BevGAN: Generative Fisheye Cross-View Transformers

Rania Benaissa, Antonyo Musabini, Rachid Benmokhtar, Manikandan Bakthavatchalam
and Xavier Perrotton
Valeo, Brain Division, anSWer AI R&D Center, Créteil 94000, France

Keywords: Generative Models, Image Generation, Image-to-Image Translation, Driving Assistance Systems, Autonomous Vehicles.

Abstract: Current parking assistance and monitoring systems synthesize Bird Eye View (BEV) images to improve drivers visibility. These BEV images are created using a popular perspective transform called Inverse Perspective Mapping (IPM), which projects pixels of surround-view images captured by onboard fisheye cameras onto a flat plane. However, IPM faces challenges in accurately representing objects with varying heights and seamlessly stitching together the projected surround-views due to its reliance on rigid geometric transformations. To address these limitations, we present BevGAN, a novel geometry-guided Conditional Generative Adversarial Networks (cGANs) model that combines multi-scale discriminators along with a transformers-based generator that leverages fisheye cameras calibration and attention-mechanisms to implicitly model geometrical transformations between the views. Experimental results demonstrate that BevGAN outperforms IPM and state-of-the-art cross-view image generation methods in terms of image fidelity and quality. Compared to IPM, we report an improvement of $+6.2db$ on PSNR and $+170\%$ on MS-SSIM when evaluated on a synthetic dataset depicting both parking and driving scenarios. Furthermore, the generalization ability of BevGAN on real-world fisheye images is also demonstrated through zero-shot inference.

1 INTRODUCTION

Automotive Surround-view Systems (SVS) assist vehicles in navigating through unpredictable real-world scenarios, reducing risks of accidents by making real-time, high-confidence decisions.

SVS involves four wide-angle fisheye lens cameras known as the surround-view cameras (SVCs), which are already mounted in vehicles and offer a large field of view (up to 195°). Numerous manufacturers like BMW, Mercedes, Toyota and Hyundai, integrate such systems into their around-view monitors (TrueCar, nd). The market size of these systems was estimated around USD 2734.6 million in 2022 with an expected increase to USD 43815.1 million by 2031 (Business Research Insights, nd).

Despite that, images captured by fisheye cameras exhibit strong radial distortions that are particularly intensified with larger fields of view. It falls short in providing an accurate and comprehensive understanding of the distant environment since mapping real-world coordinates onto perspective views alters the objects appearance leading to hindered views and occlusions between objects (see Figure 1). Moreover, the scarcity of open datasets featuring fisheye images often deters their use in tasks beyond around view montiors, such as perception systems.

Conversely, Bird Eye View (BEV) perception has proven to tremendously enhance the vehicles perceptual capabilities by using a configuration of six pinhole cameras mounted around the vehicle. BEV perception provides rich semantic information, including precise objects scaling and positioning, resulting in accurate scene representations. This is particularly pertinent in the context of parking assistance systems where BEV images are showcased to assist the driver, demanding high fidelity and an accurate representation of the real scene.

Direct acquisition of BEV images is challenging due to the need for costly equipment (such as drones or helicopters) to properly setup acquisition sensors. Alternatively, current parking assistance systems utilize an established perspective transform called *Inverse Perspective Mapping (IPM)* (Mallot et al., 1991). This transform projects pixels of surround-view images into a flat plane by computing a homography matrix that relies on the camera calibrations. IPM images are used in many perception tasks, such as parking slots detection (Zhang et al., 2018; Do and Choi, 2020; Wang et al., 2023) and Simultaneous Localization and Mapping (SLAM) tasks (Lee et al., 2023).

One significant drawback of IPM lies in its assumption that the world is flat since objects with

Benaissa, R., Musabini, A., Benmokhtar, R., Bakthavatchalam, M. and Perrotton, X.
BevGAN: Generative Fisheye Cross-View Transformers.
DOI: 10.5220/0013122200003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 2*, pages 153-163
ISBN: 978-989-758-737-5; ISSN: 2184-433X

heights (e.g. vehicles, trees and poles) are severely distorted and thus occlude other objects that are not directly in the camera's line of sight. Additionally, errors in camera calibrations contribute to IPM's failure to seamlessly merge the projected perspective views (see Figure 1f).

