Domain-Aware Diffusion for Synthetic-to-Real Data Augmentation

Salahidine Lemaachi[1,*][a], Anaïs Druart[2][b] and Nicolas Winckler[2][c]
[1]*DataDog, Paris, France*
[2]*Computer Vision Laboratory, Ipsotek-Atos, Grenoble, France*

Keywords: Generative Models, Diffusion, Domain Shift, Domain Adaptation, Synthetic Data, Conditional Control.

Abstract: Synthetic data is increasingly used to train deep models when large-scale annotated real datasets are unavailable, but performance often degrades due to the domain gap between synthetic and real images. We propose a diffusion-based framework for synthetic-to-real style transfer that produces realistic images while preserving semantic structure. Our method builds on latent diffusion models with ControlNet and introduces three key ideas. First, we design a dual-control representation that fuses segmentation maps with Canny edges, ensuring both semantic layout fidelity and fine-grained detail preservation while improving efficiency by avoiding multiple control passes. Second, we introduce *domain-aware prompting*, where lightweight tokens (synthetic" or real") are added to prompts to control domain style in image translation. Third, we adopt an iterative refinement loop in which generated images with artifacts are progressively reintroduced into training, allowing the model to correct its own errors. Experiments on GTA-to-Cityscapes show that our approach reduces the domain gap, improves mean IoU, and trains significantly faster than GAN-based baselines. Our code and data are available at https://github.com/bds-ailab/syn2real.

1 INTRODUCTION

Deep learning models continue to scale in size and capability, and their training increasingly relies on large datasets. Synthetic data offers a promising alternative to real-world data collection, as it can be generated in unlimited quantities and avoids regulatory or privacy constraints. However, models trained exclusively on synthetic data often fail to generalize when evaluated on real-world benchmarks, primarily due to the *domain shift* between synthetic and real images. Even small discrepancies in texture, lighting, or object variability can substantially degrade performance on downstream tasks.

Two main families of approaches are commonly explored to address this problem: *feature-level domain adaptation* and *pixel-level domain adaptation*. In feature-level methods, the objective is to align feature embeddings across domains, either explicitly by minimizing distributional distances such as Maximum Mean Discrepancy or correlation distance (Scheck et al., 2021), or implicitly through adversarial discriminators trained to distinguish source and target features (Tsai et al., 2018). While effective, these approaches can disrupt semantic consistency, sometimes mapping features of one object class in the source domain to another in the target domain.

Pixel-level adaptation instead translates images from the source domain into the style of the target domain while preserving their semantic structure. Adversarial methods have been widely used for this purpose. For instance, Shrivastava et al. (Shrivastava et al., 2017) proposed adapting synthetic images from simulators into realistic ones using a GAN-based framework. Although successful in certain settings, adversarial training often struggles with high-resolution images and complex layouts, and is notoriously unstable due to mode collapse.

Diffusion models have recently emerged as a more stable and higher-quality alternative for image generation. For instance, (Dhariwal and Nichol, 2021) showed that diffusion models outperform GANs on image synthesis tasks, offering superior fidelity and diversity. Building on these advances, ControlNet (Zhang et al., 2023) extends latent diffusion models with spatial conditioning, allowing them to respect structure such as edges, depth, or segmentation maps.

[a] https://orcid.org/0009-0002-3003-3404
[b] https://orcid.org/0000-0003-1916-8615
[c] https://orcid.org/0009-0005-9211-2572
* Work performed while at Eviden/Ipsotek (Atos Group).

Lemaachi, S., Druart, A. and Winckler, N.
Domain-Aware Diffusion for Synthetic-to-Real Data Augmentation.
DOI: 10.5220/0014307900004084

In *Proceedings of the 21st International Conference on Computer Vision Theory and Applications (VISAPP 2026) - Volume 1*, pages 63-74
ISBN: 978-989-758-804-4; ISSN: 2184-4321

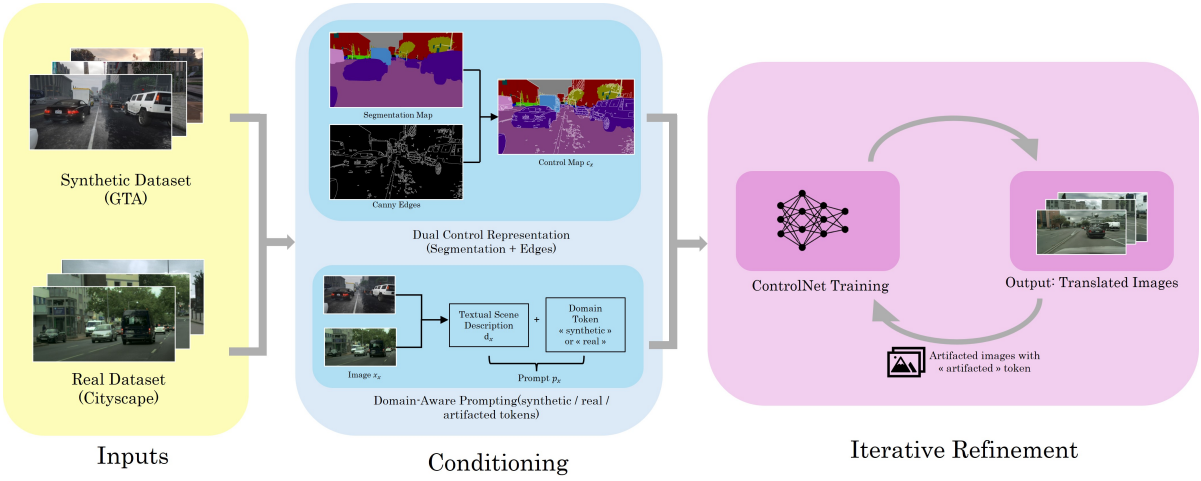

Figure 1: Overview of our synthetic-to-real translation pipeline. The framework combines dual-control conditioning, domain-aware prompting, and iterative refinement within ControlNet. Synthetic images from GTA and real images from Cityscapes are used to build dual control representations by fusing segmentation maps and Canny edges with added noise. In parallel, prompts are augmented with domain tokens ("real" or "synthetic"). The model is iteratively refined by reintroducing artifacted samples labeled with an additional "artifacted" token.

