (seeds) to images via a stochastic process. To edit real images, one must invert them into the model's seed-space [30, 32, 44, 46]. We adopt DDIM's deterministic sampling and inversion processes [41]; formal details are provided in the supplementary material. Deterministic DDIM sampling defines an injective mapping from seed-space to image-space. Similarly, DDIM inversion maps images back to seed-space.

The DM's backbone iteratively decodes the seed across the diffusion steps [17]. Editing methods typically intervene during this decoding by fine-tuning the decoder's weights [23, 36], modifying the decoder's condition input [29, 50], or injecting cross-attention elements across processes of different images [15, 32, 44]. In all cases, edits occur during the transformation of the latent seed into image.

Prior work [41] shows seed interpolation yields smooth image transitions. As illustrated in Figure 2, spherical interpolation (slerp [39]) between two inverted seeds produces semantically meaningful transitions: e.g., aging (row 1), fog density (row 2), and hair length (row 3). This suggests that seed-space is structurally informative and supports meaningful semantic operations. For instance, the young man (first row) appears progressively older as the interpolation parameter t approaches 1. Similarly, this gradual transformation is reflected in the increasing fog density and changing hair length in the second and third rows, respectively. This illustrates that seed-space encodes structured, interpretable information. To quantify this, we train ResNet18 [14] classifiers on both images and their DDIM-inverted seeds (using Stable Diffusion 2.1 [35]) across classification tasks. As shown in Table 1, seed-based classifiers perform nearly as well as image-based ones, confirming that seed-space retains significant semantic information across scene (time of day), object (dog/cat), and sub-object (age) level attributes. In this work, we embrace this observation and further leverage this structure to perform image translation within the DDIM inverted seed space directly, before sampling.

3.2 *StS*: I2IT in Diffusion Seed Space

We aim to perform unpaired I2IT within the seed-space of a pre-trained diffusion model by leveraging the information encoded in the DDIM inverted seeds. Consequently, we train a dedicated translation model that learns a mapping between seeds corresponding to images from a source domain A to seeds corresponding to images from a target domain B. We train

Task	seeds	images
Day/Night	98.37%	98.47%
Cat/Dog	90.10%	98.53%
Older/Younger	92.60%	97.90%

Table 1: **Classifier Accuracy Comparison.** Classifiers are trained once on image inputs and once on their corresponding inverted seeds. The tasks are day/night, cat/dog, and older/younger (using the *BDD100k* [49], *AFHQ* [6], and *FFHQ* [22] datasets, respectively). More details can be found in the supplementary material.

our network, referred to as *sts-GAN*, over a set of DDIM-inverted seeds from the source and target domains, using the CycleGAN architecture [51] and training strategy. clasiffier-free guidance (CFG) scale $\omega = 1.0$ is used to accurately invert the unpaired source and target domain training images to the seed-space.

Figure 1 presents a diagram depicting our method. At inference time, we first encode the input source image $x_0^{(A)}$ to the Stable Diffusion (SD) latent space, yielding $z_0^{(A)}$, and apply DDIM inversion (with a source-domain-referred prompt) to obtain a seed $z_T^{(A)}$. Next, we translate $z_T^{(A)}$ to a target-domain-referred seed $z_T^{(B)}$ using our *sts-GAN*. Finally, we sample $z_T^{(B)}$ using the same pre-trained SD model (with a target-domain-referred prompt), yielding the final denoised code $z_0^{(B)}$, which is decoded to the resulting image $x_0^{(B)}$.

While *sts-GAN* successfully translates source-referred seeds into target-referred ones, DDIM sampling these seeds without CFG typically results in images suffering from a lack of local semantic effects, despite the use of a target domain-referred prompt (*e.g.*, "A clear night", for the day-to-night translation). For example, as demonstrated in the "$\omega = 1.0$" columns of Figure 3, a day-to-night translation of automotive images might lack car lights, street lights, and reflections (left), or retain some daytime-like shadows on the road surface (right). To encourage such domain-specific effects, we employ CFG with $\omega = 5.0$, in conjunction with the same target-referred prompt.