In commercial vehicles, the coverage range of the BEV images is limited to a very short distance around the ego-vehicle (i.e., ±5 meters) to diminish image deformations. Consequently, IPM is primarily employed for parking applications rather than broader driving scenarios which limits the drivers visibility of their surroundings and prevents them from having a satisfactory user experience (Musabini et al., 2021).

This work aims to overcome these challenges and improve the driver experience across both parking and driving scenarios. To the best of our knowledge, this is the first attempt to generate BEV images using Generative Adversarial Networks (GAN) by exploiting BEV features that are directly obtained from distorted surround-view fisheye images. Our main contributions are :

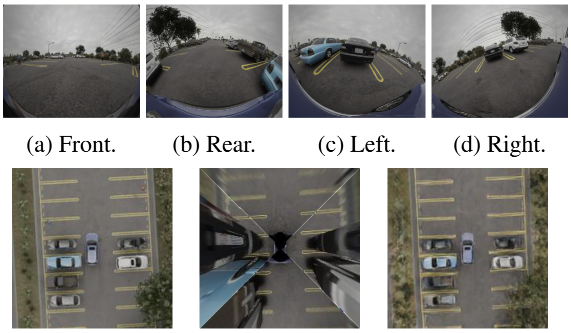
- A novel geometry-guided Conditional Generative Adversarial Networks (cGANs) model referred to as *BevGAN*. The proposed architecture leverages robust mechanisms for features extraction and inherently acquires mappings from individual camera views to infer BEV features representation. The latter are then transformed into BEV images all in a GAN-like framework.
- *BevGAN* generates BEV images of significantly higher quality compared to IPM and generative state-of-the-art methods. Furthermore, BevGAN generation capabilities span across diverse driving scenarios, while the BEV coverage range is expanded to a distance of $\pm 12m$ around the ego-vehicle.

2 RELATED WORK

2.1 Vision-Centric BEV Perception

Vision-centric BEV perception focuses on transforming one or multiple view inputs into BEV representations that are subsequently used to perform crucial downstream perception tasks (i.e. objects detection, lanes and maps segmentation, etc...). Current literature is divided into two main streams: geometry-based and network-based transformations (May et al., 2022).

Geometry-based transformations explicitly leverages the geometric properties of physical cameras to establish a natural projection relationship between the views. A pioneering work is IPM, as introduced in Section 1. It efficiently transforms the views through a homography matrix that is derived from intrinsic and extrinsic parameters of the camera. However, it falls short in meeting the intricacies of real-world driving as it relies on a rigid flat plane assumption that causes noticeable distortions on objects lying above the flat plane (for example vehicles, buildings, pedestrians). Despite other attempts to reduce the distortions by performing IPM on objects footprint segmentation maps (Loukkal et al., 2021; Can et al., 2020) and feature maps (Reiher et al., 2020) to align with the flat ground assumption, it is still present.



(a) Front. (b) Rear. (c) Left. (d) Right.

(e) Ground truth. (f) IPM. (g) BevGAN.

Figure 1: We propose a geometry-guided generative adversarial framework (BevGAN) that generates a BEV image from four surround-view fisheye images. Compared to IPM(1f), BevGAN fuses surround-view images into a unified BEV grid where the positioning of each object in the scene is known through a learnt transform. In (1g), no distortions are observed and the objects appearances are preserved thanks to GANs capabilities in generating unseen areas, resulting on a scene that is more faithful to the ground truth (1e).

Network-based methods implicitly incorporate camera geometry using neural networks to learn a cross-view mapping (Yang et al., 2021; Zou et al., 2023). In particular, transformers-based networks are widely used in dense and sparse perception tasks for their expressiveness (May et al., 2022). These methods utilize cross-attention between the BEV queries and the input image features to transform the views (May et al., 2022; Zhou and Krahenbuhl, 2022; Jiachen et al., 2022; Bartoccioni et al., 2022; Yang et al., 2023).