In this work, we propose a diffusion-based synthetic-to-real style transfer method designed to preserve semantic annotations while improving realism. While our approach follows the objective of domain adaptation, it leverages limited target-domain annotations during generative model fine-tuning and therefore does not correspond to a purely unsupervised setting. Our approach introduces two main ideas. First, we design a dual-control representation that combines segmentation maps and Canny edges, thus providing both high-level semantic layouts and fine-grained object details. Second, we incorporate a *domain-aware prompting* mechanism, where lightweight tokens (e.g., "synthetic" or "real") encode the domain style directly in the prompt. To further enhance the realism of the generated data, we adopt an *iterative refinement loop*, where generated images, including those with artifacts, are progressively reintroduced into training so that the model learns to correct its own errors.

An illustration of our pipeline is shown in Figure 1. Our contributions can be summarized as follows:

- We propose a dual-control representation combining segmentation maps and edges to improve structural and semantic consistency in synthetic-to-real style transfer.
- We introduce domain-aware prompting with lightweight tokens to guide diffusion models toward specific domain styles.
- We develop an iterative refinement loop that reduces artifacts and enhances realism, yielding improved performance for downstream segmentation tasks on Cityscapes (Cordts et al., 2016).

2 RELATED WORK

2.1 Feature-Level Adaptation

Feature-level domain adaptation aims to align latent feature distributions between source and target domains, either by minimizing statistical distances such as Maximum Mean Discrepancy (MMD) (Scheck et al., 2021) or by training adversarial discriminators (Tsai et al., 2018). While effective in reducing domain gaps, such methods can introduce semantic inconsistencies, as alignment is performed without explicit class-level supervision, sometimes blending visually similar yet semantically distinct categories. Pixel-level approaches, by contrast, operate directly in image space under structural constraints, which generally helps preserve spatial coherence and semantic consistency, making them well suited for synthetic-to-real translation.

2.2 GAN-Based Style Transfer for Domain Adaptation

Pixel-level domain adaptation has been extensively studied using generative models. On the adversarial side, (Shrivastava et al., 2017) adapted simulator-rendered images into realistic ones with GANs. CycleGAN (Zhu et al., 2017) introduced cycle consis-

tency to enable unpaired translation, and CyCADA (Hoffman et al., 2017) extended this idea specifically for domain adaptation. These methods demonstrated promising results but often struggled with high-resolution images and complex layouts. Moreover, adversarial training remains prone to instability and structural hallucinations.

More recently, hybrid approaches combining adversarial and diffusion principles have emerged (Isola et al., 2018). CycleGAN-Turbo (Parmar et al., 2024) integrates the efficiency of Stable Diffusion Turbo (Sauer et al., 2024) with a CycleGAN-style framework. By leveraging diffusion priors within a GAN-like architecture, it improves training stability and sample fidelity compared to purely adversarial models. However, its benchmarks mainly involve tasks where domains differ primarily in texture or style (e.g., horse-to-zebra, day-to-night). In more complex synthetic-to-real settings, we observed that it can still alter semantic layouts, whereas our dual-control diffusion approach better preserves structural fidelity.

2.3 Diffusion-Based Approaches

Diffusion models have recently become a powerful alternative to adversarial generation, outperforming GANs in both fidelity and diversity, as shown in (Goodfellow et al., 2014), (Ledig et al., 2017), and (Dhariwal and Nichol, 2021). Their foundation was originally established by (Sohl-Dickstein et al., 2015), who formulated data generation as the reversal of a gradual noising process. This idea was later revitalized through score-based generative modeling by (Song et al., 2019), which estimated the data density gradient and employed Langevin dynamics for iterative sample generation as an alternative to likelihood-based training. A practical and scalable discrete-time formulation was then introduced with the Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020). Subsequently, (Song et al., 2020) unified these approaches within a continuous-time stochastic differential equation (SDE) framework. Finally, Latent Diffusion Models (LDMs) (Rombach et al., 2022) improved efficiency by operating in a compressed latent space, making high-resolution image synthesis tractable.

In parallel, research has also focused on conditioning mechanisms that guide diffusion models toward desired structures or semantics. By incorporating external information – such as textual prompts, visual features, or spatial priors – into the denoising process, these models enable controllable generation with fine-grained control over both content and layout.

Conditioning Mechanisms. Semantic conditioning provides high-level control over the content and style of the generated image, typically implemented through cross-attention mechanisms (Rombach et al., 2022) that inject contextual embeddings, such as text or visual features (Ye et al., 2023), into the denoising network. In contrast, structural conditioning constrains the spatial layout of the generation process by injecting explicit priors into the diffusion trajectory. Methods such as SDEdit (Meng et al., 2022) and ControlNet (Zhang et al., 2023) achieve this by providing structural inputs – such as contours, depth, or segmentation maps – that anchor the denoising process to a given geometry. Our work builds on this line by jointly leveraging domain-aware textual prompts for semantic guidance and a unified control representation for spatial guidance, enabling efficient dual-conditioning while preserving both semantic coherence and structural fidelity.