The cyclic consistency mechanism employed during *sts-GAN* training enforces structural similarity between the source and the output *within the seed space*. However, this similarity might not be maintained as the translated seed $z_T^{(B)}$ is sampled back to the image space. This issue becomes more pronounced when using CFG, as the extrapolation amplifies the accumulated errors from the DDIM inversion [30]. Consequently, even if the translation from $z_T^{(A)}$ to $z_T^{(B)}$ is perfect in seed space, the final image $x_0^{(B)}$ may significantly deviate from the structure and content of the source image. To address this, we employ ControlNet [50] to enforce structural similarity between the source image and the final output throughout the sampling trajectory.

The "$\omega = 5.0$" column of Figure 3 demonstrates that spatially-guided conditional sampling enhances the target-domain appearance, introducing the missing effects, while remaining faithful to the source image's structure. Additional discussion and quantitative evaluation of the CFG scale factor are provided in the supplementary material.

4 Experiments

We evaluate our method through extensive experiments on automotive unpaired image translation tasks, comparing against leading GAN-based and globally-constrained DM-based ap-

Figure 3: **Day-to-night translation with *StS* using different CFG-scales.** While achieving a global night-time appearance, a low CFG-scale ($\omega = 1$) may result in lack of local domain-related semantic effects (middle). Using a higher CFG-scale ($\omega = 5$) introduces these important effects (right). The same prompt "A clear night" is used in both columns.

proaches. While GAN-based methods are currently considered state-of-the-art for lighting and weather translation in automotive scenes, our results demonstrate the efficacy of diffusion models in these challenging tasks. We present quantitative results on the Day-to-Night task and qualitative results across multiple image translation tasks, followed by an ablation study of our method's components. Code and models will be available upon publication.

4.1 Implementation Details

We evaluate unpaired I2IT tasks using the Berkeley DeepDrive *BDD100k* [49] and *DENSE* [2] datasets. Our implementation uses Stable Diffusion (SD) 2.1 [35] at 512×512 resolution as the diffusion backbone. For *sts-GAN*, we employ a modified *ResNet18* encoder with 4-channel input to match SD's latent space, omitting the final normalization layer. We trained *sts-GAN* following the methodology of Zhu et al. [51]. Note that due to the low dimensionality of SD's latent space (8 times smaller than the image space along each axis, with an additional channel), training *sts-GAN* in the latent space is significantly faster than training GANs in the image space. Due to SD's limited performance on automotive datasets, we finetune both SD 2.1 and its corresponding ControlNet [50] on the *BDD100k* training set using Diffusers' [45] default scheme. Both DDIM sampling and DDIM inversion use 20 timesteps, with CFG-scales of $\omega = 1.0$ for inversion and $\omega = 5.0$ for forward sampling.

4.2 Baselines

We compare our Day-to-Night translation on *BDD100k* against GAN-based methods (CycleGAN [51], MUNIT [19], TSIT [20], AU-GAN [24], and CycleGAN-Turbo [33]), using the provided day2night checkpoints for AU-GAN and CycleGAN-Turbo. We trained the other models for 100 epochs using the provided public code with default hyperparameters, and selected the best checkpoints.

As mentioned in Section 2, while state-of-the-art image editing techniques deliver outstanding results for object-level edits or relatively straightforward images, they often fail when applied to global edits of complex scenes. Diffusion-based methods, in particular, struggle to balance high fidelity to the source image with achieving the desired modifications in such challenging scenarios. Qualitative examples of these limitations are provided in the supplementary material. For our comparisons, we selected diffusion-based baselines with global-constraints, which are more suitable for global edits (as discussed in Section 2). Specifically, we quantitatively evaluate our performance against SDEdit [29] with varying strength parameters (0.5, 0.7, 0.9) and Plug-and-Play (PnP) [44]. We also include compar-

	GAN baselines					Diffusion baselines				StS (ours)
	CycleGAN	MUNIT	TSIT	AU-GAN	CycleGAN-turbo	SDEdit 0.5	SDEdit 0.7	PnP	ControlNet	
FID ↓	19.908	52.152	21.315	14.426	16.840	73.494	48.757	61.617	35.091	16.384
MMD ↓	58.395	260.081	56.484	45.970	49.845	242.001	161.666	172.808	95.171	41.344
KID ↓	4.539	12.968	4.446	3.985	4.215	12.097	9.185	9.575	6.340	3.718
SSIM ↑	0.469	0.308	0.3929	0.463	0.431	0.661	0.603	0.768	0.493	0.505

Table 2: **Quantitative comparison to other methods.** Day-to-Night translation on the *BDD100k* dataset. For each metric, the best and second-best scores are shown in blue and red, respectively.

isons to a pure ControlNet [50], which is designed for image synthesis using a combination of textual and spatial conditions. To ensure a fair comparison, we utilize our fine-tuned U-Net for zero-shot diffusion-based methods (SDEdit and PnP) as well as ControlNet when working with the automotive datasets. For all these methods, we use the default settings of 50 timesteps and CFG-scale $\omega = 7.5$ during inference.