However, capitalizing on the rich BEV representations to generate BEV images remains a nascent field of research as we report only one work (Gieruc et al., 2024) that exploits TriPlane representations to generate BEV images.

Moreover, only few prior works consider fisheye camera geometry (Samani et al., 2023) (Musabini et al., 2024) or use a combination of both fisheye and pinhole cameras (Pham et al., 2024).

2.2 Generative Adversarial Networks

GANs were introduced to synthesize new images from a complex, high-dimensional training distribution (Goodfellow et al., 2020). It consists of two networks trained adversarially : a generator G that learns to generate photo-realistic images from a noise vector and a discriminator D that learns to distinguish between the generated and ground truth images.

Many variants of GANs use novel networks architectures and loss functions to leverage specific tasks like super-resolution (Ledig et al., 2017), text-to-image translation (Zhang et al., 2017) or images editing (Pan et al., 2023).

Among Conditional Generative Adversarial Networks (cGANs) applications, image-to-image translation (also called *cross-domain image translation*) achieved prominent success in mapping images of a source domain X to a target domain Y. Early endeavors (Zhu et al., 2017; Isola et al., 2017; Wang et al., 2018) concentrated on transforming aligned source and target domains, where geometrical transformations between the domains views are insignificant.

Recent approaches (Jain et al., 2021; Zhu et al., 2018) exploit semantic segmentation or geometry information (such as depth or homography estimation) to handle unaligned domains. BridgeGAN (Zhu et al., 2018) generates a BEV image of a driving scene from a single frontal view image by incorporating the homography image as an intermediate view to a multi-GAN framework designed such that frontal, homography and bird eye views share the same feature representation. However, this method highly depends on homography view and distortions are still present. A video-to-video translation model (Jain et al., 2021) was also introduced to generate BEV driving sequences from egocentric RGB videos using an additional branch that estimates the optical flow map to ensure temporal consistency between the generated frames. However, this method produces blurrier images over time due to accumulated generations errors and some inconsistencies related to global structure of the scene are also reported.

Shifting to semantic-guided approaches, the generator in (Regmi and Borji, 2019; Wu et al., 2023) is forked to produce both target-view images and segmentation maps to learn rich semantic features through the optimization of alignment losses. In (Ren et al., 2021; Tang et al., 2020a; Tang et al., 2019; Ren et al., 2023), the generator is conditioned by real target-view segmentation maps to learn shared features that progressively align domains. Overall, semantic-guided frameworks mostly depend on target-view segmentation maps at inference time or have heavy, complex architectures that hardly learn efficient mappings between the views.

To date, none of the existing methods handle fisheye images and multiple surround-view images fusion. In this work, we formulate our task as a cross-view image-to-image translation task and propose a geometry-based model that involves four surround-view fisheye images (i.e. front, rear, left and right fisheye images). Instead of relying on hard geometry transforms, it learns an estimate of the 3D objects locations in the scene to construct pertinent BEV mappings. The generative capabilities of our model allow to translate these mappings into a BEV image that faithfully represents the scene.

3 METHODOLOGY

3.1 The Proposed Cross-View Transformers-Based GAN

Our goal is to generate a high-fidelity BEV image considering a set of four surround-view fisheye images $(\mathbf{I_k} \in \mathbb{R}^{W\times H\times 3})_{k=1}^{4}$ each having a corresponding camera intrinsics matrix $\mathbf{K_k} \in \mathbb{R}^{3\times 3}$, extrinsic rotation matrix $\mathbf{R_k} \in \mathbb{R}^{3\times 3}$ and translation vector $\mathbf{t_k} \in \mathbb{R}^{3}$ relative to the center of the ego-vehicle.

To accomplish that, we introduce a novel geometry-guided cross-view image-to-image translation GAN architecture, referred to as BevGAN. Using cameras calibration and attention mechanisms, it implicitly models geometrical transformations between the views to produce rich BEV features representation.