Controlled Generation with Structured Priors. Diffusion models have also been explored for structured vision tasks such as semantic segmentation and object detection. Some studies exploit the semantic understanding captured by cross-attention maps to directly infer segmentation masks from text-to-image diffusion models, as shown in (Wang et al., 2023) and (Kawano and Aoki, 2024). Other approaches instead focus on data generation with preserved annotations. Layout-to-image frameworks condition the diffusion process on explicit spatial priors such as Canny edges, depth, or segmentation masks to synthesize images consistent with a desired layout. Among these, (Fang et al., 2024) introduced a ControlNet-based diffusion framework for object detection data augmentation, where image synthesis is guided by spatial priors and a CLIP-based filtering step ensures alignment between generated content and target class, thereby enhancing dataset quality and semantic fidelity.

Domain Adaptation and Prompt Control. Several diffusion-based approaches leverage image editing or style transfer to mitigate domain shift while preserving spatial structure. For instance, (Huang et al., 2024) proposed *BlenDA*, a diffusion-based method for unsupervised domain adaptation in object detection, blending source images with target-like translations generated by InstructPix2Pix using domain text prompts. While such approaches effectively reduce the source–target gap, they may introduce subtle variations in texture or lighting that affect annotation accuracy.

Text prompts play a central role in semantic conditioning, guiding the model toward specific concepts,

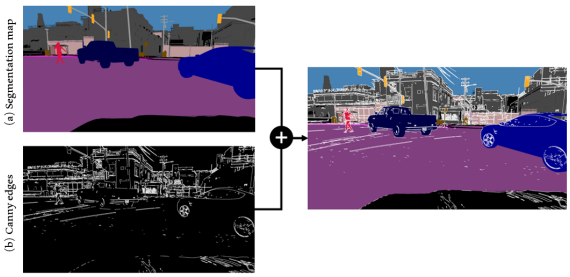

Figure 2: The chosen control form combines segmentation maps with Canny edges to inject both global semantic features and low-level detail features during the diffusion process. Data augmentation techniques are performed on control images.

styles, or domains during generation. Recent research has explored various strategies to enhance this control through prompt-based personalization. Textual Inversion (Gal et al., 2022) learns lightweight token embeddings that capture novel visual concepts, allowing them to be reintroduced in prompts to reproduce particular objects or artistic styles. DreamBooth (Ruiz et al., 2022) achieves similar personalization by fine-tuning the model on a few examples of a target subject, associating it with a unique token that enables subject-specific synthesis. Building on this principle, DATUM (Benigmim et al., 2023) extends DreamBooth to one-shot domain adaptation for semantic segmentation, where domain-specific fine-tuning enables a diffusion model to generate data consistent with the target visual domain while preserving semantic structures.

Inspired by these works, our method introduces *domain-aware prompting*, where domain-indicative tokens are incorporated into text prompts during joint fine-tuning of the ControlNet on both domains. This enables the model to associate textual domain cues with visual appearance variations, producing translations consistent with the target domain.

3 METHOD

In this section, we present our approach for synthetic-to-real style transfer based on latent diffusion models.

3.1 Conditional Control with Dual Representations

ControlNet augments pretrained text-to-image diffusion models with spatially localized, task-specific conditioning. This is achieved by injecting additional control signals such as e.g., edges, segmentation, depth. In our case, Stable Diffusion XL serves

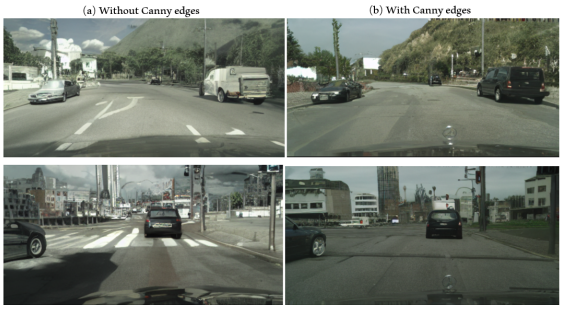

Figure 3: Initial experiments of generation results after training ControlNet with and without Canny edges.

as the backbone generator, while ControlNet provides the structural conditioning pathway.

For our use case, preserving the semantic layout of synthetic data is essential: any distortion in object placement introduces annotation noise, undermining the value of synthetic labels. Among the control signals available, segmentation maps are particularly useful, as they encode both position and class information. However, we observed that training solely on segmentation maps led to poor reconstruction of fine details with unrealistic building shapes, hallucinated road markings, and lighting and texture patterns often affected by noticeable artifacts (see Figure 3). To address this, we introduce a second control form based on Canny edges, which provide a complementary low-level structure. The standard ControlNet design uses a separate branch for each control signal, which increases the number of trainable parameters. When fine-tuning on a moderate dataset of 30k images, this overparameterization can lead to overfitting and reduced generalization. To mitigate this, we fuse both control forms into a single composite conditioning input, enabling one ControlNet to process different structural cues jointly. This unified design not only improves robustness and reduces memory usage, but also lowers computational overhead by requiring a single forward pass instead of multiple control branches.

Our approach thus employs a *single fused control image* combining segmentation maps and Canny contours, as illustrated in Figure 2. This composite conditioning preserves different structural information within one representation. As shown in Figure 3, this fusion substantially reduces hallucinated details and improves the consistency of object boundaries. Our underlying hypothesis is that the conditioning inputs (e.g., segmentation maps or edges) encode domain-invariant structure, while complementary mechanisms handle domain-specific appearance control; this separation ensures that geometric fidelity is preserved by the control signal, preparing the

ground for the domain-aware prompting introduced next.