4.3 Evaluation Metrics

We follow the standard evaluation protocol used in prior GAN-based I2IT works [4, 26, 28], employing SSIM [47] and FID [10, 16] to assess weather and lighting translation tasks. While feature-based metrics like DINO-Struct-Dist [43] have gained popularity, we found them unstable for complex automotive scenes. Given the limited size of our validation datasets (up to a few thousands samples per domain), we additionally report KID [3] and MMD [17], which are considered more suitable for smaller datasets.

4.4 Results

Quantitative results are presented in Table 2. Our method achieves the lowest MMD and KID scores and the second lowest FID score. It should be noted that the high SSIM scores achieved by SDEdit and PnP result from their frequent failure to achieve the target domain appearance, as reflected by their low FID, KID, and MMD scores. This phenomenon is explained by the inherent trade-off between achieving the desired target domain appearance and preserving the content from the source image without the cycle-consistency mechanism. For example, when increasing the strength parameter of SDEdit above 0.7, the results become increasingly disconnected from the source image (see Figure 5). Our model exhibits the best balance between target domain appearance and structure preservation compared to the baselines.

Qualitatively, Section 4.4 compares our *StS* results to both GAN-based and diffusion-based methods for the Day-to-Night task using the *BDD100k* dataset. Our model achieves the highest level of realism compared to all other methods. The GAN-based methods mostly suffer from the occurrence of artifacts, primarily manifested as random light spots that are uncorrelated with semantically meaningful potential light sources in the image (*e.g.*, car headlights, taillights, streetlights, which are commonly turned off during the day but can be turned on at night). Our model minimizes the occurrence of these artifacts and leverages the powerful semantic understanding of the diffusion model to accurately generate semantics-related target domain effects, such as light sources, light scatters, and reflections (see Figure 4). While PnP and SDEdit struggle to balance between output realism and structural preservation, our model excels in both aspects.

Figure 4: Qualitative comparison for Day-to-Night translation over the *BDD100k* dataset - GAN-based baselines

Figure 5: Qualitative comparison for Day-to-Night translation over the *BDD100k* dataset - Diffusion-based baselines

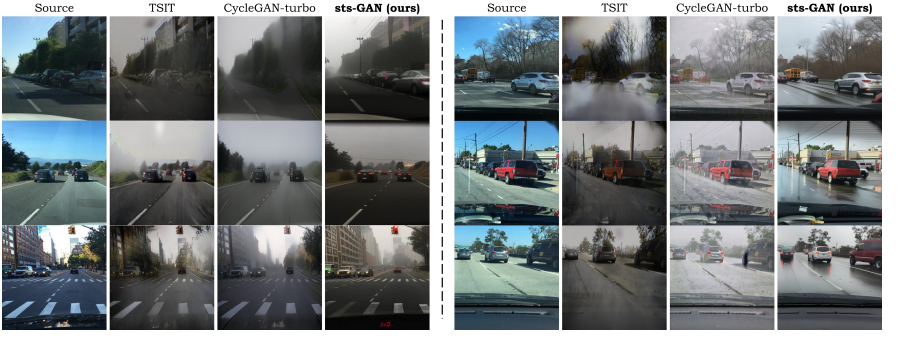

Figure 6: Qualitative comparison of weather translation- *clear2fog* (left) and *clear2rain* (right). See the suppl. material for additional examples and domains.

