2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

DiffPlace: Street View Generation via Place-Controllable Diffusion Model Enhancing Place Recognition

Ji Li[1], Zhiwei Li[2], Shihao Li[2], Zhenjiang Yu[2], Boyang Wang[2,†], Haiou Liu[2]

***Abstract*— Generative models have advanced significantly in realistic image synthesis, with diffusion models excelling in quality and stability. Recent multi-view diffusion models improve 3D-aware street view generation, but they struggle to produce place-aware and background-consistent urban scenes from text, BEV maps, and object bounding boxes. This limits their effectiveness in generating realistic samples for place recognition tasks. To address these challenges, we propose DiffPlace, a novel framework that introduces a place-ID controller to enable place-controllable multi-view image generation. The place-ID controller employs linear projection, perceiver transformer, and contrastive learning to map place-ID embeddings into a fixed CLIP space, allowing the model to synthesize images with consistent background buildings while flexibly modifying foreground objects and weather conditions. Extensive experiments, including quantitative comparisons and augmented training evaluations, demonstrate that DiffPlace outperforms existing methods in both generation quality and training support for visual place recognition. Our results highlight the potential of generative models in enhancing scene-level and place-aware synthesis, providing a valuable approach for improving place recognition in autonomous driving.**

I. Introduction

In recent years, generative models [1], [2], [3] have seen significant progress, particularly in the generation of high-quality and realistic visual content. Diffusion models [3], [4], one of the major contributors to this progress, are especially known for their stable and high-quality sample generation. Recent breakthroughs in controllable generative technologies have further enabled precise and flexible content customization. In particular, the development of multi-view diffusion models have greatly improved the synthesis of street images with 3D geometry control. Initial studies focused on generating street view images to enhance image-based bird's-eye view (BEV) perception methods, with models like BEVGen [5], BEVControl [6], and MagicDrive [7]. More recent research has expanded to generating driving scene videos [8], [9], [10], [11], aiming to enhance various aspects of the scene generation. However, generating background-consistent (or place-aware) street view images remains a challenge, as current methods rely solely on text, BEV maps, and object bounding boxes.

While these components are useful, they do not provide sufficient information for accurately modeling the background of a scene, limiting the realism of the generated content. As a result, these methods struggle to create synthetic samples that are closely resemble from real-world locations. This limitation restricts the use of generative models in place recognition tasks [12], [13], which remains one of the most important yet challenging tasks in autonomous driving and robotics. A straightforward approach to address this might involve fine-tuning text-conditioned diffusion models directly on place-ID embeddings to enhance their place-aware prompting capabilities. However, this strategy has several drawbacks. First, it compromises the ability of street view generative models to generate images from text, BEV maps, and 3D bounding boxes. Additionally, fine-tuning such models demands substantial computational resources. To overcome these challenges, we propose a novel framework for street view generation to generate background-controllable multi-view images with additional place-ID

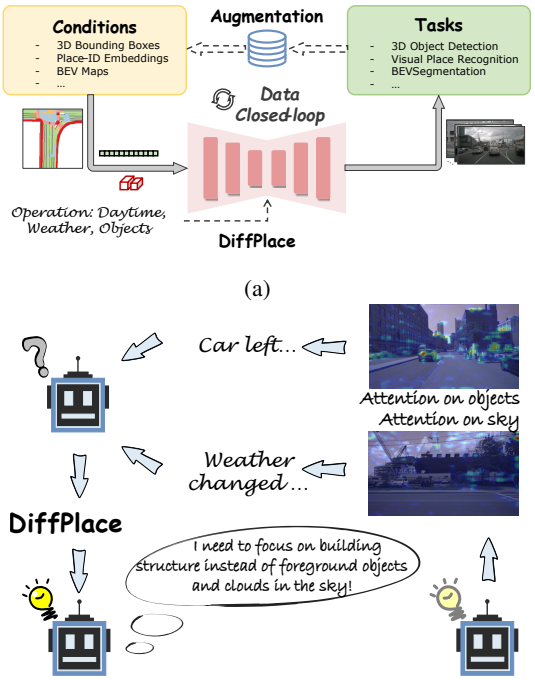

Fig. 1: **Systematic depiction of proposed DiffPlace.** (a) Generation is seen as the reverse process of perception, which generates images with the input of place-ID embeddings, bounding boxes, etc. We are the first to achieve data closed-loop for visual place recognition augmentation training. (b) The original place recognition model mistakenly focused on foreground objects and clouds in the sky. We corrected the place recognition model by augmented training on "car left" and "weather changed" situations through DiffPlace.

[1]The University of Hong Kong. [2]Beijing Institute of Technology. The research is funded by the National Natural Science Foundation of China under Grants No. 52302489 and 52172378. † is corresponding author.

Project page: https://jerichoji.github.io/DiffPlace/

embeddings. These embeddings integrate seamlessly with existing controllability.