3.2 Domain-Aware Prompting

We extend ControlNet with a lightweight prompt modification strategy that we call *domain-aware prompting*. For each image, a conditioning prompt p_x is constructed by combining the textual scene description with a domain token ("synthetic" or "real"). This allows the model to learn domain-specific style features associated with each token.

Formally, given a synthetic image x_a with control map c_a and description d_a, we form the prompt

$$p_a = \text{"a synthetic picture of "} + d_a.$$

For real images x_b, we form

$$p_b = \text{"a real picture of "} + d_b.$$

Training then reduces to a supervised denoising objective over both domains:

$$\mathcal{L}_{LDM} = \mathbb{E}_{x,\varepsilon} \left\| \varepsilon - \varepsilon_\theta(x, c_x, p_x) \right\|^2,$$

where ε_θ denotes the noise prediction network of the diffusion model.

At inference, translation is achieved simply by *swapping the domain token*. For example, a prompt with "synthetic" is replaced by "real", while keeping the same control map and description. This forces the model to generate the corresponding real-style image while preserving structure.

3.3 Iterative Refinement Loop

Despite strong results, initial translations may still contain artifacts. To improve robustness, we introduce an iterative refinement loop inspired by self-training (Zou et al., 2019). After each training round, the model is used to translate a batch of unseen synthetic images into the real domain. Generated images are relabeled with an additional token ("artifacted") and reintroduced into the training set. This enables the model to learn to avoid reproducing such artifacts in future rounds. Over successive iterations, the model progressively improves translation quality and realism.

4 EXPERIMENTS

We evaluate our method on synthetic-to-real style transfer for semantic segmentation. Our experiments are designed to answer three main questions: (i) does domain-aware prompting combined with dual-control images reduce the domain gap in distribution space? (ii) does the generated data improve downstream segmentation performance compared to synthetic-only training? (iii) how does our approach compare with state-of-the-art style transfer and domain adaptation methods?

4.1 Training Details

Datasets. We use datasets from the ICCV 2017 Visual Domain Adaptation (VisDA) challenge (Peng et al., 2017). The source domain consists of 25,000 synthetic images from GTA (Richter et al., 2016) with segmentation annotations, and the target domain consists of 3,000 annotated and 2,000 unannotated images from Cityscapes (Cordts et al., 2016). In accordance with the VisDA (Peng et al., 2017) protocol, the unannotated subset—corresponding to the test split—is used only in an unsupervised manner during training. Specifically, we employ these images solely for structural conditioning via Canny edge maps, without accessing their labels. For annotated target samples, both segmentation maps and edge conditioning are used.

Image Captioning. Text prompts are also important for the generation since they can provide multiple details about the scene or the target image. Beyond the objects' positions, and details, the model cannot infer other information like the weather, time of day (day or night), etc., solely from the control image. Thus, additional conditioning using text prompts is necessary. Another utility of using prompts is that we can alter the generated objects' appearances, such as colors, which is required in our use case to augment the data and prevent overfitting. Since annotating 30k images manually is not feasible, we opted for automatic captioning models such as the BLIP model (Li et al., 2022). We provided the model with the image along with a small prompt "a picture of" so that BLIP could generate the rest of the caption. As illustrated in Figure 4, the dataset is organized into paired image–control entries, each linked to a text prompt. These relationships are defined in a JSON metadata file that maps each image to its corresponding conditioning input and description.

Data Augmentation. To prevent overfitting, we randomly remove segmentation maps (leaving only edges), or edges (leaving only segmentation maps). To encourage robustness, we adopt a partial classifier-free guidance strategy. During training, 20% of text prompts have their scene descriptions removed, leaving only the domain tokens (e.g., synthetic" or real"). This allows the model to learn both fully conditioned

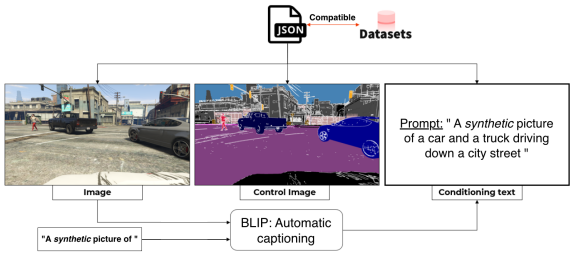

Figure 4: Illustration of our training set construction pipeline.

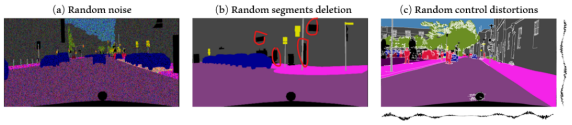

Figure 5: Data augmentation techniques on control images: (a) adding random Gaussian noise, (b) deleting random segments from the image for more generation robustness, (c) distorting the control images.

and partially unconditioned mappings. We also employed simple augmentation techniques, as illustrated in Figure 5. These techniques include adding random Gaussian noise to the segmentation maps, randomly removing some segment colors to force the model to infer the object based only on its shape and the surrounding context, and distorting the control image to handle unclear or noisy contours. Each domain was assigned a fixed pair of Canny edge thresholds. During augmentation, we applied small random perturbations around these preset values to simulate realistic variability while preserving domain-specific characteristics.

Iterative Refinement. Training is conducted in multiple rounds. In each round, the model translates a batch of synthetic images to the real domain using domain-aware prompting. Translated samples with artifacts are reintroduced into training with an additional "artifacted" token, enabling the model to learn to avoid such artifacts over time.