BevGAN overall architecture is presented in Figure 2. It comprises multi-scale discriminators and a transformers-based generator designed with an encoder and decoder built upon cross-view transformer (Zhou and Krahenbuhl, 2022) and Pix2PixHD (Wang et al., 2018) frameworks respectively.

The key elements of BevGAN are outlined in the following sections.

3.1.1 BEV Grid Construction

BEV grid representations are constructed by adapting the dense query-based encoder of Cross-view Transformers (Zhou and Krahenbuhl, 2022) to take input images acquired from four fisheye cameras, instead of the original six pinhole cameras.

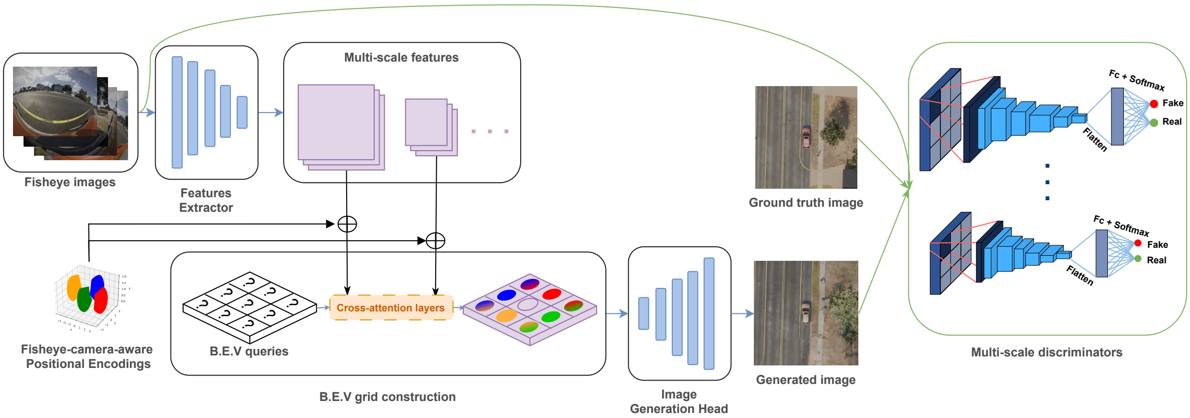


Figure 2: **BevGAN architecture overview**. Image features across multiple resolutions are extracted from surround-view fisheye images. Combined with Positional encodings (see Figure 3) obtained from cameras position and intrinsics, the BEV grid is constructed through a serie of cross-attention layers and passed to the decoder to generate the final BEV image. The BEV image along with ground truth image are given to multi-scale discriminators for the discrimination process.

First, a features extractor builds up multi-scale features ϕ_k for each input image (see Figure 2) (Zhou and Krahenbuhl, 2022). Two different resolutions are considered, each processed independently to be passed into a cross-view attention mechanism. The latter matchs up a BEV representation with input images features by implicitly modeling the geometry transformation between the views.

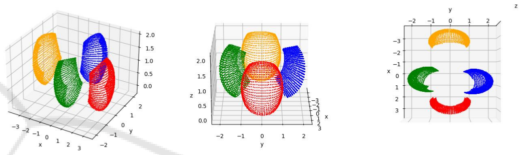
Starting from the lowest resolution, learnt BEV embeddings are refined through the projection of higher resolutions to better estimate the 3D location of each object in the scene.

3.1.2 Fisheye-Camera-Aware Positional Encoding

For each image coordinate $x_i^{(I)}$, the angle of incidence $\alpha_{k,i}$ emanating from each fisheye camera lens is computed based on its distance $r_{k,i}$ from the principal point (center of projection) and the radial distortions coefficients of each camera $(c_{k,l})_{l=1}^{4}$.

$$r_{k,i} = c_{k,1}\alpha_{k,i} + c_{k,2}\alpha_{k,i}^2 + c_{k,3}\alpha_{k,i}^3 + c_{k,4}\alpha_{k,i}^4 \qquad (1)$$

Equation 1 depicts the relationship between the angle of incidence $\alpha_{k,i}$ and the distance $r_{k,i}$ following the Polynomial FishEye Transform (PEFT). The computed polynomial root, $\alpha_{k,i}$, encodes the depth information of a new world coordinate. The latter is multiplied by the inverse of the camera extrinsic rotation matrix R_k and translation vector t_k to obtain the direction vector $d_{k,i}$ which represents a fisheye camera positional encoding of the image coordinate $x_i^{(I)}$. Figure 3 illustrates a multi-view 3D visualization of the fisheye camera positional encodings derived from a single-frame scene.



