2025 International Joint Conference on Neural Networks (IJCNN) | 979-8-3315-1042-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/IJCNN64981.2025.11228158

Enhancing Synthetic Image Realism with Controlled Diffusion Models

Iqra Nosheen[1], Muhammad Ali Farooq[2], Peter Corcoran[2], Cathy Ennis[3], Michael G. Madden[1]
[1]School of Computer Science, University of Galway, Galway, Ireland
[2]School of Engineering, University of Galway, Galway, Ireland
[3]School of Computer Science, Maynooth University, Dublin, Ireland
{i.nosheen1, muhammadali.farooq, peter.corcoran}@universityofgalway.ie
cathy.ennis@mu.ie, michael.madden@universityofgalway.ie

***Abstract*—In this work, we present an innovative approach utilizing ControlNet-based diffusion models along with upscaling capabilities for domain adaptation and quality refinement of 3D modelled synthetic datasets, focusing on autonomous vehicle applications. A significant domain gap often exists between synthetic and real-world data, hindering the applicability of deep learning models trained on synthetic data for real-world scenarios. Our methodology leverages the strengths of Controlled Augmentation by simultaneously utilizing multiple ControlNet signals, including edge detection, depth information, segmentation maps, and tile resampling. To improve how synthetic data aligns with the desired domain specifications, these signals guide the generative process, and we also incorporate text-guided prompts extracted via Large Language Models (LLMs), to improve control over the synthesis of desired features and attributes. We test the approach on diverse environmental conditions from the VKITTI dataset, a well-known 3D modelled synthetic dataset generated in Unity for autonomous driving research. The refined data is validated using quantitative metrics including FID, SSIM, and LPIPS, and is also evaluated on downstream machine learning tasks of object detection and classification, using YOLO-v8 to ensure its utility and effectiveness. Experimental analysis demonstrates the effectiveness of this method in improving the realism and usability of synthetic data. Our approach contributes to fields that require high-quality data synthesis and domain adaptation. The experimental work, along with ControlNet models used in this project is available online.[1]**

***Index Terms*—Diffusion models, domain adaptation, image realism, synthetic data, autonomous vehicles**

I. Introduction

In many real-world applications such as autonomous vehicles, the acquisition of datasets presents challenges due to constraints related to cost, time limitations, and restricted accessibility [1]. To mitigate these challenges, the generation of synthetic data through advanced 3D modelling techniques has emerged as a pragmatic alternative. This synthetic data, produced in a controlled virtual environment, can be used to train deep learning models. However, a domain gap often exists between synthetic and real-world data, which affects the performance of these models. This domain gap can limit the direct applicability of models trained on synthetic data to real-world scenarios. Eliminating this gap is essential to realise the potential of deep learning models.

To bridge this domain gap and enhance the realism of virtual data, several techniques have been proposed. The techniques can be broadly divided into three types; domain adaptation, domain randomization and photo-realistic rendering [2]. Domain adaptation involves aligning the distributions of synthetic data to make it look more similar to real-world data. Domain randomization involves adding much diversity in the data so that the real-world data becomes the subset of synthetic data. Photo-realistic rendering uses advanced graphics techniques to create realistic virtual environments. Traditionally, GAN-based approaches are commonly used to improve the adaptability of models by altering the style of source images to target domain. However, these approaches face various limitations. Firstly, these approaches often fail to preserve local details of the source data and results in semantic inconsistencies between the refined images and their corresponding labels [3]. Secondly, these approaches may struggle to generate high-fidelity and realistic data in diverse weather and illumination conditions. Thirdly, GAN-based methods typically require training from scratch on task-specific datasets, which limits their ability to generalise across diverse scenarios [4]. Fourthly, these approaches often exhibit training challenges including instability and mode collapse [5].

Recently, diffusion models have gained popularity in generating high-fidelity data. These models are trained on large datasets and can be fine-tuned for specific tasks. For example, Stable Diffusion can be fine-tuned with DreamBooth for a specific application. Diffusion models are a promising alternative for various tasks that have been tackled with GAN-based methods, such as image restoration, image generation, image editing and style transfer. However, converting a data distribution across different domains using diffusion models is challenging, since standard diffusion models typically convert from a Gaussian distribution rather than a specific domain distribution. It is crucial to maintain the semantics of the source domain data during conversion, to ensure accurate transformation to the target domain. To address these issues, our approach combines multiple ControlNet annotators, text-

Supported by the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) Grant No. 18/CRT/6224, Taighde Eireann Grant IRCLA/2023/1992, and ADAPT - Centre for Digital Content Technology.

[1]https://github.com/MAli-Farooq/Enhancing-Synthetic-Image-Realism

Authorized licensed use limited to: Maynooth University Library. Downloaded on November 07,2026 at 15:24:40 UTC from IEEE Xplore. Restrictions apply.