In this implementation, and for the sake of simplicity, we assume that all generations produced after a given training round still exhibit deviations from the true target-domain distribution. Consequently, instead of filtering or manually selecting faulty samples, we annotate the entire set of generated images from each round with the "artifacted" token. This strategy encourages the model to learn a refined distinction between the target domain and its current generative distribution. Over successive iterations, this process implicitly redefines the decision boundary between domains, allowing the diffusion model to progressively align its output distribution with that of real target images.

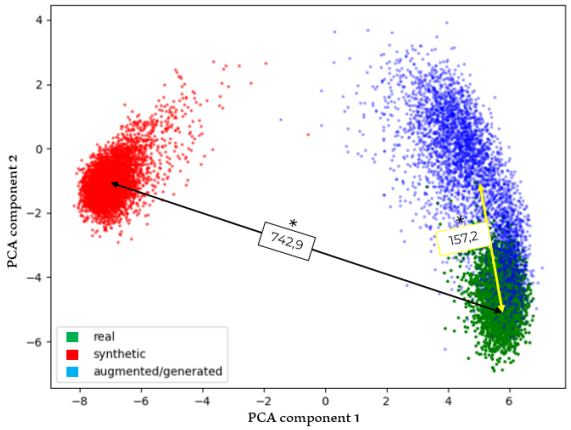

Figure 6: Semantic Similarity Analysis: The translated source images exhibit a closer alignment with the real image distribution. The computed distance is measured using the Maximum Mean Discrepancy (MMD) on CLIP (Jayasumana et al., 2024) embeddings prior to projection.

Implementation Details. All experiments use ControlNet built on top of the Stable Diffusion XL (SDXL) (Podell et al., 2023) backbone. Training is performed with a learning rate of 4×10^{-5}, a batch size of 2, and 10 gradient accumulation steps. Images are resized to 512×1024. On an A100 GPU (40GB), one round of training with 20,000 steps takes under 8 hours.

4.2 Analysis of Generated Images

To evaluate the semantic alignment of our translated images, we adopt the CLIP-based Maximum Mean Discrepancy (CMMD) metric (Jayasumana et al., 2024). CMMD computes the Maximum Mean Discrepancy (MMD) between CLIP (Radford et al., 2021) embeddings of real, synthetic, and translated images, providing a perceptually grounded measure of distributional similarity that better reflects visual semantics than traditional FID. Unlike FID, which is sensitive to image degradations and can vary non-monotonically with visual quality, CMMD exhibits a stable, monotonic behavior consistent with human perceptual judgments. Figure 6 shows that our translations significantly reduce the domain gap, with a 4.7× reduction in MMD compared to the original synthetic data.

We also examine the diversity and controllability of our translations by varying prompt attributes, such as vehicle colors, and sampling with different random seeds. This test assesses how well the model follows prompt modifications when translating into the real domain, while maintaining the semantic layout imposed by the control maps. As shown in Figure 7, the model reliably reflects textual changes without al-

Figure 7: Generation samples under different prompts and seeds. Objects change appearance while the scene layout remains consistent.

tering scene geometry. Remarkably, despite relying on a simple, automatically generated captioning strategy during training, the model still preserves a strong linkage between the text description, the image appearance, and the underlying layout, enabling realistic and controllable data augmentation.

To complement our evaluation of the generated images, we assessed their quality from a texture-oriented perspective within a distortion-aware embedding space. In this space, images with similar texture or degradation patterns are placed close together, irrespective of their semantic content. We used ARNIQA (leArning distoRtion maNifold for Image Quality Assessment) (Agnolucci et al., 2024) to extract such embeddings. ARNIQA learns a manifold of common image distortions – such as blur, noise, and compression artifacts – using deep features whose distances correspond to perceptual similarity in texture and degradation. This makes it particularly suited for evaluating texture fidelity independently of semantic information, offering a complementary measure to content-driven metrics.

In addition to distortion embeddings, ARNIQA provides a regression module that maps each embedding to an image quality score trained from human perceptual judgments. When applying this regressor to our generated images, the overall score distribution centered around 0.7, indicating generally good quality. However, because the regressor reflects subjective human preferences, it tends to penalize blur or sensor noise even when such characteristics are realistic. In our case, these properties are consistent with target-domain imagery. For instance, as shown in Figure 9, ARNIQA assigned a score of about 0.3 to a generated image (left), despite it accurately replicating the appearance of real dashcam images (right), which naturally contain blur and compression artifacts. Thus, although the regressor assigns lower scores, these images remain desirable for reducing domain shift by matching the target distribution.

To more effectively exploit ARNIQA, we therefore prioritized the analysis of distortion embeddings directly rather than relying solely on the regressor's

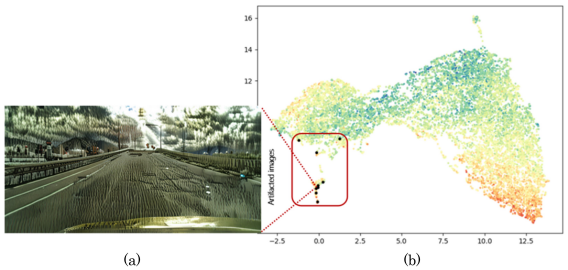

Figure 8: (a) Generated artifacted image on ARNIQA distortion embeddings projection (b): Small cluster of artifacted generations, colored by the attributed quality score.