Our work introduces **DiffPlace**, a method that leverages visual place recognition networks and a place-ID controller to enable place-aware scene synthesis. The core innovation of the place-ID controller involves linear projection, perceiver transformer, and contrastive learning to map place-ID embeddings into the fixed CLIP space. Through place-aware control, the synthesized images maintain consistent background information but allow modifications to foreground objects and weather conditions. These generated images ensure that the generated scenes retain key background elements while allowing for variability in foreground details, thus they can be used to augment place recognition training. We summarize our contributions as follows:

- To the best of our knowledge, we are the first to propose a place-controllable diffusion model for generating augmented data to improve the performance of existing place recognition methods. Our approach enables fine-grained control over both objects-level and scene-level synthesis as verified by extensive quantitative experiments.
- Our method generates high-fidelity street view images that preserve background consistency across varying weather conditions and different foreground objects by simply adjusting the prompts. Using our synthetic images as augmented data significantly enhances the training of place recognition models, demonstrating superior performance.

II. Related Work

A. Latent Diffusion Models

The task of text-to-image (T2I) generation focuses on creating realistic images from textual descriptions. Diffusion models, a class of probabilistic generative models, introduce noise to data in a gradual manner and then learn to reverse this process in order to generate samples [3]. Early approaches framed this challenge as a sequence-to-sequence problem. In recent advancements, Denoising Diffusion Probabilistic Models (DDPM) [4] have demonstrated substantial success in addressing the T2I task. The image quality has been further enhanced by leveraging the strong image generation capabilities of diffusion models, or by improving the alignment between text and images through powerful text encoders [14]. For instance, DALL-E [15] employs text tokens to generate discrete image embeddings via VQ-VAE. Further improvements have been made in subsequent works, utilizing more sophisticated architectures like encoder-decoder frameworks [16], and hierarchical transformers [17].

These models have gained considerable attention in recent years due to their impressive performance across a variety of applications, setting new benchmarks in areas such as video generation [18], and 3D content creation [19]. In order to further improve the controllability of image generation, techniques like ControlNet [20] and GLIGEN [21] have been introduced, enabling the use of various control signals such as depth maps, segmentation maps, canny edges, and sketches.

B. Visual Place Recognition

Visual Place Recognition (VPR) has progressed from traditional feature-based approaches, such as RootSIFT [22] to more advanced deep learning methods. Early CNN-based models [12] achieved notable results, but recently, Vision Transformers (ViTs) [23], [24] have emerged, providing enhanced performance owing to their capacity to grasp long-distance relationships. Visual Foundation Models (VFMs) like CLIP [14] and DINOv2 [25] have gained popularity, with solutions such as AnyLoc [26] using dense local features for zero-shot VPR. Although these models perform well in certain contexts, they struggle when faced with significant time gaps or environmental changes. Fine-tuned models like SALAD [23] enhance accuracy but require increased feature dimensionality and higher memory consumption. Other approaches, such as CricaVPR [27], integrate trainable adapters within ViT architectures to prevent catastrophic forgetting, though they still confront computational difficulties. Through early efforts [28], [29] have explored GAN models for visual place recognition under changing environments, existing VPR methods remain challenged by changing foregrounds, lighting, and weather conditions. In this paper, we focus on using diffusion models to generate richer training data, incorporating diverse foreground objects and varying weather conditions. This approach aims to enhance the training of any existing VPR algorithms, and improve their robustness in real-world applications.

C. Street View Image generation

This task can be seen as the reverse of perception tasks, aiming to produce images from inputs like bounding boxes. Previous techniques employed GANs [30] or diffusion models [31], [32] to synthesize images based on 2D layouts. These approaches commonly encoded the layout into a conditional image, which was subsequently processed through downsampling and upsampling alongside the data. More contemporary work, such as [5] has leveraged VQ-VAE to generate multi-view urban street view images from BEV layouts.

At the same time, approaches such as BEVControl [6], MagicDrive [7], and DrivingDiffusion [8] have incorporated layout specifications to further refine the image generation process. The essence of diffusion-based generative models resides in their capacity to grasp the complexities of the environment. By harnessing the potential of these models, [33], [34], [35] then introduce 3D Gaussian Splatting [36] representations to render novel-view images in generated scenes. Unlike prior approaches [35] that scale up to video generation models with substantial computational cost, our method instead focuses on advancing image generation on controllability to generate auxiliary datasets for visual place recognition.