Fig. 1: Refined image results using ControlNet conditioning and SR-based upscaling methodology explained in Algorithm 1. Each block shows input (top row) and refined images (bottom row) for Fog, Overcast, Rain, and Sunset weather conditions.

guided prompts, and a super-resolution up-scaling operation to enhance the quality of synthetic image database as demonstrated in Figure 1. This preserves semantic structures effectively, while enhancing image quality and the realism of virtual synthetic data, particularly in challenging weather conditions. The core contributions of this paper are as follows:

1) We propose an approach that uses Controlled Augmentation by employing multiple ControlNet signals, including Canny edges, semantic maps, tile adaptation, and additional guidance in the form of text prompts for each weather condition, to generate refined images based on the structure and semantics of synthetic source images.
2) We apply a second-stage upscaling technique using the R-ESRGAN 4x super-resolution algorithm, to further improve image quality.
3) We perform extensive experiments on benchmark datasets to demonstrate the effectiveness of the proposed approach. Results indicate that the generated samples closely resemble the target samples and reduce the domain gap between the source and target domains.

II. Background/ Literature Review

A. GAN-Based Approaches for Domain Adaptation

To reduce the domain gap between synthetic and real world data, traditional approaches are focused on using generative models to adapt the real-world style in synthetic images [6]–[8]. These methods involve training GANs to translate synthetic images into more realistic ones by learning the underlying distributions of both domains. The DIVIDE framework [9] leveraged GAN-based adversarial training to create domain invariant geometric space for improving the realism of synthetic data. However, this framework struggles to make synthetic data appear realistic when the real-world scenes are significantly different form the synthetic training data. AugGAN-Det introduced by [10] used Cycle-object Consistency loss for instance-aware image-to-image translation to improve object detection across various weather conditions. Unlike previous methods that required pixel-level semantic segmentation, this approach used bounding box annotations, which are easier to acquire and ensure object preservation during translation. However, the framework faced difficulty in preserving very small objects in images. Regarding diverse weather conditions, different studies have used generative

Authorized licensed use limited to: Maynooth University Library. Downloaded on November 07,2026 at 15:24:40 UTC from IEEE Xplore. Restrictions apply.

models to synthesize clear-weather images into various challenging conditions such as rainy [11], foggy, snowy [12] and night-time scenarios [10]. The authors in [13] focused on reducing the domain gap by using a dynamic adaptation model and uncertainty-based sample selection for effective transfer learning across multi-source domains, thereby mitigating the impact of domain shifts. In [14], the authors proposed the Online Boosting Adaptive Learning method, which employed the AdaCOSA algorithm to handle covariate shift and concept drift to improve the prediction accuracy by dynamically adjusting the weights of relevant data across multiple domains. GAN-based approaches require access to large amount of target domain data for effective training of generative models. They often produce visible artifacts, such as colour anomalies [15] and face challenges in preserving the semantic details of source images during translation to the target domain [3], [16]. However diffusion models are pre-trained on large corpora of real world data, can be effectively fine-tuned on domain-specific tasks [17].

B. Diffusion Models for Domain Adaptation

Diffusion Models are generative models that are inspired by non-equilibrium thermodynamics [18]. They learn to denoise from a common distribution, such as Gaussian noise, to a specific data distribution. These models were proposed by [19] and have become popular for solving computer vision tasks [1], [3], [20]. Recent works like DreamBooth [21], Textual Inversion [22], and ControlNet [23] enabled personalised data generation with fine-tuning. In the context of domain adaptation, different studies [24], [25] have demonstrated the potential of diffusion models in aligning the source and target domains by capturing the underlying data distributions. [24] used pre-trained text-to-image diffusion models to create a synthetic target dataset based on target domain descriptions, while [26] adapted diffusion techniques to change image styles between domains. [25] developed a label-conditioned diffusion model to generate labelled data for the target domain.

ControlUDA [4] utilized large-scale text-to-image diffusion models with ControlNet to synthesize high-fidelity images under various weather scenarios. The study used real-world datasets as source and target domain and generated a dataset of different weather conditions using images from a clear weather to improve the semantic segmentation performance. ZoDi [17], a zero-shot domain adaptation method was proposed to synthesize target-like images from source images using layout-to-image diffusion models and stochastic inversion to maintain content and layout. It transformed clear weather images from the Cityscapes dataset into various adverse weather conditions such as nighttime, snowy, rainy, and foggy environments and trained segmentation model with these synthesized images to improve model's performance. Stable Diffusion combined with Low-Ranked Adapters (LoRA) and ControlNet was employed [27] to transform synthetic plant images into realistic ones. This method showed better results than CycleGAN, maintaining image geometry and realism. A data augmentation with diffusion models (DATUM) was proposed [24] that leveraged text-to-image diffusion models to generate synthetic target datasets that are stylistically and contextually aligned to the target domain. This used a single target image to fine-tune a pre-trained text-to-image diffusion model using multiple cropped regions to capture the target domain's appearance and generate new data.

Unlike existing studies which are focused on generating new data samples based on the source dataset, in this study we have used diffusion model to improve the realism of 3D modelled virtual synthetic data. We used the publicly available VKITTI dataset, which includes images with different weather conditions, and we experimented with each condition separately to assess the realism of diffusion-based refined images.