Figure 9: ARNIQA quality scores for two examples from the real and generated datasets. The regressor exhibits a subjective bias, assigning lower scores to noisier images.

outputs. Similar to our use of CLIP embeddings (Radford et al., 2021), we projected the ARNIQA embeddings into a lower-dimensional space using PCA and identified clusters corresponding to poor-quality generations as shown in Figure 8. These clusters contained images with texture-related artifacts, although they did not always coincide with the lowest regressor scores, further illustrating the regressor's subjectivity. The frequency of such low-quality generations was approximately 1 percent, as only 128 out of 12.5k images exhibited these distortions. Nonetheless, we elected to filter this cluster from the dataset to prevent potential negative effects on downstream models trained with this data.

Comparison with Style Transfer Models We compare against CycleGAN-Turbo (Parmar et al., 2024), a hybrid GAN–diffusion model based on Stable Diffusion Turbo, trained using the authors' official scripts. As illustrated in Figure 10, while both models successfully learned to translate the synthetic images into a more realistic style, our method better preserves semantic layouts, while CycleGAN-Turbo often alters background objects or hallucinates structures despite parameter tuning.

4.3 Comparison with Domain Adaptation Approaches

To evaluate downstream benefits, we train DeepLabv3 on translated data and test on Cityscapes. While our

Figure 10: Visual comparison of CycleGAN-Turbo (Parmar et al., 2024) and our approach, both trained on our datasets. Our method preserves layout and annotations while improving realism.

approach follows the spirit of domain adaptation, it does not fall under the purely unsupervised setting: the generative model was fine-tuned using target-domain annotations to better preserve semantic structures in the translated images. Nevertheless, since the objective remains to bridge the source–target gap for improved segmentation on real data, we compare our method against leading feature level domain adaptation baselines from the VisDA challenge (Peng et al., 2017), including AdaptSegNet (Tsai et al., 2018). Table 1 reports the corresponding results, where our method achieves higher mean IoU across most classes with faster convergence. Note that lower IoU for classes such as *bus*, *truck* and *wall* is consistent with the strong class imbalance of Cityscapes. (Hoyer et al., 2023)

Table 2 reports the segmentation performance obtained when training a DeepLabv3 segmenter on datasets generated by different style transfer methods. For our experiments, we evaluate two backbone configurations – ResNet-50 and ResNet-101 – to ensure consistency with prior work. Three variants of CycleGAN-Turbo (Parmar et al., 2024) are included, each trained with a distinct identity-loss weight λ_{Idt} that balances style fidelity and content preservation. Our ControlNet-based diffusion model consistently outperforms these variants, yielding higher mean IoU and producing translations that better preserve structural coherence.

For reference, we also compare with CyCADA (Hoffman et al., 2017), a cyclic GAN framework explicitly designed to preserve semantic layouts. The original CyCADA results were reported using a DeepLab-ResNet-101 backbone; we therefore report both of our configurations (R50 and R101) for a fair qualitative comparison. Across both backbones, our method achieves up to a +6-point improvement in mean IoU over GAN-based baselines. The benefit is particularly evident for the *sky* class, where CycleGAN-Turbo attains 0.565 IoU and CyCADA 0.690, compared to 0.864 and 0.916 for our model with ResNet-50 and ResNet-101 respectively. These results indicate that our diffusion-based translation preserves semantic integrity and minimizes the annotation distortions often introduced by cyclic GANs.

Figure 11 provides qualitative segmentation results. Segmenters trained on our translated images achieve smoother and more accurate predictions compared to those trained on source (GTA dataset) and GAN-generated data. As shown in Figure11, the segmenters trained on CycleGAN-generated data struggle with certain classes, such as sidewalks (in pink). In contrast, the segmenter trained on our generated data mostly produces cleaner and more accurate segmentations, closely matching the ground truth annotations. The black labels in the annotation maps represent unlabeled classes that we ignored in this simplified study to benchmark the segmenter's performance on the key classes most prevalent in the dataset. It is worth noting that the training scripts for all cases, as well as the model architecture, remained identical, with no additional techniques applied when training on our data. Additionally, the dataset sizes were equal, as we consistently transformed the entire synthetic dataset (25k images) with each generative model.

4.4 Efficiency Analysis

Beyond accuracy, training efficiency is critical. CycleGAN-Turbo requires 38 hours on cropped 256×256 images to complete 20,000 steps, while our model trains in under 8 hours on full-resolution 512×1024 images. This highlights the advantage of leveraging ControlNet fine-tuning instead of training large generative-discriminative models.

4.5 Ablation Study

To quantify the contribution of each design choice, we conduct an ablation study. Table 3 reports segmentation performance when training without iterative refinement, with only segmentation maps, or with only edges. Removing either component degrades performance, confirming that both dual-control conditioning and iterative refinement are necessary.

5 SUMMARY AND OUTLOOK

This work presented a diffusion-based framework for synthetic-to-real style transfer that unifies spatial and semantic conditioning through dual-control

Table 1: Segmentation mean IoU comparison with feature level domain adaptation methods. DeepLabv3 with a ResNet101 backbone was trained on translated datasets generated by our diffusion model, which was fine-tuned using target-domain annotations to enhance translation quality.

Method	Building	Bus	Car	Person	Road	Sidewalk	Sky	Terrain	Truck	Veg	Wall	Mean	FW-Mean
AdaptSegNet (Tsai et al., 2018)	0.799	0.354	0.737	0.585	0.865	0.36	0.756	0.333	**0.325**	0.834	0.234	0.562	0.702
Microsoft research Asia team (Zhang et al., 2018)	0.747	0.251	0.80	0.559	0.870	0.385	0.904	0.326	0.234	0.721	**0.237**	0.548	0.687
University of Oxford team (Peng et al., 2017)	0.534	**0.364**	0.755	0.578	0.853	0.424	0.617	0.273	0.275	0.771	0.173	0.511	0.635
University of California team (Peng et al., 2017)	0.702	0.180	0.780	0.470	0.872	0.333	0.903	0.286	0.248	0.772	0.136	0.516	0.679
Ours	**0.889**	0.216	**0.848**	**0.588**	**0.912**	**0.562**	**0.916**	**0.335**	0.153	**0.863**	0.13	**0.588**	**0.763**

Table 2: Mean IoU comparative results of the segmentation score on different generated datasets.