(a) Side view. (b) Front view. (c) Top view.

Figure 3: A multi-view 3D visualization of fisheye-camera positional encodings derived from a single-frame scene. The vehicle's center is positioned at pixel coordinates $(0,0)$ and the z-axis depicts height above ground level. The positional encoding of each camera is represented by a color : red for front camera, yellow for rear camera, blue for left camera and green for right camera.

The direction vector $d_{k,i}$ is encoded, using an MLP shared across all k views, into a D-dimensional positional embedding $\delta_{k,i} \in \mathbb{R}^{128}$.

3.1.3 Image Generation Head

The proposed decoder is constructed upon the global generator of Pix2PixHD (Wang et al., 2018) which has proven to be effective for high-resolution image-to-image translation and style transfer tasks (Johnson et al., 2016). It consists of a sequence of 9 residual blocks followed by a transposed convolutional backend (which is composed of 3 blocks of 3×3 convolution of stride $\frac{1}{2}$, instance normalization and ReLU layers).

3.1.4 Multi-Scale Discriminators

In order to generate coherent scene representation with finer details, it is necessary to design a discriminator that has a large receptive field. Inspired by previous works (Wang et al., 2018; Tang et al., 2020b), we adopt multi-scale discriminators which comprise three identical convolutional PatchGAN classifiers (Isola et al., 2017) that operate at different scales. For each discriminator $D_k, k = \{1,2,3\}$, we downsample input images by a factor of 2^{k-1}.

3.2 Optimization Scheme

Three optimization loss functions are considered in BevGAN : i) a least-square adversarial loss L_{lsgan} (Mao et al., 2017) , ii) a feature matching loss L_{FM} (Isola et al., 2017) and iii) a VGG perceptual loss L_{VGG} (Isola et al., 2017). The generator G and discriminators $D_k, k = \{1,2,3\}$ are alternately optimized according to the following optimization problem :

$$\begin{aligned} L_{BevGAN} = arg \min_{G,D} \sum_{k=1}^{3} L_{lsgan}(G, D_k) + \\ \lambda(\sum_{k=1}^{3} L_{FM}(G, D_k) + L_{VGG}(G)) \end{aligned} \quad (2)$$

where λ controls the importance given to the visual losses L_{FM} and L_{VGG}.

4 EXPERIMENTS

4.1 Parallel Domain Dataset

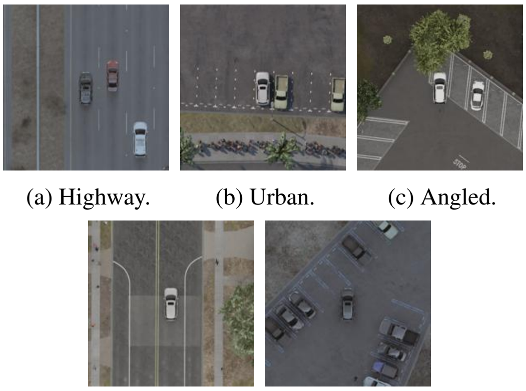
The dataset used in this work was procedurally generated using the synthetic data platform developed at Parallel domain (Parallel Domain Plateform, nd). To ensure visual diversity in the images, three different scenario types were generated:

1. Highway scenario
2. Urban scenario with high pedestrian density
3. Parking scenarios with the following parking slot variations :
   - Angled parking slots
   - Parallel parking slots
   - Perpendicular slots

Each scenario contains high fidelity renderings (and all the associated annotations) for four surround-view cameras and a BEV camera viewing the scene orthogonally as illustrated in Figure 4.