III. Proposed Methodology

This section details the adapted methodology in this work. Figure 2 shows the comprehensive block diagram representation for refining the synthetic datasets using Controlled Augmentation by employing ControlNet diffusion models as first stage processing and image super resolution as second stage processor.

A. Data Acquisition

For this study, we have utilized combination of synthetic and real-world datasets to thoroughly evaluate the effectiveness of our proposed approach for domain adaptation and quality refinement of autonomous vehicles datasets. For the source domain, the VKITTI 2 [28] dataset is used as a virtual synthetic data that is generated in Unity and comprises of 5 sequences of driving scenes with each having diverse weather conditions. For target domain, we have used the ACDC dataset [29] as a real-world dataset that provides data under multiple weather conditions including rain, fog, snow and nighttime. Additionally, we have also used the KITTI real-world dataset [30], a standard benchmark in the field of autonomous driving research, to assess the quality of refined overcast and sunset weather conditions.

B. ControlNet for Domain Adaptation and Data Refinement

The ControlNet-based methodology for domain adaptation and data refinement leverages conditional control mechanisms within neural networks to guide image generation and manipulation. It allows fine-tuned control over pre-trained models by introducing additional conditional inputs, such as images, text, and other data, which specify the desired characteristics or modifications. To integrate ControlNet with Stable Diffusion, the process begins by converting each input conditioning image (such as a depth map or edge map) from its original size of 512 × 512 into a 64 × 64 feature space vector compatible with Stable Diffusion. This is achieved using a small network $\mathcal{E}$ consisting of four convolutional layers, each with 4 × 4 kernels and 2 × 2 strides. The layers use ReLU activation and have 16, 32, 64, and 128 channels, respectively. The network is initialized with Gaussian weights and trained together with the full model. This network encodes the image-space condition $\boldsymbol{c}_\mathrm{i}$ into a feature space conditioning vector c_f, defined as:

$$c_\mathrm{f} = \mathcal{E}\left(\boldsymbol{c}_\mathrm{i}\right). \tag{1}$$

Authorized licensed use limited to: Maynooth University Library. Downloaded on November 07,2026 at 15:24:40 UTC from IEEE Xplore. Restrictions apply.

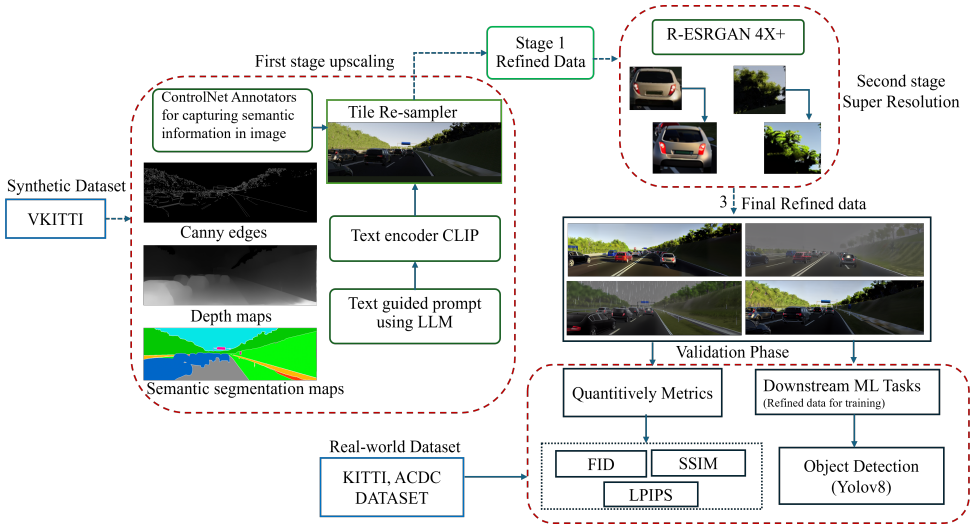


Fig. 2: Architecture of adapted synthetic up scaling methodology.

The conditioning vector c_{f} is subsequently fed into ControlNet for additional processing. In our proposed methodology as shown in Figure 2, in our first stage up-scaling block, we incorporate edge maps c_e, depth maps c_d, and segmentation maps c_s as input conditions for distinct environmental conditions to guide the domain adaptation and image refinement process on low resolution input image x_{LR}. These input modalities provide essential structural and contextual information, which enhance the model's ability to adapt to new domains and refine image details. Additionally, a tile resampling operation is applied to the input data, allowing for finer spatial resolution and improved consistency in the generated outputs. This first stage of ControlNet pipeline is explained in Algorithm 1 steps 1 - 4. Further, we have used text-guided prompts that are manually designed and refined via GPT-4o LLM which are then fed to the text encoder as an additional conditioning input. This enables the model to incorporate contextual information from textual descriptions, further refining the output and ensuring that the generated images align closely with the desired attributes. The main focus is to supplement additional details and refine the data quality of roadside objects using ControlNet. This includes enhancing textual features of objects such as cars, buses, vans, trucks, street poles, and trees lining the roadside. By incorporating ControlNet annotators as input conditions, we provide rich contextual information that aids in the accurate refinement of these objects. The input image is processed through encoding functions E_e, E_d, E_s, which are designed to extract and encode the relevant features from edge, depth, and segmentation maps. These encoded features are then resampled using a tile resampling operation to ensure consistent spatial resolution, resulting in a combined conditioning vector c. This vector is subsequently fed into ControlNet, enabling precise control over the adaptation and refinement processes, thereby improving the overall quality and detail of the generated data.

