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

Or Greenberg[1,2]
or.greenberg@gm.com
Eran Kishon[1]
eran.kishon@gm.com
Dani Lischinski[2]
danix3d@gmail.com

[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.

© 2025. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

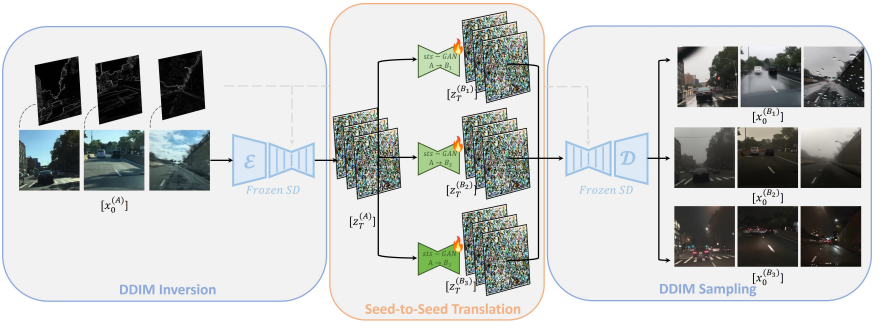

Figure 1: **Seed-to-Seed Translation** addresses the unpaired Image-to-Image Translation task from the source domain A to some target domain B_i by performing the translation in the seed-space of a pretrained diffusion model. The source image $x_0^{(A)}$ is first inverted to a corresponding seed $z_T^{(A)}$. 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 im ages via a stochastic process. To edit real images, one must invert them into the mode seed-space [30, 32, 44, 46]. We adopt DDIM's deterministic sampling and inversion pro cesses [41]; formal details are provided in the supplementary material. Deterministic DDIM sampling de nes 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]. Edit ing methods typically intervene during this decoding by ne-tuning the decoder's weights [23, 36], modifying the decoder's condition input [29, 50], or injecting cross-attention ele- ments across processes of different images [15, 32, 44]. In all cases, edits occur during t 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 seman cally 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 seman operations. For instance, the young man (rst row) appears progressively older as the inte polation parameter t approaches 1. Similarly, this gradual transformation is re ected in the increasing fog density and changing hair length in the second and third rows, respectivel This illustrates that seed-space encodes structured, interpretable information. To quanti this, we train ResNet18 [14] classi ers on both images and their DDIM-inverted seeds (us ing Stable Diffusion 2.1 [35]) across classi cation tasks. As shown in Table 1, seed-base classi ers perform nearly as well as image-based ones, con rming that seed-space retai signi cant semantic information across scene (time of day), object (dog/cat), and sub-objec (age) level attributes. In this work, we embrace this observation and further leverage thi 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 dedicated translation model that learns a mapping between seeds corresponding to imag 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: Classi er Accuracy Comparison. Classi ers are trained once on image inputs and once on their corresponding inverted seeds. The tasks are day/night, cat/dog, 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. clasif er-free guidance (CFG) scale $w = 1.0$ 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 rst 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 nal 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 ("A clear night", for the day-to-night translation). For example, as demonstrated in the "$w = 1.0$" columns of Figure 3, a day-to-night translation of automotive images might lack car lights, street lights, and re ections (left), or retain some daytime-like shadows on the road surface (right). To encourage such domain-speci c effects, we employ CFG with $w = 5.0$, 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 ampli es 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 nal image $x_0^{(B)}$ may signi cantly deviate from the structure and content of the source image. To address this, we employ ControlNet to enforce structural similarity between the source image and the nal output throughout the sampling trajectory.

The "$w = 5.0$" 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

Figure 3: Day-to-night translation with StS using different CFG-scales. While achieving a global night-time appearance, a low CFG-scale ($w = 1$) may result in lack of local domain-related semantic effects (middle). Using a higher CFG-scale ($w = 5$) 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 lightin and weather translation in automotive scenes, our results demonstrate the ef cacy of diffu sion models in these challenging tasks. We present quantitative results on the Day-to-Nig task and qualitative results across multiple image translation tasks, followed by an ablatio 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 modi ed ResNet18 encoder with 4-channel input to match SD's latent space, omitting the nal 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 ax with an additional channel), training sts-GAN in the latent space is signi cantly faster than training GANs in the image space. Due to SD's limited performance on automotive datasets we netune 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 $w = 1.0$ for inversion and $w = 5.0$ 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 hyperparameter and selected the best checkpoints.

As mentioned in Section 2, while state-of-the-art image editing techniques deliver out standing results for object-level edits or relatively straightforward images, they often fail when applied to global edits of complex scenes. Diffusion-based methods, in particula struggle to balance high delity to the source image with achieving the desired modi ca- tions in such challenging scenarios. Qualitative examples of these limitations are provide in the supplementary material. For our comparisons, we selected diffusion-based baselin with global-constraints, which are more suitable for global edits (as discussed in Section 2 Speci cally, 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 BDD100kdataset. For each metric, the best and second-best scores are shown in blue red, respectively.

isons to a pure ControlNet [50], which is designed for image synthesis using a combinat of textual and spatial conditions. To ensure a fair comparison, we utilize our ne-tuned Net for zero-shot diffusion-based methods (SDEdit and PnP) as well as ControlNet wh working with the automotive datasets. For all these methods, we use the default setting 50 timesteps and CFG-scale $w = 7.5$ 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 tasl While feature-based metrics like DINO-Struct-Dist [43] have gained popularity, we four them unstable for complex automotive scenes. Given the limited size of our validati datasets (up to a few thousands samples per domain), we additionally report KID [3] a MMD [12], which are considered more suitable for smaller datasets.

4.4 Results

Quantitative results are presented in Table 2. Our method achieves the lowest MMD KID scores and the second lowest FID score. It should be noted that the high SSIM sco achieved by SDEdit and PnP result from their frequent failure to achieve the target dom appearance, as re ected by their low FID, KID, and MMD scores. This phenomenon is e plained by the inherent trade-off between achieving the desired target domain appear and preserving the content from the source image without the cycle-consistency mechan 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 exh the best balance between target domain appearance and structure preservation compa 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 BDD100kdataset. Our model achieves the highest level of realism compared to all other methods. The GAN-based methods mo suffer from the occurrence of artifacts, primarily manifested as random light spots that 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 turned on at night). Our model minimizes the occurrence of these artifacts and leverages powerful semantic understanding of the diffusion model to accurately generate seman related target domain effects, such as light sources, light scatters, and re ections (see ure 4). While PnP and SDEdit struggle to balance between output realism and struct preservation, our model excels in both aspects.