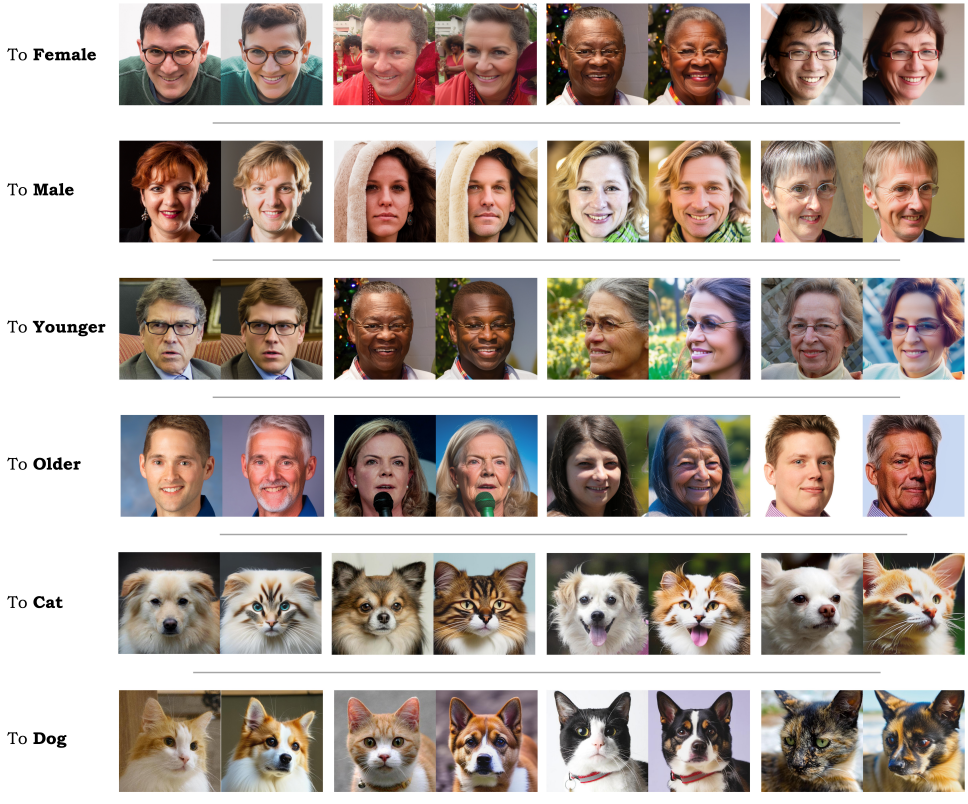

Figure 7: Non-automotive translations on the *FFHQ* [22] (faces) and *AFHQ* [6] (cats/dogs) datasets. In each pair, the left image is the source and the right image is the translated version.

Figure 6 qualitatively demonstrates our model's performance for weather translations, compared to SoTA GAN-based methods. Specifically, we experimented with Clear-to-Foggy and Clear-to-Rainy translations. Additional results, as well as night-time examples, are provided in the supplementary material. To train *sts-GAN* for these weather translations, we utilized clear and rainy images from *BDD100k* (both day and night) and foggy images from both the "light fog" and "dense fog" splits of the *DENSE* dataset [1]. While primarily focused on automotive translations, our model is versatile enough to be generalized for classic object-level edits. Qualitative examples showcasing representative non-automotive applications are provided in Figure 7 and are extended in the supplementary material.

4.5 Ablation Study

We evaluate the contribution of each component in our method through controlled ablations (see Table 4). Starting from a baseline ControlNet initialized with a random seed (RS), we incrementally introduce two components: (1) inverted seed initialization (Inv), and (2) seed-to-seed translation via sts-GAN (ST). Our complete model **StS**, which combines both components, achieves the best results across all metrics, notably improving appearance metrics (FID, KID, MDD) while preserving structure. This confirms that the sts-GAN enables

CFG-scale	FID ↓	MMD ↓	KID ↓	SSIM ↑
1.0	25.955	67.404	5.364	0.549
3.0	17.454	45.540	3.988	0.526
5.0	16.384	41.344	3.718	0.505

Table 3: **CFG Ablation study.** Balance between content preservation and target domain appearance via CFG-scale.

Method	FID ↓	MMD ↓	KID ↓	SSIM ↑
ControlNet+RS	35.091 (+114%)	95.171 (+130%)	6.340 (+70%)	0.493 (+2%)
ControlNet+Inv	49.572 (+202%)	411.060 (+894%)	14.981 (+296%)	**0.756 (-49%)**
ControlNet+RS+ST	21.316 (+30%)	67.650 (+63%)	5.5456 (+49%)	0.450 (+11%)
ControlNet+Inv+ST (*StS*)	**16.384**	**41.344**	**3.718**	0.505

Table 4: **Ablation study - Model Components.** Day-to-Night translation over *BDD100k*.

accurate domain transfer while retaining content fidelity. In contrast, using only inverted seeds leads to poor target appearance due to low editability, while using ST over random seeds suffers from structural degradation.