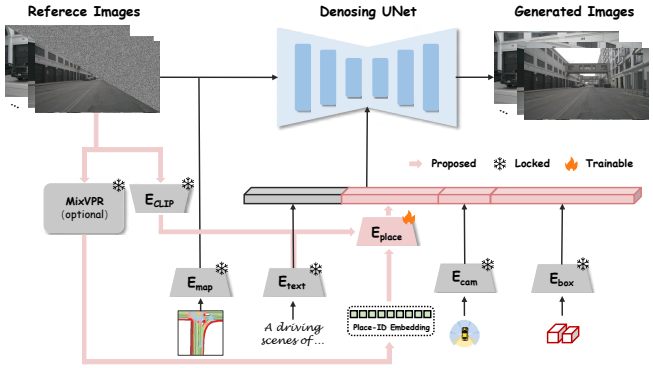

Fig. 2: **Overview of the DiffPlace pipeline.** The input scene representation $S = \{MAP,\ BOX,\ TEXT,\ PLACE\_ID\}$ is processed by dedicated encoders: E_{map}, E_{text}, E_{place}, E_{cam}, and E_{box}. The resulting encoded features are concatenated and fed into the U-Net via cross-attention mechanisms to generate multi-view consistent images with controllable background and foreground elements. An optional visual place recognition network (MixVPR [37]) is utilized to extract place-ID embeddings, enabling enhanced place-aware synthesis.

III. Methods

While advanced generative methods can create high-quality images of driving scenes, their impact on downstream perception tasks such as place recognition (shown in Fig. 1a), remains limited. We believe this is mainly due to the inadequate control over the generated background information, which is vital for effective scene understanding and place recognition. As depicted in Fig. 2, various strategies are implemented to inject information into multi-view diffusion models. We propose a place-ID controller that maps place-ID embedding from the place recognition network to align with the CLIP image space. Key components of this approach include place-ID encoding, attribute perceiver transformer, and contrastive learning strategy.

In Section 3-A, we begin by presenting the basic notions of multi-view diffusion. Subsequently, in Section 3-B, we unveil our comprehensive diffusion architecture, which incorporates bespoke designs aimed at bolstering the place-controllability of the diffusion models. Lastly, we provide a detailed implementation for adding control of place features to the diffusion model.

A. Preliminary

Latent Diffusion models are intended to capture a probability distribution, denoted as $p_\theta(x_0) = p_\theta(x_{0:T})\, dx_{1:T}$, where x_0 signifies the data and $x_{1:T} := x_1, \ldots, x_T$ represent latent variables. This joint distribution is defined by a Markov chain [4], specifically referred to as the reverse process:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{1}$$

with $p(x_T) = \mathcal{N}(x_T; 0, I)$ and $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I)$. Here, $\mu_\theta(x_t, t)$ is a trainable component, while the variance σ_t^2 consists of untrained time-dependent constants. The aim is to learn μ_θ for generation purposes.

To achieve this, the forward process is constructed:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{2}$$

where

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{3}$$

and β_t are constants. The DDPM approach demonstrates that by defining:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} x_t - \beta_t \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t), \tag{4}$$

with α_t and $\bar{\alpha}_t$ being constants derived from β_t and ϵ_θ functioning as a noise predictor, we can learn ϵ_θ by minimizing the following loss function:

$$\mathcal{L}_{\text{base}} = \mathbb{E}_{t, x_0, \epsilon} \left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t \right) \right\|^2, \tag{5}$$

where ϵ is a random variable sampled from $\mathcal{N}(0, I)$.

To maintain viewpoint consistency, multi-view diffusion models, such as those proposed in [38], typically utilize cross-view attention modules. As expressed in Eq. 6, given the sparse camera layout in driving environments, each cross-view attention mechanism enables the target view in accessing information from its neighboring left and right views, The target view states then consolidates this information through a skip connection. The cross-view attention Attention_{cv} computation can be described as follows:

$$\text{Attention}_{cv}(Q_t, K_i, V_i) = \text{softmax}\left(\frac{Q_t K_i^T}{\sqrt{d}}\right) \cdot V_i, \quad i \in \{l, r\} \tag{6}$$

In this context, t, l, and r refer to the target view, left view, and right view, respectively.

B. Overall Architecture

Our latent multi-view diffusion model consists of a VAE encoder E, a denoising U-Net, and a VAE decoder D. While description text, BEV maps, and 3D geometric information similar to MagicDrive [7] provide useful context, they do not offer precise guidance for generating the background. To address this, we introduce an additional controller E_{place}, for place-ID, which expands the original input-output pair. Let $S = \{MAP,\ BOX,\ TEXT,\ PLACE\_ID\}$ represent the components of a driving scene encompassing the ego vehicle, where *MAP* is a binary map depicting a $w \times h$ meter area of the road in Bird's Eye View (BEV), with c semantic classes. *BOX* $= \{(c_i, b_i)\}_{i=1}^{N}$ indicates the locations of 3D bounding boxes for each object within the scene. At the scene level, *TEXT* includes textual descriptions offering fundamental context about the scene (such as weather and time of day). Additionally, *PLACE_ID* offers more detailed background information of the scene. Given the camera pose $P = [K, R, T]$ (which includes intrinsic parameters, rotation,

and translation), the generator $G(\cdot)$ aims to synthesize multi-view consistent images that incorporate foreground (object-level), midground (map-level), and background (scene-level) information.