4.2 Evaluation Metrics

Drawing from previous works (Wu et al., 2023; Ren et al., 2023; Tang et al., 2020a; Regmi and Borji, 2019; Zhu et al., 2018; Ren et al., 2021), the following quantitative metrics have been selected to assess the quality of the generated images.



(a) Highway. (b) Urban. (c) Angled.

(d) Parallel. (e) Perpendicular.

Figure 4: Illustration of BEV samples from each scene type of Parallel Domain dataset.

- **Peak Signal-to-Noise Ratio (PSNR)** operates at the pixel-level. It is defined as the ratio between the maximum of a signal f and the power of corrupting noise that affects the fidelity of its representation.
- **Multi-Scale Structural SIMilarity (MS-SSIM)**(Wang et al., 2003) value ranges from 0 to 1. MS-SSIM compares two images patches at multiple scales based on their luminance, contrast and structure.
- **Sharpness Difference (SD)** measures the similarity between real and generated images in terms of sharpness by computing the difference between their gradients.
- **Learned Perceptual Image Patch Similarity (LPIPS)** computes the difference between the feature maps patches of the real and generated images extracted at different layers of a pre-trained model (a VGG-network, AlexNet or SqueezeNet...) using L1-distance.
- **Frechet Inception Distance (FID)**(Heusel et al., 2017) uses frechet distance to measure differences in density between two distributions at the deepest layers of an Inception-v3 classifier.

4.3 Implementation Details

In all experiments, we use an *EfficientNet-V2 Medium* (Tan and Le, 2021) as the features extractor and set the weight $\lambda = 10$ (see Equation 2). Other networks parameters remain unchanged. BevGAN is trained on a *40G NVIDIA A100 GPU* for 80 epochs following the splitting scheme established in Table 1. We use AdamW optimizer and the one-cycle learning rate scheduler (Smith and Topin, 2018) with a learning rate of 10^{-3} for both generator and discriminators.

Table 1: Training and test sets splitting scheme for Parallel Domain Dataset.

Scenes	Highway	Urban	Parking		
			Angled	Parallel	Perpendicular
Training set images	5350	6850	6720	8625	5350
Test set images	650	650	745	975	650
Total	6000	7500	7465	9600	6000

The batch size is set to 9 which is the largest size that the model can accommodate within the GPU capacity.

Surround-view fisheye images are resized to 640×528 and BEV images are cropped and resized such that it covers an area of $\pm 12m$ around the ego-vehicle with a corresponding pixels resolution of 200×200 for a BEV grid size of 25×25.

4.4 Results

4.4.1 State-of-the-Art Comparison

The proposed BevGAN is compared to three methods : IPM, Pix2PixHD and PanoGAN. IPM is the cutting-edge method deployed in parking assistance and monitoring systems to transform multiple inputs into a single BEV image.

Pix2PixHD originally generates high-resolution images from a single input of a different domain (for example, edges-to-image or labels-to-image translation (Wang et al., 2018)). Because BevGAN employs similar decoder and discriminators, contrasting the two models gives an assessment of the transformer-based encoder ability to construct a detailed BEV grid.

On the other hand, PanoGAN is the state-of-the-art method for cross-view image-to-image translation. It is designed to handle more significant geometrical transformations since it synthesizes a panorama image from a single BEV image.

Pix2PixHD and Panogan are adapted and trained on Parallel Domain dataset following the training scheme in 1, using the same hyper-parameters as set in their respective papers. Both models initially take one single input image. To accommodate to our task, the four fisheye images were arranged in a 2 by 2 grid then fed to the models as a single input.

Quantitative Evaluation. The quantitative evaluation results obtained on Parallel Domain dataset are presented in Table 2. It is noticeable that BevGAN achieves the best scores on all presented metrics. In comparison to IPM, an improvement of $+6.2db$ on PSNR, $+170\%$ on MS-SSIM and $+22\%$ on SD is observed. Moreover, BevGAN outperforms both Multi-input PanoGAN and Multi-input Pix2PixHD by a large margin as we report an improvement of $+2.66db$ and $+1.87db$ on PSNR and $+35\%$ and $+20\%$ on MS-SSIM respectively. These results indicate that BevGAN is able to produce BEV images with higher quality than other leading methods. Scores achieved on high-level evaluation metrics (LPIPS and FID) demonstrate the expressiveness of the BEV representations obtained with BevGAN as it allows the generation of diverse, high-quality images.