C. Upscaling using Super Resolution

In the second stage, the processed data using Controlled Augmentation is further up-scaled using single image super resolution. For this purpose, we have utilized Residual Enhanced Super-Resolution Generative Adversarial Networks (R-ESRGAN 4x+) as illustrated in Figure 2, in the second-stage Super Resolution block. R-ESRGAN 4x+ is an advanced super-resolution technique that enhances the quality of low-resolution images by increasing their resolution by a factor of four. It employs a generative adversarial network with residual-in-residual dense blocks, which helps in preserving finer details and textures while effectively reducing noise and artifacts.

Incorporating the second-stage super-resolution algorithm, R-ESRGAN 4x+ enhances the detail and refinement of objects in our system, including cars, buses, vans, trucks, street poles, and trees lining the roadside. This integration ensures that even subtle features are more accurately preserved, improving the overall fidelity of the generated images. The overall process is summarized in Algorithm 1. It can be observed that process executes by passing low-resolution input images (x_{LR}) through a five-stage refinement process, leveraging both visual and text-based conditioning. First, visual conditioning vectors for edge (c_e), depth (c_d), and segmentation (c_s) features are generated by encoding the input image using specific encoders (E_e, E_d, E_s) and resampling them with *TileResample* to ensure spatial consistency. Simultaneously, a text-guided conditioning vector (c_{text}) is created by encoding a text prompt (t) with a CLIP encoder. These visual and textual features are concatenated to form the combined conditioning vector (c_{final}), which is then input into *ControlNet* to refine the image and adapt it to the desired domain, producing x_{refined}. Finally, the refined image undergoes super-resolution enhancement using the *R-ESRGAN 4x+* network, resulting in a high-quality output image (x_{SR}). This method combines advanced encoding, domain adaptation, and super-resolution techniques to achieve superior synthetic image refinement.

With the combination of ControlNet for domain adaptation and data refinement, along with text-guided conditioning and further incorporating R-ESRGAN 4x+ for super-resolution, our approach achieves enhanced image quality and detail

Authorized licensed use limited to: Maynooth University Library. Downloaded on November 07,2026 at 15:24:40 UTC from IEEE Xplore. Restrictions apply.

in synthetic imaginary. As a result it helps in enhancing object detection and recognition accuracy, which is critical for autonomous vehicular applications. Figure 3 shows the refined images that are semantically consistent with their source images.

D. Validation of Refined Data

The refined data is validated using both quantitative metrics and downstream machine learning tasks. The quantitative metrics include Fréchet Inception Distance (FID), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). FID assesses the quality of images by comparing the distributions of features between real and generated datasets, with lower values indicating higher similarity and better image fidelity. SSIM evaluates the structural similarity between two images by comparing their luminance, contrast, and structure. LPIPS measures the perceptual similarity between images by comparing deep features extracted from neural networks and reflects human visual perception [4]. These metrics aim to ensure the quality of refined data, and alignment with human perception of image quality.

Algorithm 1 ControlNet Scaling and SR up-scaling Methodology for Synthetic Data Refinement

Input: Low-resolution input image x_{LR}

Steps: Five stage process

1) **Conditioning Vector:**
$$c_e = \text{TileResample}(E_e(c_e, x_{\text{LR}}))$$
$$c_d = \text{TileResample}(E_d(c_d, x_{\text{LR}}))$$
$$c_s = \text{TileResample}(E_s(c_s, x_{\text{LR}}))$$
2) **Text-Guided Conditioning Vector:**
$$c_{\text{text}} = \text{CLIP}(t)$$
3) **Combined Conditioning Vector:**
$$c_{\text{final}} = [c, c_{\text{text}}]$$
4) **Refined Image:**
$$x_{\text{refined}} = \text{ControlNet}(x_{\text{LR}}, c_{\text{final}})$$
5) **Super-Resolved Image:**
$$x_{\text{SR}} = \text{R-ESRGAN 4x+}(x_{\text{refined}})$$

Output: Super resolved image x_{SR}

IV. Experimental Results

The experimental work was carried out on a workstation machine equipped with A6000 GPU with video memory of 48 GB. PyTorch library was used to set up the environment.