C. Place-ID Encoding

Place recognition networks are specifically designed to extract distinctive features of a given place. In our framework, we can leverage any existing visual place recognition method. We optionally take MixVPR [37] as example, and adopt an implementation that utilizes a ResNet-50 backbone in combination with all-MLPs aggregation, producing a discriminative place-ID embedding with a dimension of 4096.

To ensure the dimension of the place-ID embedding aligns with other conditions, we implement two trainable linear projection layers. These layers operate on the place-ID embedding and output a sequence of features Z with length N_S (set to 4 in our implementation), matching the dimensionality of other prompt embeddings. It has been observed that using two linear layers results in better performance compared to employing an MLP with successive blocks [39].

We do not apply a masking mechanism to the place-ID embedding, as the aggregation of various scenes consistently yields a fixed-length place-ID embedding, thereby eliminating the need for random masking operations. Moreover, we refrain from using a multi-view approach in place recognition, as we believe it could disrupt the integrity of place information. Instead, we rely on camera parameter encoding and cross-view attention in the U-Net to maintain consistency across multiple views during the generating phase.

D. Attribute Perceiver Transformer

We employ a perceiver-based transformer to map the embedding with the aid of attributes in reference images. Prior to attribute extraction using the CLIP Image Encoder, we mask out the foreground objects using bounding boxes and the sky regions following the method in [40], thereby directing the network's focus to the street scene and minimizing interference from the other conditions. This approach ensures that the attention mechanism is concentrated solely on the relevant portions of the scene.

After being processed through a linear projector, the place-ID embeddings are enhanced by the CLIP image features from reference images through several cross-attention layers (we implement 3 layers with a dimension of 1024 in this study). Given the query place-ID embeddings Z and the CLIP image features c_I, the output of the cross-attention mechanism is given by the following expression:

$$Z = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{7}$$

where $Q = ZW_q$ represents the query matrix for the place-ID embeddings, and $K = c_I W_k$, $V = c_I W_v$ denote the key and value matrices derived from the CLIP hidden states.

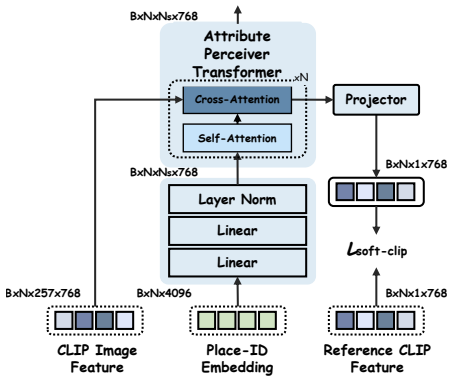

Fig. 3: **Details of the proposed place-ID controller.** (a) Place-ID embeddings are projected via trainable linear layers to align with other conditions; (b) Attribute perceiver transformer interacts place-ID embeddings Z with CLIP image features c_I; (c) A contrastive loss $\mathcal{L}_{\text{SoftCLIP}}$ is applied to align place-ID embeddings with the CLIP latent space.

E. Contrastive Learning

We also integrate contrastive learning into the ControlNet training process, which aids in mapping place-ID embeddings into the CLIP space, thereby facilitating the alignment of place-ID conditions with the original text descriptions.

Contrastive learning is a potent technique for learning representations across different modalities by maximizing the cosine similarity of positive pairs while minimizing it for negative pairs. Previous studies have shown the efficacy of contrastive learning, even in the context of neural data [41]. In our method, we apply contrastive learning to align additional place-ID embedding conditions with the static CLIP image space. Specifically, we use a projector to transform features with a shape of $N \times C$ to match the length of CLIP image tokens (after pooling in our implementation). These tokens are then normalized, and the SoftCLIP loss is computed between them. It is important to note that this strategy is only used during the training phase.

Our SoftCLIP loss is inspired by knowledge distillation [42], which suggests that the softmax probability distribution generated by a strong teacher model provides a more effective teaching signal for the student model than hard labels. To create soft labels, we first calculate the dot product of the CLIP embeddings within a batch. The SoftCLIP loss is then computed as our contrastive objective between the CLIP-to-CLIP and place-ID-to-CLIP matrices, as follows:

$$\mathcal{L}_{\text{SoftCLIP}} = \sum_{i=1}^{N}\sum_{j=1}^{N} \left[\frac{\exp\left(\frac{t_i \cdot t_j}{\tau}\right)}{\sum_{m=1}^{N}\exp\left(\frac{t_i \cdot t_m}{\tau}\right)} \cdot \log\left(\frac{\exp\left(\frac{p_i \cdot t_j}{\tau}\right)}{\sum_{m=1}^{N}\exp\left(\frac{p_i \cdot t_m}{\tau}\right)}\right)\right] \tag{8}$$

Therefore, the total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda \cdot \mathcal{L}_{\text{SoftCLIP}} \tag{9}$$

where $\lambda = 0.1$.