Model	Backbone	Building	Bus	Car	Person	Road	Sidewalk	Sky	Terrain	Truck	Veg	Wall	Mean	FW-Mean
CycleGAN-Turbo (Parmar et al., 2024) $\lambda_{idt} = 1$	R50	0.801	**0.260**	0.788	**0.521**	0.883	0.441	0.519	0.263	0.107	0.739	0.195	0.502	0.694
CycleGAN-Turbo $\lambda_{idt} = 1.5$	R50	0.826	0.215	0.808	0.456	0.909	0.397	0.590	0.239	0.121	0.728	0.187	0.498	0.706
CycleGAN-Turbo $\lambda_{idt} = 2$	R50	0.821	0.207	0.778	0.481	0.903	0.446	0.586	0.241	0.102	0.748	**0.202**	0.501	0.635
Ours	R50	**0.860**	0.212	**0.810**	0.501	**0.935**	**0.586**	**0.864**	**0.320**	**0.146**	**0.831**	0.158	**0.565**	**0.758**
CyCADA (Hoffman et al., 2017)	R101	0.779	**0.256**	0.745	**0.628**	0.791	0.331	0.69	0.267	**0.209**	0.815	**0.234**	0.5222	0.6646
Ours	R101	**0.889**	0.216	**0.848**	0.588	**0.912**	**0.562**	**0.916**	**0.334**	0.153	**0.863**	0.13	**0.588**	**0.7635**

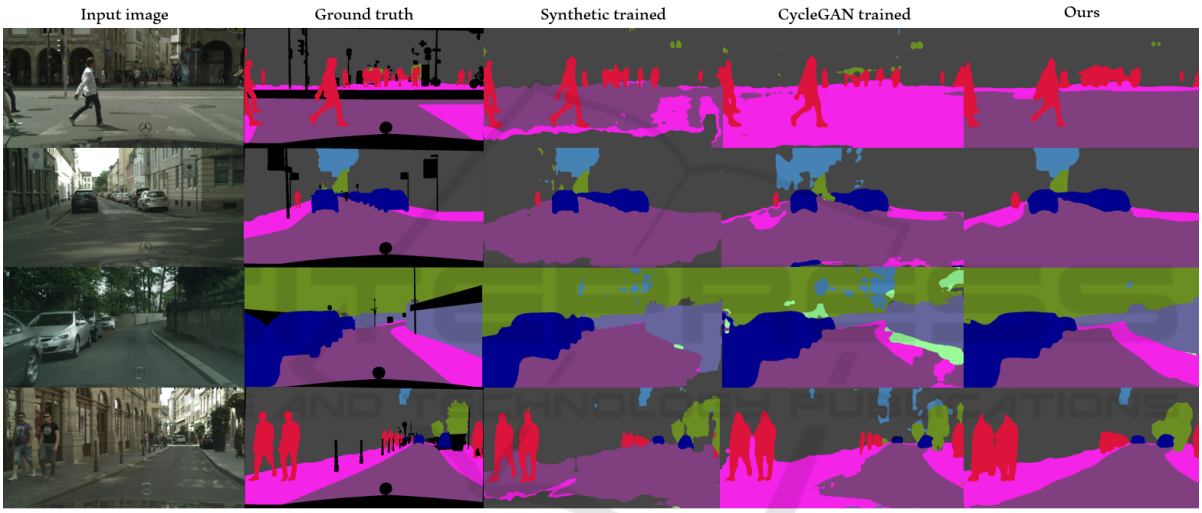

Figure 11: Visual comparison of the segmentation inference results when trained on the different datasets. DeepLabv3 with a ResNet-50 backbone was used as the segmenter.

Table 3: Ablation study on the impact of iterative refinement and control maps for ControlNet training.

Configuration	Backbone	Building	Bus	Car	Person	Road	Sidewalk	Sky	Terrain	Truck	Veg	Wall	Mean	FW-Mean
No it. Refinement	R50	0.852	0.144	0.823	0.500	0.923	0.522	0.651	0.240	0.102	0.756	0.144	0.514	0.734
Segmentation only	R50	0.856	0.170	0.809	**0.527**	0.859	0.399	0.798	0.237	0.109	0.816	0.131	0.519	0.714
Canny only	R50	0.855	0.125	**0.824**	0.503	0.889	0.449	0.668	0.164	0.135	0.726	0.137	0.498	0.713
Ours Complete	R50	**0.860**	**0.212**	0.810	0.501	**0.935**	**0.586**	**0.864**	**0.320**	**0.146**	**0.831**	**0.158**	**0.565**	**0.758**

inputs and domain-aware prompting. The proposed strategy enables consistent generation across domains while maintaining realistic details and controllable structure. A progressive refinement loop further enhances visual quality and reduces artifacts, demonstrating that diffusion-based control mechanisms can effectively bridge domain gaps.