A. Image Refinement Results

In the image refinement process, we utilized specific annotators that were tailored to each weather condition and further integrating it with the tile resampling technique helps to consistently delivers the high-quality results, ensuring that image detail are maintained across various environmental scenarios. For the sunset scenario, a Canny edge map was used as the ControlNet condition, guided by a text prompt from the GPT-4o engine as can be seen in 3 . Similarly, a depth map was used as the ControlNet condition for the overcast scenario, also guided by a text prompt.For challenging weather conditions which includes raining and foggy weather, we incorporate semantic segmentation as a control signal within ControlNet along with more customized short textual prompt according to the weather condition. This approach ensures that essential features, such as vehicles and environmental elements, remain distinct and represented precisely, despite adverse visual conditions. Further in this pipeline we have utilized the De-noising Diffusion Implicit Modelsimage sampler, selecting a final sampling step size of 32 to optimize the balance between image quality and computational efficiency. Finally the second stage upscaling operation, R-ESRGAN 4X+ model enhances the quality of images by utilizing its advanced generative adversarial architecture. It effectively reconstructs high-resolution images from low-resolution inputs, preserving fine details and textures while reducing artifacts in synthetic data. The image refinement results for all the weather are demonstrated in Figure 3 and Figure 1. The complete image refinement results for all the weather conditions and can be found on our GitHub page. Figure 3 shows that for complex environmental conditions like rain and fog, specialized text prompts guided the refinement process. These prompts enhance the diffusion model's output by providing better control over image refinement, preventing loss of precision and context. We generated text prompts using GPT-based models to extract precise, condition-specific descriptors. Incorporating these keywords enhanced the robustness of the data refinement pipeline, improving the capture of intricate environmental details and variability for better synthetic data quality.

B. Quantitative Results

In our experiments, we used the VKITTI dataset as the source dataset and the KITTI and ACDC datasets as the target datasets. The original KITTI dataset contains images captured under clear weather conditions, as well as some samples of sunset and overcast scenes. To assess the realism of fog and rain refined images, we utilized the ACDC dataset, which contains data from diversified weather conditions including rain and foggy weather. These target domain datasets encompass a variety of scenes, from which we randomly selected a scene to evaluate the similarity between the refined synthetic data and the corresponding real-world samples. The results of quantitative metrics are shown in Table I, which demonstrate that our refined synthetic data exhibits better fidelity and structural similarity to real images in all weather conditions. We also compared the results of our approach against several state-of-the-art techniques for quality refinement cross validation using the FID metric. We used results reported in [9] for comparison. Table II compares our ControlNet + Upscaling method with various state-of-the-art unpaired image-to-image translation techniques, where our method achieves the lowest FID score (44.14). Unlike GAN-based synthetic-to-real adaptation methods such as T2Net and GASDA, our method benefits from control mechanisms and upscaling techniques, leading to high-fidelity image generation.

Authorized licensed use limited to: Maynooth University Library. Downloaded on November 07,2026 at 15:24:40 UTC from IEEE Xplore. Restrictions apply.