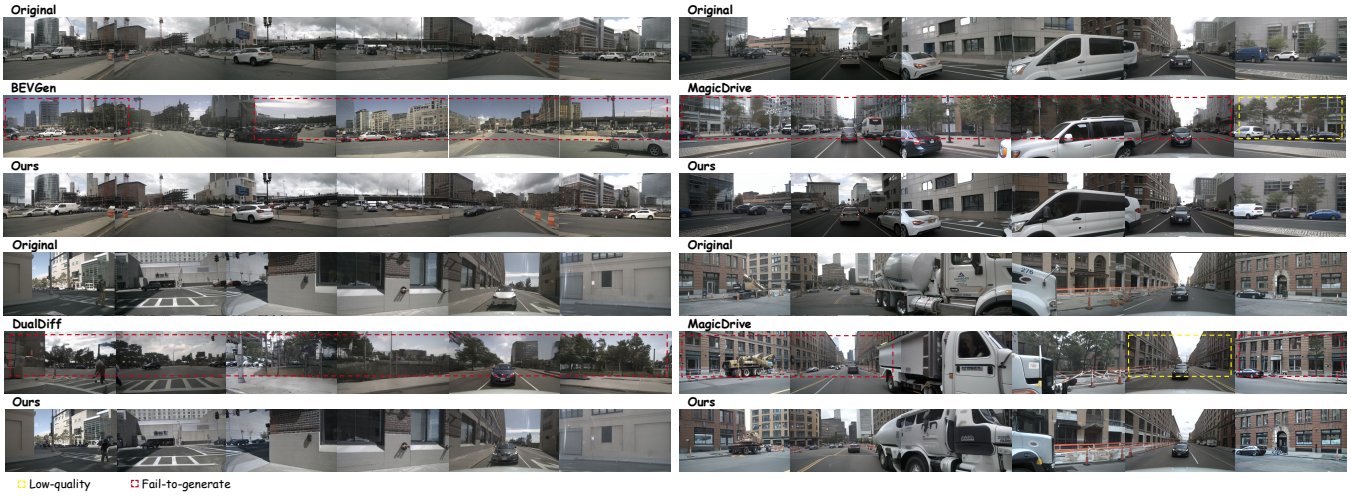

Fig. 4: **Realism and controllability validation.** Our method demonstrates significantly better control over place features, particularly in the background, compared to BEVGen, MagicDrive and DualDiff. We highlight some background areas in low-quality (yellow) and fail-to-generate (red) for comparison. All scenes are from the nuScenes validation set.

IV. Experiments

We experimentally showcase the effectiveness of DiffPlace in enhancing place-aware controllability while preserving objects and weather controllability. Our approach achieves state-of-the-art performance in augmented training support for 3D object detection and place recognition benchmarks. We conduct extensive experiments and analyses to validate our design.

A. Experiments Setups

1) Dataset and Baselines: We conduct experiments on the nuScenes dataset to assess the effectiveness of DiffPlace. This dataset is widely used in 3D object detection and place recognition for autonomous driving. We follow standard methods for generating street view images, utilizing 700 street view scenes for training and 150 for validation. Moreover, our baselines are BEVGen [5], BEVControl [6], MagicDrive [7] and [43], the state-of-the-art approaches for controllable street view generation.

Additionally, we present training support for place recognition experiments on the Pitts30k-test dataset [44],which comprises images sourced from Google Street View and encompasses 8,000 database and 8,000 queries. The Pittsburgh dataset poses notable challenges owing to variations in viewpoint and lighting conditions. To examine the training support, synthesis images are generated from the training set of nuScenes dataset as augmentation data for training 3D object detection and place recognition models.

2) Evaluation Metrics: We assess the realism of the generated images using the Frèchet Inception Distance (FID). The images are generated in alignment with the validation set annotations, and we utilize pre-trained detection and recognition models on generated data to evaluate both image quality and control accuracy. To evaluate object generation accuracy controlled by the input 3D bounding boxes, we use the metrics such as mean Average Precision (mAP) and nuScenes Detection Score (NDS) in [45]. For place controllability, we adopt the same evaluation metrics as [37], where the Average Recall @1 and @5 (AR@1 and AR@5) are calculated. The criteria for considering a query image as successfully retrieved are met when at least one of the top-ranked reference images, either the top-1 or within the top-5, is situated within a 25-meter proximity to the query image.

TABLE I: **Comparison of generation fidelity with street view generation methods.** On the generated images, We test the performance of MixVPR trained on the nuScenes training set. Conditions for data synthesis are from nuScenes validation set, which are not seen.