Experiments on the GTA-to-Cityscapes benchmark demonstrate the effectiveness of this framework. We focus on this benchmark because it combines complex urban geometry with dense pixel-level labels, making it particularly suitable for assessing annotation preservation under synthetic-to-real translation, while leaving broader cross-dataset validation to future work. Models trained on data augmented with our synthetic-to-real translations achieved consistently higher mean IoU with fewer resources. The generated images align closely with real data in CLIP embedding space and outperform GAN-based baselines such as CycleGAN-Turbo (Parmar et al., 2024) and CyCADA (Hoffman et al., 2017) in both realism and downstream segmentation accuracy. While our

method relies on semantic annotations from the target domain – representing a supervised form of domain adaptation – it provides a practical and scalable balance between control, stability, and visual fidelity.

Future work will focus on reducing the reliance on target-domain annotations to move toward fully unsupervised adaptation, for example by leveraging foundation models for automatic mask generation or weak supervision, such as SAM-like segmentation frameworks (Kirillov et al., 2023). Beyond urban scenes, the same dual-control diffusion principle could be extended to other domains that provide structured priors, such as medical imaging, robotics simulation, or satellite imagery. Furthermore, integrating domain generalization objectives, such as regularizing latent representations across different synthetic domains or styles, may further enhance robustness. Future work will also investigate the role of textual conditioning more deeply. In this paper, we deliberately adopted a simple BLIP-based captioning strategy, as our primary objective was to use prompts mainly as a vehicle for the domain-specific token driving the style transfer. Despite this minimalist approach, the model successfully learned to associate textual cues with visual appearance while preserving the underlying layout (Figure 7, 12). However, the prompting mechanism could be significantly strengthened. One direction is to introduce richer or perturbed prompts, or to learn domain embeddings through techniques such as textual inversion or lightweight fine-tuning, enabling the model to better disentangle content from style. Another promising avenue is to make captions more structured and exhaustive: instead of free-form captions, the captioning model could be queried with predefined questions—e.g., What is the weather? Is it day or night? Which objects are present and what are their attributes?—and the resulting answers concatenated into a detailed scene description. This would provide the diffusion model with a clearer and more expressive semantic signal, potentially improving controllability and reducing ambiguity during translation.

Figure 12: Additional generation samples using our model.

ACKNOWLEDGEMENTS

We would like to express our deep gratitude to the members of our team for their valuable feedback and the enriching discussions that have greatly contributed to the advancement of this project. This research was done during the E2CC project. This Project supported by the EC in the framework of its Programme on Next Generation Cloud Infrastructure and Services (IPCEI-CIS). Aid granted under the grant contracts between Bpifrance and the parties in relation to the IPCEI -CIS E2CC PLATFORM, in accordance with the decision of the Prime Minister dated 12/04/2024. Competent authority: Ministry of Economics, Finance and Industrial and Digital Sovereignty, France.

REFERENCES

Agnolucci, L., Galteri, L., Bertini, M., and Del Bimbo, A. (2024). Arniqa: Learning distortion manifold for image quality assessment. In *Proceedings of*

the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 189–198.

Benigmim, Y., Roy, S., Essid, S., Kalogeiton, V., and Lathuilière, S. (2023). One-shot unsupervised domain adaptation with personalized diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 698–708.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ..., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*.

Fang, H., Han, B., Zhang, S., Zhou, S., Hu, C., and Ye, W.-M. (2024). Data augmentation for object detection via controllable diffusion models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1246–1255.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models.

Hoffman, J., Tzeng, E., Park, T., Zhu, J. Y., Isola, P., Saenko, K., ., and Darrell, T. (2017). Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Hoyer, L., Dai, D., Wang, Q., Chen, Y., and Van Gool, L. (2023). Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. *International Journal of Computer Vision*, 131:2070–2096.

Huang, T., Huang, C.-C., Ku, C.-H., and Chen, J.-C. (2024). Blenda: Domain adaptive object detection through diffusion-based blending. *arXiv preprint arXiv:2401.09921*.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2018). Image-to-image translation with conditional adversarial networks.

Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., and Kumar, S. (2024). Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kawano, Y. and Aoki, Y. (2024). Maskdiffusion: Exploiting pre-trained diffusion models for semantic segmentation. *arXiv preprint arXiv:2403.11194*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. (2022). Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*.

Parmar, G., Park, T., Narasimhan, S., and Zhu, J. Y. (2024). One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*.

Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., and Saenko, K. (2017). Visda: The visual domain adaptation challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ..., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Richter, S. R., Vineet, V., Roth, S., and Koltun, V. (2016). Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 102–118.

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

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2022). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*.

Sauer, A., Lorenz, D., Blattmann, A., and Rombach, R. (2024). Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer.

Scheck, T., Grassi, P., A., and Hirtz, G. (2021). Unsupervised domain adaptation from synthetic to real images for anchorless object detection. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, pages 319–327.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 2107–2116.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2019). Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, pages 11918–11930.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Tsai, H., Y., Hung, C., W., Schulter, S., Sohn, K., Yang, H., M., Chandraker, and M. (2018). Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7472–7481.

Wang, J., Li, X., Zhang, J., Xu, Q., Zhou, Q., Yu, Q., Sheng, L., and Xu, D. (2023). Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*.

Ye, C., Zhang, J., Liu, P., Zhang, J., Sun, X., Gao, Y., Liu, Z., Liu, Z., Zhang, L., Jin, Q., and Li, H. (2023). Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zhang, L., Rao, A., and Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Zhang, Y., Qiu, Z., Yao, T., Liu, D., and Mei, T. (2018). Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6810–6818.

Zhu, J. Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zou, Y., Yu, Z., Liu, X., Kumar, B. V. K., and Wang, J. (2019). Confidence regularized self-training. In *In Proceedings of the IEEE/CVF international conference on computer vision*, pages 5982–5991.