We also ablate the CFG-scale (Table 3), finding that $\omega = 5.0$ yields the best trade-off between structural preservation and domain consistency. For further analysis and qualitative examples, please refer to the supplementary material.

5 Conclusions and Future Work

We propose a novel framework for unpaired image-to-image translation that leverages the generative power of pretrained diffusion models (DMs) through seed-space manipulation. Our architecture combines a task-optimized GAN, responsible for unpaired translation in seed-space, with a frozen DM for encoding and image synthesis. This hybrid design allows for structure-preserving translations of complex scenes, outperforming both GAN- and DM-based baselines across various automotive tasks. To mitigate structural drift during sampling, we incorporate ControlNet for spatial guidance. In future work, we aim to replace ControlNet with more general structural control mechanisms. Additionally, we plan to explore whether recent advances in inversion techniques can improve seed quality and structural fidelity through more accurate reconstructions.

Our method is general and may benefit from integration with emerging DiT-based diffusion backbones (e.g., SD3.x, FLUX). It is also well-suited for distilled models that require fewer sampling steps, where manipulations along the shorter sampling trajectory can be strengthened by a meaningful seed. More broadly, we believe that seed-space manipulations represent a versatile paradigm that extends beyond unpaired image translation, with the potential to open new directions across diverse generative tasks.

References

[1] Saleh Altakrouri, Sahnius Bt Usman, Norulhusna Binti Ahmad, Taghreed Justinia, and Norliza Mohd Noor. Image to image translation networks using perceptual adversarial loss function. In *2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 89–94. IEEE, 2021.

[2] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus

Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11682–11692, 2020.

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[5] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6306–6314, 2018.

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[9] Ujjal Kr Dutta. Seeing objects in dark with continual contrastive learning. In *European Conference on Computer Vision*, pages 286–302. Springer, 2022.

[10] Maurice Fréchet. Sur la distance de deux lois de probabilité. In *Annales de l'ISUP*, volume 6, pages 183–198, 1957.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[13] Xi Guo, Zhicheng Wang, Qin Yang, Weifeng Lv, Xianglong Liu, Qiong Wu, and Jian Huang. GAN-based virtual-to-real image translation for urban scene semantic segmentation. *Neurocomputing*, 394:127–135, 2020.

[14] 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.

[15] 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.

[16] 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.

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

[18] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *Proc. ICLR*, 2018.

[19] 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.

[20] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. TSIT: A simple and versatile framework for image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 206–222. Springer, 2020.

[21] Jingxuan Kang, Bin Zang, and Weipeng Cao. Domain adaptive semantic segmentation via image translation and representation alignment. In *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 509–516. IEEE, 2021.

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[23] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023.

[24] Jeong-gi Kwak, Youngsaeng Jin, Yuanming Li, Dongsik Yoon, Donghyeon Kim, and Hanseok Ko. Adverse weather image translation with asymmetric and uncertainty-aware GAN. *arXiv preprint arXiv:2112.04283*, 2021.

[25] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. ControlNet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pages 129–147. Springer, 2025.

[26] Jie Liang, Hui Zeng, and Lei Zhang. High-resolution photorealistic image translation in real-time: A Laplacian pyramid translation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9392–9400, 2021.

[27] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

[28] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] 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.

[30] 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.

[31] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, Feb 2023.

[32] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[33] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024.

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[35] 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.

[36] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023.

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[38] Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. It is all about where you start: Text-to-image generation with seed selection. *arXiv preprint arXiv:2304.14530*, 2023.

[39] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985.

[40] 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.

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[42] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, 2019.

[43] 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.

[44] 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.

[45] 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.

[46] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023.

[47] 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.

[48] Feilong Wu and Suya You. Semi-supervised semantic segmentation via image-to-image translation. In *Automatic Target Recognition XXXI*, volume 11729, pages 100–106. SPIE, 2021.

[49] 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.

[50] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[51] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.