Method	FID↓	Place Recognition		Reference
		AR@1↑	AR@5↑	
BEVGen [5]	25.6	31.2	60.8	RAL2024
BEVControl [6]	24.8	-	-	ICLR2024
MagicDrive [7]	16.2	35.9	64.1	ICLR2024
DualDiff [43]	**11.0**	48.7	68.9	Arxiv2025
DiffPlace (Ours)	13.4	**57.6**	**75.4**	-

3) Implementation Details: We use the CLIP ViT-L/14 [14] as a frozen image encoder to extract attribute features in the place-ID controller. Experiments are time-consuming as the driving scenes involve 6 different views. Therefore, our DiffPlace leverages pre-trained weights from MagicDrive, which is implemented using ControlNet and Stable Diffusion v1.5, to reduce training costs. For the place recognition task, we trained a MixVPR model on the nuScenes training set to generate original place-ID embeddings, following the hyperparameter settings from [37]. In training support experiment, we train this model on the same training set, augmented with generated data.

During training, we only optimize the place-ID controller while keeping the original fixed. The model is trained with a constant learning rate of 1×10^{-4} and a linear warm-up for the first 3000 iterations. The training process is conducted on 4 NVIDIA 3090 GPUs with a gradient descent batch size of 24. We use AdamW as the optimizer, with a weight decay of 0.01. To reconcile discrepancies in the object detection task,

we use a resolution of 224×400. For place recognition, we resize the images to 320×320 to meet the required input size. For image reconstruction, we use 20 denoising timesteps with the UniPC [46] multi-step noise scheduler.

B. Quantitative Results

1) Realism and Controllability Validation: To validate the effectiveness of DiffPlace in producing realistic imagery, we utilize the nuScenes validation dataset to synthesize street view images and present the relevant metrics in Table I. DiffPlace outperforms the baseline approach in terms of image fidelity, achieving notably lower FID scores. Although there exists an inherent discrepancy between place-ID embeddings and pre-trained models, our refinement technique successfully bridges this gap, resulting in highly realistic and convincing outputs. Regarding controllability, while maintaining the same effect in the 3D object detection task, DiffPlace dramatically exceeds baseline results in place recognition task. This is attributed to the distinct place-ID controller which significantly boosts generation precision on background buildings.

As shown in Figure 4, our method is capable of generating background buildings that closely resemble those in the ground truth. Furthermore, Table I demonstrates that our generated images achieve significantly higher average recall on the validation set—outperforming MagicDrive by 21.7% and 11.3%, and exceeding DualDiff by 8.9% and 7.6%, respectively. Compared with DualDiff, although it produces images with higher photorealism, the recognition results from MixVPR validate the effectiveness of our background (scene-level) controllability, as our generated images are more easily recognized by the pre-trained model. Notably, the validation set was not seen during training.

2) Visualization Experiments: In addition, we utilize the t-SNE and Euclidean Distance of the place-ID embedding (in the shape of 4096) generated by MixVPR on the original nuScenes validation set and synthetic images to conduct visualization experiments. As shown in Fig. 5a and Fig. 5b, compared with MagicDrive, the t-SNE results of ours show a smaller gap, which indicates the images generated by our method can generate images with more-aligned place feature with original ones. Then, the Euclidean Distance between original and synthetic ones in Fig. 5c also revealed that our method has better controllability on place features. Additionally, we extract the attention scores within the cross-attention mechanism influenced by the place-ID embedding to visualize its role during the generation process.

Fig. 6 vividly demonstrates the contribution of place-ID embedding enables dynamic adjustment of attention toward background regions, thereby facilitating the integration of place-aware features.

3) Training Support: DiffPlace is capable of generating augmented data with precise annotations and consistent place-IDs, thereby bolstering the training process for autonomous vehicles' perception and place recognition tasks. For 3D object detection and place recognition, we augment the dataset with an equivalent number of images as in the original, maintaining consistent training iterations and batch sizes to ensure fair comparisons with the baseline. To optimize data augmentation, we randomly modify the weather to "heavily rainy with wet road" and alter half of the bounding boxes in each synthetic scene.

(a) (b) (c)

Fig. 5: **Descriptors visualization results.** (a) and (b) are t-SNE Visualization of place-ID Embedding between nuScenes validation set and synthetic images of ours and MagicDrive's; (c) Euclidean Distance of place-ID Embedding between nuScenes validation set and synthetic images.

Fig. 6: **Place-ID embedding contributed cross-attention visualization in generation process.**

TABLE II: **Training Support for place recognition.** The results are reported on Pitts30k-test set.

Data	MixVPR		CricaVPR	
	AR@1↑	AR@5↑	AR@1↑	AR@5↑
w/o synthetic data	83.5	90.3	90.9	96.0
MagicDrive	84.2	91.1	90.3	95.7
Ours	**89.7**	**95.2**	**92.9**	**96.8**

As shown in Fig. 7, our method preserves the consistency of background buildings across variations in foreground objects and weather conditions. Table II highlights the beneficial impact of DiffPlace-generated data in training place recognition models including MixVPR [37] and CricaVPR [27], with AR@1 and AR@5 surpassing the baseline by 5.5%, 4.1% for MixVPR and 2.6%, 1.1% for CricaVPR, respectively. The inherent consistency in place-ID representation between the generated and original images, combined with increased data diversity, provides strong augmentation support for training place recognition networks.