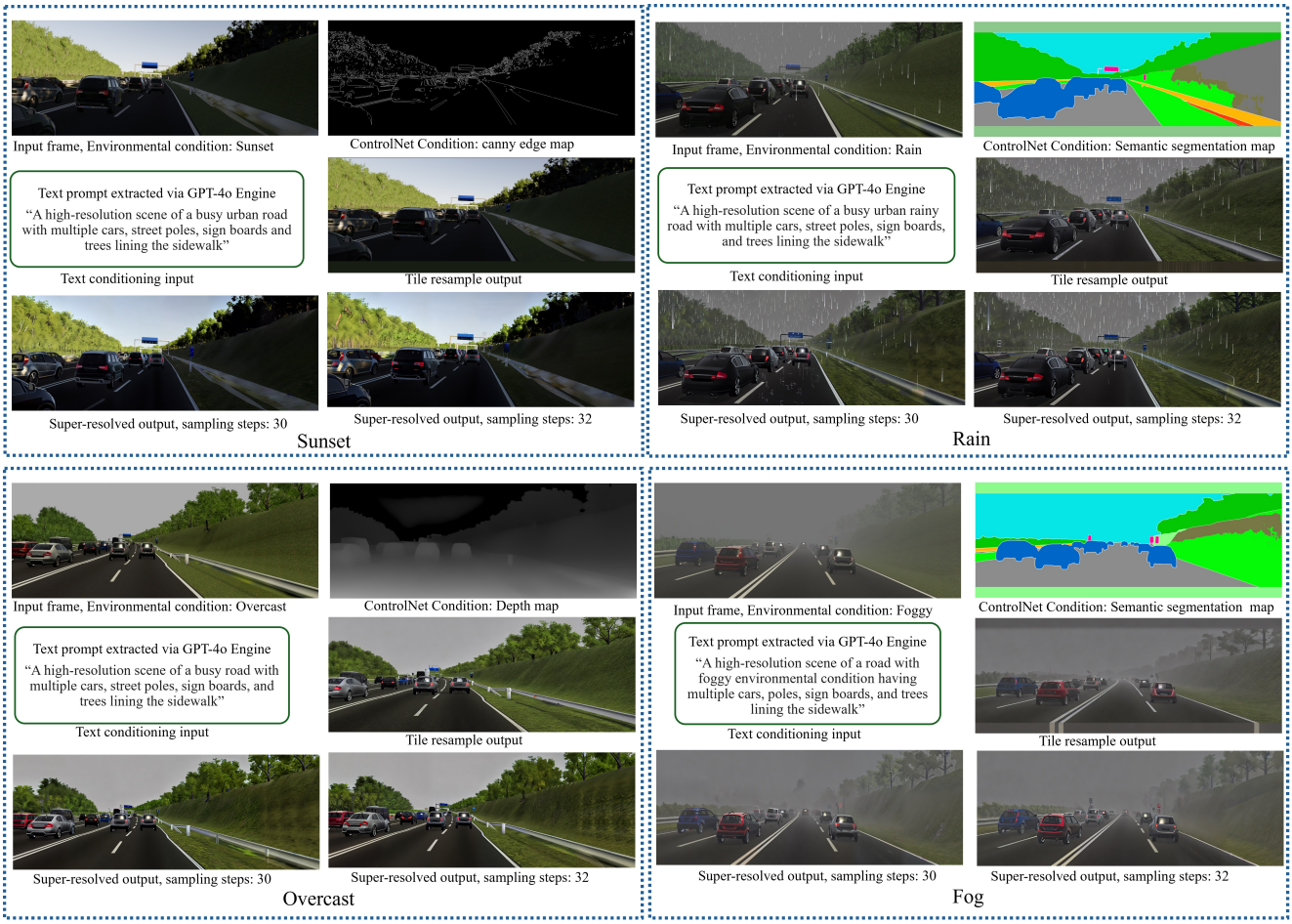


Fig. 3: Synthetic image refinement for four test samples across varied weather conditions. Each block shows: (1) the low-resolution (x_{LR}) input, (2) ControlNet annotator results, (3) the text conditioning prompt, (4) the first-stage refining output (ControlNet tile re-sampler), (5) the super-resolution (SR) result at 30 sampling steps, and (6) the final super-resolved (x_{SR}) image at 32 steps.



Fig. 4: Comparison of object detection results on real-world ACDC dataset. The first row presents the results of a model trained on original VKITTI data and the second row presents results of model trained with refined data.

C. Object Detection and Classification Results

To validate the synthetic data refinement pipeline, the refined dataset that was generated via this pipeline was evaluated for their impact on model accuracy and robustness in downstream object detection and classification tasks. For this purpose we have used YOLOv8 nano model and fine-tuned it on three distinct classes including car, truck, and van. The nano model, characterized by a significantly reduced number of model parameters, is optimized for enhanced speed and computational efficiency. We trained this model using optimal hyper parameter settings with batch size of 8. During this experimental phase the performance of models fine-tuned on the refined VKITTI dataset was compared with models fine-tuned on the original VKITTI dataset. Each weather condition from the VKITTI dataset was split into training, validation, and testing subsets and only training data was refined using upscaled ControlNet to validate its effectiveness over original VKITTI. Table III shows improvements when using refined VKITTI data over the original VKITTI data using different metrics including mean Average Precision (mAP), Precision, and Recall. The refined data achieved better results than the original dataset across all weather conditions. In addition to

Authorized licensed use limited to: Maynooth University Library. Downloaded on November 07,2026 at 15:24:40 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Quantitative comparison of diffusion-based refined VKITTI images vs original VKITTI images

Dataset1 (Synthetic)	Dataset2 (Real)	FID↓	SSIM↑	LPIPS↓
VKITTI Fog	ACDC Fog	55.211	0.476	0.590
Refined VKITTI Fog	ACDC Fog	**48.359**	**0.488**	**0.578**
VKITTI Rain	ACDC Rain	46.698	0.391	0.580
Refined VKITTI Rain	ACDC Rain	**44.202**	**0.432**	**0.574**
VKITTI Sunset	KITTI	54.111	0.266	0.608
VKITTI Sunset Refined	KITTI	**49.627**	**0.339**	**0.616**
VKITTI Overcast	KITTI	69.645	0.310	0.631
VKITTI Overcast Refined	KITTI	**61.918**	**0.316**	**0.627**

TABLE II: FID scores for different methods

Method	FID↓
T2Net [8]	102.16
GASDA [7]	89.70
DIVIDE [9]	76.85
CUT [31]	78.04
ILVR [31]	81.49
SDEdit [32]	80.73
ControlNet + Upscaling (Ours)	**44.140**

testing on the VKITTI data, we also evaluated our approach on real-world data to further validate its effectiveness for real world scenarios. Table III presents the object detection results on real-world KITTI data, demonstrating that models fine-tuned with refined VKITTI data achieve better performance compared to those fine-tuned with the original VKITTI data, as also illustrated in Figure 4. The first row shows results from a model trained on the original VKITTI data, with examples of false detections and missed detections in challenging conditions like rain and snow. In contrast, the second row shows results of a model trained on refined data, with correct detections even in adverse weather conditions.