Table 3 presents quantitative evaluation of BevGAN on Parallel Domain dataset for each scene type. It is worth mentioning that BevGAN performs better on highway scenes in contrast to urban and parking scenes. In fact, parking and urban scenes simulate more complex and detailed environments, featuring high pedestrians density in urban areas and a variety of parking slot types and markings in parking scenarios (see Figure 4). Moreover, only a small number of scenes include challenging conditions (like night and snow scenes). Plus, these scenes are predominant in parking scenarios making it difficult for the model to effectively learn their characteristics.

Qualitative Evaluation. Figure 5 illustrates qualitative results obtained with leading methods on various scenes of Parallel Domain dataset. It is observable that BevGAN produces more realistic BEV images across all scenarios closely matching the corresponding ground truth images. Compared to IPM, all generative methods provide distortion-free representations of the generated objects. Although, BevGAN is able to identify all objects in the scene and represents them more accurately in terms of shape, color, and sharpness.This holds true even in complex scenarios (see row 2 and 5) where textured objects with varied sizes (vehicles, pedestrians, etc...) and various road and parking slots markings, are present. Furthermore, BevGAN synthesizes a scene where invisible regions (ie. regions that are originally invisible in surround-view images) are consistent with respect to the visible parts of the scene (see the vehicles roofs, vehicles occluded by the blue vehicle in row 5).

4.4.2 Ablation Study

An ablation study was conducted on Parallel Domain dataset to assess the effectiveness of the proposed

Table 2: Quantitative evaluation of BevGAN against state-of-the-art methods on Parallel Domain Dataset.

Method	PSNR (↑)	MS-SSIM (↑)	LPIPS (↓)	SD (↓)	FID (↓)
IPM	17.14	0.30	0.47	0.22	249.46
Multi-input PanoGAN	20.68	0.60	0.40	0.23	176.80
Multi-input Pix2PixHD	21.47	0.67	0.29	0.20	114.33
BevGAN (ours)	23.34	0.81	0.17	0.18	52.7

Table 3: Quantitative evaluation of BevGAN on Parallel Domain Dataset based on scenes type.

Scenes		PSNR (↑)	MS-SSIM (↑)	LPIPS (↓)	SD (↓)	FID (↓)
Urban		23.11	0.8	0.18	0.23	89.29
Highway		25.54	0.85	0.13	0.13	42.00
Parking	Angled	23.5	0.81	0.19	0.20	89.75
	Parallel	23.26	0.83	0.19	0.14	113.87
	Perpendicular	21.75	0.82	0.17	0.13	116.30

Table 4: Ablation study of the proposed BevGAN on Parallel Domain Dataset.

Variant	Description	PSNR (↑)	MS-SSIM (↑)	LPIPS (↓)	SD (↓)	FID (↓)
A	generator + VGG	18.76	0.67	0.32	0.18	108.82
B	generator + discriminator + VGG	21.5	0.72	0.23	0.20	67.20
C	generator + discriminator + FM	22.65	0.77	0.22	0.19	83.12
BevGAN	generator + discriminator + all losses	23.34	0.81	0.17	0.18	52.7

BevGAN model. Table 4 provides a quantitative comparison of various BevGAN variants. *Variant A* includes the generator only, which is trained in a supervised manner using input fisheye images and corresponding BEV images. In this configuration, the optimization focuses solely on minimizing the VGG loss since that both adversarial and FM losses depend on the discriminator (see Equation 2). Compared to this variant, BevGAN demonstrates a significant increase in performances on all evaluation metrics with a difference of $4.58db$ in PSNR and a increase of $+21\%$ in MS-SSIM. These results confirm that the use of a generative model is more adequate for the task.