Moreover, we visualize the feature maps extracted by the state-of-the-art place recognition method CricaVPR on the images from nuScenes test set. As illustrated in Fig. 8, after augmented training , the network shifts its attention towards background buildings, which are more reliable cues for place recognition. This demonstrates that the enhanced attention aligns better with semantic landmarks, indicating the effectiveness of our DiffPlace framework in improving place recognition models.

Furthermore, as presented in Table III, DiffPlace achieves a marginal improvement over BEVFusion in the CAM+LiDAR setting [45], which is competitive with MagicDrive. This demonstrates that our method enhances place

Fig. 7: **Place controllablility under weather and objects edit.**

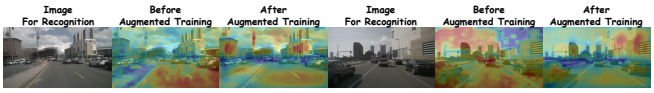

Fig. 8: **Attention map extracted by CricaVPR on nuScenes test images.**

recognition task without compromising its effectiveness in the original task.

C. Ablation Study

As found in [39], [47], we also find that using two linear projector can perform as well as cascaded MLPs, while reducing computational complexity. DiffPlace utilizes perceiver transformer to encode place-ID embedding. To demonstrate the efficacy, we train a model that directly takes the dimensional projection of place-ID embedding, denoted as "w/o perceiver transformer" in Table IV. Due to the limited information in place-ID embedding, supplementary cross-attention mechanisms are crucial for enhancing the model's ability to precisely capture background information, as demonstrated by the disparity in recall rate performance. Additionally, we evaluate the generation results without contrastive loss. In this case, without external force to make place conditions close to the CLIP space, the stability of training phase is reduced and it easily leads to bigger gap that reduces controllability.

D. Limitations

We show representative failure cases in Figure 9. While DiffPlace generally performs well in generating driving scenes with consistent architectural backgrounds, it struggles in environments dominated by dense vegetation. This may be because the pre-trained place recognition models used

TABLE III: **Training Support for 3D Object Detection.** The results are reported on nuScenes validation set.

Data	mAP↑	NDS↑
w/o synthetic data	64.92	69.42
w/ MagicDrive	67.28	70.14
Ours	67.71	70.58

TABLE IV: **Ablation study on perceiver transformer in controller and contrastive learning strategy.** The results are recognition performance of MixVPR on synthesis images.

Data	AR@1↑	AR@5↑
w/o perceiver transformer	46.3	66.5
w/o contrastive loss	51.1	70.2
cascaded MLPs	57.2	75.1
Ours	**57.6**	**75.4**

to extract place-ID embeddings lack the fundamental ability to capture unstructured features such as foliage, making it more difficult for the generator to synthesize plant-heavy scenes. Additionally, when the reference images are captured under extremely low-light conditions, the generated results often fail to match the ground truth. This can be attributed to two factors: (1) the pre-trained Diffusion model is not optimized for generating scenes under very dark lighting, and (2) the control signals, including the place-ID and reference image, do not provide sufficient information to guide faithful synthesis.

Fig. 9: **Failure cases of DiffPlace.**

V. Conclusion

In this work, we introduced DiffPlace, the first place-controllable diffusion model for generating augmented data to enhance place recognition performance. Extensive experiments indicate DiffPlace enables precious control over scene synthesis by place-ID controller, preserving background consistency across varying weather and foreground objects in generation. The augmented training experiments demonstrate that our method significantly enhances place recognition performance, yielding a 5.5% improvement in AR@1 and a 4.1% improvement in AR@5 for MixVPR, as well as gains of 2.6% and 1.1% for CircaVPR, respectively. These results highlight the effectiveness of DiffPlace in advancing place-aware street view generation and supporting robust place recognition. Future work may include training a street view video generative model with controllable background, midground and foreground on large-scale datasets.

References

[1] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[5] A. Swerdlow, R. Xu, and B. Zhou, "Street-view image generation from a bird's-eye view layout," *IEEE Robotics and Automation Letters*, 2024.

[6] K. Yang, E. Ma, J. Peng, Q. Guo, D. Lin, and K. Yu, "Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout," *arXiv preprint arXiv:2308.01661*, 2023.

[7] R. Gao, K. Chen, E. Xie, L. Hong, Z. Li, D.-Y. Yeung, and Q. Xu, "Magicdrive: Street view generation with diverse 3d geometry control," *arXiv preprint arXiv:2310.02601*, 2023.

[8] X. Li, Y. Zhang, and X. Ye, "Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model," in *European Conference on Computer Vision*. Springer, 2025, pp. 469–485.