TABLE III: Comparison of object detection results of models trained on original VKITTI images vs refined images

Dataset (VKITTI)	mAP50↑	mAP50-95↑	Precision↑	Recall↑
Original Fog	0.753	0.564	0.910	0.688
Refined Fog	**0.872**	**0.757**	**0.965**	**0.803**
Original Overcast	0.775	0.577	0.914	0.676
Refined Overcast	**0.899**	**0.788**	**0.961**	**0.834**
Original Rain	0.838	0.617	0.904	0.776
Refined Rain	**0.906**	**0.805**	**0.946**	**0.857**
Original Sunset	0.884	0.643	0.917	0.797
Refined Sunset	**0.956**	**0.862**	**0.958**	**0.906**

D. Computational Cost

In this subsection, we evaluate the computational efficiency of the proposed image refinement pipeline, focusing on inference time and overall feasibility. Table IV presents inference times per image for each refinement stage and total processing time for both single and batched inputs. Stage 1 requires 1.51–1.62 seconds per image, while Stage 2 takes 2.24–2.35 seconds for 30 sampling steps, and 2.31–2.48 seconds for 32 steps. Total batch inference times average between 1.93 and 2.08 minutes. Rainy conditions slightly increase processing time, whereas sunset and overcast conditions are more efficient. FLOPS metric indicates how many floating-point calculations a system can perform in a single second. In our case, the model's FLOPS range from 8.01 to 8.64. Despite minor variations, the method maintains a favorable trade-off between computation and output quality, ensuring its suitability for autonomous driving scenarios.

V. Discussion and Limitations

Our experimental results demonstrate the effectiveness of ControlNet-based diffusion models combined with upscaling techniques to bridge the domain gap between synthetic and real-world data. Despite the promising results, our current approach has some limitations that point to opportunities for future work. One is that generative AI techniques in general do not retain the original semantics of road signs. Also, because each image is processed in isolation rather than as a sequence of video frames, vehicle models and colours are not temporally consistent. Similarly, brake lights may be rendered as on, even if they were off in the original image. For the rainy weather condition, the technique tended to generate noisy and somewhat unrealistic raindrops. All of the above can be seen by inspection of Fig. 1. and Fig. 3. Another limitation is the domain gap between control signals and the actual VKITTI annotations. For example, the depth or segmentation maps extracted from the control signals might be different the VKITTI annotations. Lastly, the use of the super-resolution model improves the quality of images, but requires significant computational resources to generate refined images.

VI. Conclusions and Future Work

In this paper, we proposed an innovative methodology that integrates domain adaptation, text-guided conditioning, and multi-stage super-resolution to enhance the fidelity and detail of synthetic datasets, particularly in the context of roadside object detection. By employing techniques including ControlNet for precise domain adaptation and R-ESRGAN 4x+ for high-quality image refinement, our methodology significantly improves the representation of complex objects such as vehicles and street infrastructure in synthetic images. This framework not only narrows the gap between synthetic and real-world data but also contributes to the advancement of autonomous vehicle applications, where accurate and detailed visual information is critical. In future work, we plan to address the limitations identified earlier, such as improving object consistency through the training of a custom ControlNet model and optimizing the pipeline to reduce processing time. We also aim to incorporate more sophisticated text-guided models and Large Language Model advances for more nuanced control over synthetic data. These improvements should enable more accurate and detailed data generation, and could also be applied to 3D tasks such as object detection and scene reconstruction.

Acknowledgments

This work is carried out with the financial support of the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. Additionally, the research work is jointly funded by Taighde Eireann – Research Ireland under Grant IRCLA/2023/1992 and the ADAPT Centre for Digital Content Technology.

Authorized licensed use limited to: Maynooth University Library. Downloaded on November 07,2026 at 15:24:40 UTC from IEEE Xplore. Restrictions apply.

TABLE IV: Computational cost table showing average inference time per image and total inference time per batch (50 samples) for executing multi-stage refinement pipeline.

VKITTI/Weather Condition	**Average Inference Time Per Image (Seconds)**				**Total Inference Time**	**FLOPS**
	Stage 1 Refinement: ControlNet	**Stage 2 Refinement: Upscaling using SR (ESRGAN4X)**	**Sampling Steps 30**	**Sampling Steps 32**	**(Minutes)/Batch (50 Samples)**	**(x10^9)**
Sunset	1.51	+0.29 = 1.80	+0.44 = 2.24	+0.07 = 2.31	≈1.93	8.64
Rain	1.62	1.86	2.35	2.48	≈2.08	8.01
Overcast	1.54	1.74	2.27	2.36	≈1.98	8.42
Fog	1.57	1.79	2.31	2.38	≈1.99	8.38

REFERENCES

[1] R. He, S. Sun, X. Yu, C. Xue, W. Zhang, P. Torr, S. Bai, and X. Qi, "Is synthetic data from generative models ready for image recognition?" *CoRR*, vol. abs/2210.07574, 2022.

[2] A. Cordier, P. Gutierrez, and V. Plessis, "Improving generalization with synthetic training data for deep learning based quality inspection," *arXiv preprint arXiv:2202.12818*, 2022.

[3] D. Peng, P. Hu, Q. Ke, and J. Liu, "Diffusion-based image translation with label guidance for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2023, pp. 808–820.