[9] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-driven world models for autonomous driving," *arXiv preprint arXiv:2309.09777*, 2023.

[10] B. Deng, R. Tucker, Z. Li, L. Guibas, N. Snavely, and G. Wetzstein, "Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[11] Z. Wu, J. Ni, X. Wang, Y. Guo, R. Chen, L. Lu, J. Dai, and Y. Xiong, "Holodrive: Holistic 2d-3d multi-modal street scene generation for autonomous driving," *arXiv preprint arXiv:2412.01407*, 2024.

[12] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[13] J. Li, Q. Liu, B. Wang, H. Liu, and Y. Han, "Rangeplace: A hierarchical range image transformer for lidar-based place recognition," *IEEE Transactions on Intelligent Vehicles*, 2024.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[15] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[17] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[18] S. Bahmani, I. Skorokhodov, A. Siarohin, W. Menapace, G. Qian, M. Vasilkovsky, H.-Y. Lee, C. Wang, J. Zou, A. Tagliasacchi, *et al.*, "Vd3d: Taming large video diffusion transformers for 3d camera control," *arXiv preprint arXiv:2407.12781*, 2024.

[19] P. Wang and Y. Shi, "Imagedream: Image-prompt multi-view diffusion for 3d generation," *arXiv preprint arXiv:2312.02201*, 2023.

[20] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[21] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 511–22 521.

[22] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2911–2918.

[23] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 658–17 668.

[24] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, and H. Wang, "R2former: Unified retrieval and reranking transformer for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 370–19 380.

[25] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[26] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," *IEEE Robotics and Automation Letters*, 2023.

[27] F. Lu, X. Lan, L. Zhang, D. Jiang, Y. Wang, and C. Yuan, "Cricavpr: Cross-image correlation-aware representation learning for visual place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 772–16 782.

[28] H. Hu, H. Wang, Z. Liu, C. Yang, W. Chen, and L. Xie, "Retrieval-based localization based on domain-invariant feature learning under changing environments," in *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2019, pp. 3684–3689.

[29] H. Porav, W. Maddern, and P. Newman, "Adversarial training for adverse conditions: Robust metric localisation using appearance transfer," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1011–1018.

[30] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2337–2346.

[31] F. Shen, L. Zhou, K. Kuecuekaytekin, G. B. F. Eskandar, Z. Liu, H. Wang, and A. Knoll, "W-controluda: Weather-controllable diffusion-assisted unsupervised domain adaptation for semantic segmentation," *IEEE Robotics and Automation Letters*, pp. 1–8, 2025.

[32] S. Sun, Z. Gu, T. Sun, J. Sun, C. Yuan, Y. Han, D. Li, and M. H. Ang, "Drivescenegen: Generating diverse and realistic driving scenarios from scratch," *IEEE Robotics and Automation Letters*, vol. 9, no. 8, pp. 7007–7014, 2024.

[33] Z. Yu, H. Wang, J. Yang, H. Wang, Z. Xie, Y. Cai, J. Cao, Z. Ji, and M. Sun, "Sgd: Street view synthesis with gaussian splatting and diffusion prior," *arXiv preprint arXiv:2403.20079*, 2024.

[34] R. Gao, K. Chen, Z. Li, L. Hong, Z. Li, and Q. Xu, "Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes," *arXiv preprint arXiv:2405.14475*, 2024.

[35] Y. Yan, Z. Xu, H. Lin, H. Jin, H. Guo, Y. Wang, K. Zhan, X. Lang, H. Bao, X. Zhou, *et al.*, "Streetcrafter: Street view synthesis with controllable video diffusion models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 822–832.

[36] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, July 2023.

[37] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "Mixvpr: Feature mixing for visual place recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 2998–3007.

[38] Y. Shi, P. Wang, J. Ye, M. Long, K. Li, and X. Yang, "Mvdream: Multi-view diffusion for 3d generation," *arXiv preprint arXiv:2308.16512*, 2023.

[39] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[40] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[41] P. Scotti, A. Banerjee, J. Goode, S. Shabalin, A. Nguyen, A. Dempster, N. Verlinde, E. Yundler, D. Weisberg, K. Norman, *et al.*, "Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[42] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[43] H. Li, Z. Yang, Z. Qian, G. Zhao, Y. Huang, J. Yu, H. Zhou, and L. Liu, "Dualdiff: Dual-branch diffusion model for autonomous driving with semantic fusion," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025.

[44] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *The International journal of robotics research*, vol. 27, no. 6, pp. 647–665, 2008.

[45] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[46] W. Zhao, L. Bai, Y. Rao, J. Zhou, and J. Lu, "Unipc: A unified predictor-corrector framework for fast sampling of diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[47] C. Feng, Z. Chen, A. Holynski, A. A. Efros, and A. Owens, "Gps as a control signal for image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025.