[4] F. Shen, L. Zhou, K. Kucukaytekin, Z. Liu, H. Wang, and A. Knoll, "Controluda: Controllable diffusion-assisted unsupervised domain adaptation for cross-weather semantic segmentation," *arXiv preprint arXiv:2402.06446*, 2024.

[5] Y. Jia, L. Hoyer, S. Huang, T. Wang, L. V. Gool, K. Schindler, and A. Obukhov, "Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control," in *Synthetic Data for Computer Vision Workshop at CVPR 2024*, 2023.

[6] A. Lopez-Rodriguez and K. Mikolajczyk, "Desc: Domain adaptation for depth estimation via semantic consistency," *Int. J. Computer Vision*, vol. 131, no. 3, pp. 752–771, 2023.

[7] S. Zhao, H. Fu, M. Gong, and D. Tao, "Geometry-aware symmetric domain adaptation for monocular depth estimation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, pp. 9788–9798.

[8] C. Zheng, T.-J. Cham, and J. Cai, "T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks," in *Proc. European Conf. Computer Vision (ECCV)*, 2018, pp. 767–783.

[9] D. Shim and H. J. Kim, "Divide: Learning a domain-invariant geometric space for depth estimation," *IEEE Robotics and Automation Letters*, 2024.

[10] C.-T. Lin, J.-L. Kew, C. S. Chan, S.-H. Lai, and C. Zach, "Cycle-object consistency for image-to-image domain adaptation," *Pattern Recognition*, vol. 138, p. 109416, 2023.

[11] M. Hnewa and H. Radha, "Integrated generative-model domain-adaptation for object detection under challenging conditions," in *2022 IEEE 95th Vehicular Technology Conf. (VTC2022-Spring)*. IEEE, 2022, pp. 1–5.

[12] T. Rothmeier and W. Huber, "Let it snow: On the synthesis of adverse weather image data," in *2021 IEEE Int. Intelligent Transportation Systems Conf. (ITSC)*. IEEE, 2021, pp. 3300–3306.

[13] W. Zhang, Z. Lv, H. Zhou, J.-W. Liu, J. Li, M. Li, Y. Li, D. Zhang, Y. Zhuang, and S. Tang, "Revisiting the domain shift and sample uncertainty in multi-source active domain transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 751–16 761.

[14] E. Yu, J. Lu, B. Zhang, and G. Zhang, "Online boosting adaptive learning under concept drift for multistream classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 15, 2024, pp. 16 522–16 530.

[15] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are gan generated images easy to detect? a critical analysis of the state-of-the-art," in *2021 IEEE Int. Conf. Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[16] X. Guo, Z. Wang, Q. Yang, W. Lv, X. Liu, Q. Wu, and J. Huang, "Gan-based virtual-to-real image translation for urban scene semantic segmentation," *Neurocomputing*, vol. 394, pp. 127–135, 2020.

[17] H. Azuma, Y. Matsui, and A. Maki, "Zodi: Zero-shot domain adaptation with diffusion-based image transfer," *arXiv preprint arXiv:2403.13652*, 2024.

[18] A. Karn, S. Kumar, S. K. Kushwaha, and R. Katarya, "Image synthesis using gans and diffusion models," in *2023 IEEE Int. Conf. Contemporary Computing and Communications (InC4)*, vol. 1. IEEE, 2023, pp. 1–6.

[19] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conf. on Machine Learning*. PMLR, 2015.

[20] D. Peng, Q. Ke, A. Ambikapathi, Y. Yazici, Y. Lei, and J. Liu, "Unsupervised domain adaptation via domain-adaptive diffusion," *IEEE Transactions on Image Processing*, 2024.

[21] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510.

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

[23] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2023, pp. 3836–3847.

[24] Y. Benigmim, S. Roy, S. Essid, V. Kalogeiton, and S. Lathuilière, "One-shot unsupervised domain adaptation with personalized diffusion models," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2023, pp. 698–708.

[25] Y. Zhang, S. Chen, W. Jiang, Y. Zhang, J. Lu, and J. T. Kwok, "Domain-guided conditional diffusion model for unsupervised domain adaptation," *arXiv preprint arXiv:2309.14360*, 2023.

[26] K. Song, L. Han, B. Liu, D. Metaxas, and A. Elgammal, "Diffusion guided domain adaptation of image generators," *arXiv preprint arXiv:2212.04473*, 2022.

[27] Z. Hartley, R. J. Lind, M. P. Pound, and A. P. French, "Domain targeted synthetic plant style transfer using stable diffusion lora and controlnet," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2024, pp. 5375–5383.

[28] Y. Cabon, N. Murray, and M. Humenberger, "Virtual kitti 2," *arXiv preprint arXiv:2001.10773*, 2020.

[29] C. Sakaridis, D. Dai, and L. V. Gool, "Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2021, pp. 10 765–10 775.

[30] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[31] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 319–345.

[32] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," *arXiv preprint arXiv:2108.01073*, 2021.

Authorized licensed use limited to: Maynooth University Library. Downloaded on November 07,2026 at 15:24:40 UTC from IEEE Xplore. Restrictions apply.