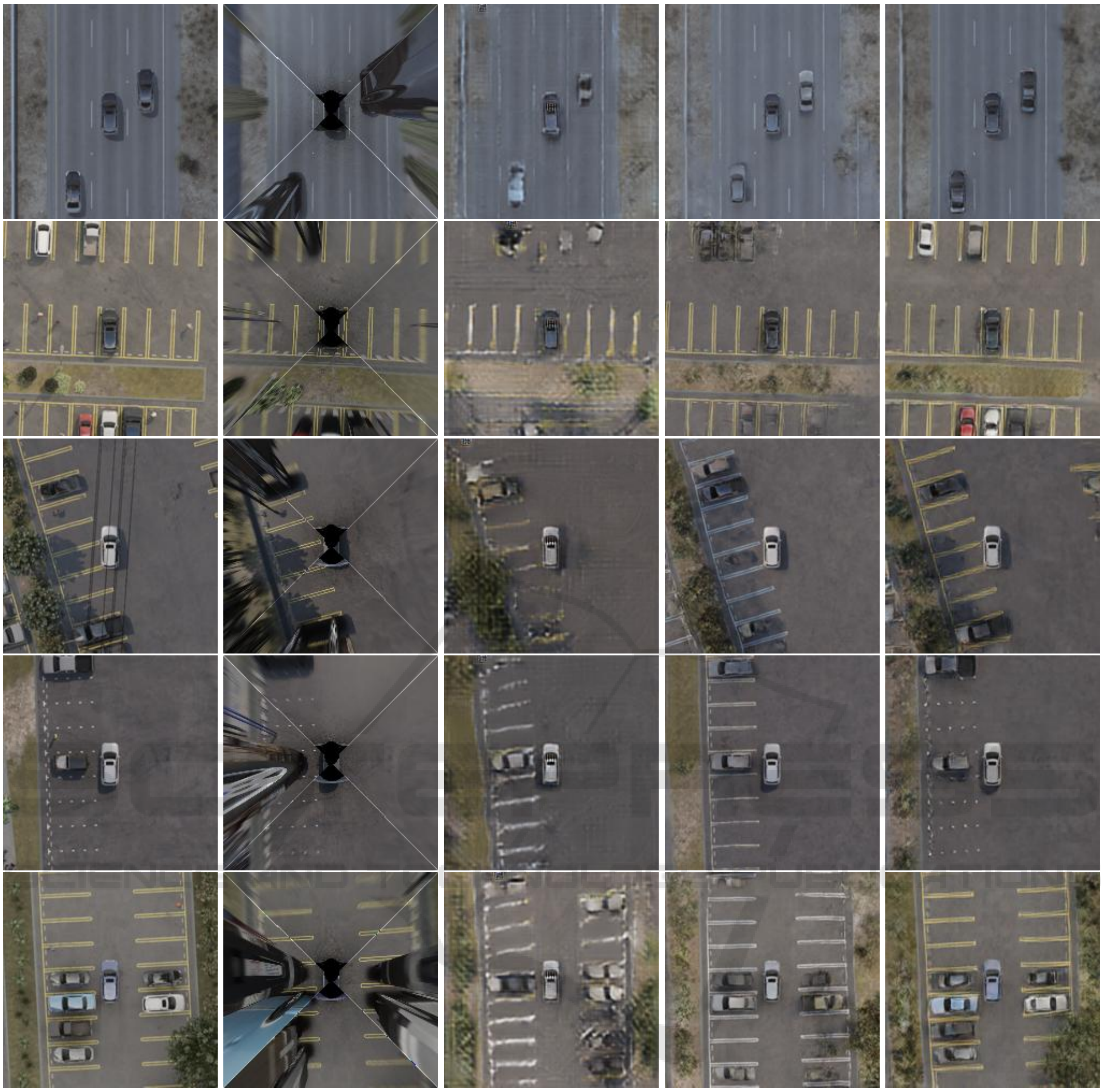
The two remaining variants (*variants B and C*) include both the generator and discriminator. They evaluate the impact of each visual loss on the quality of the generated images by employing only one visual loss at a time. Results from *Variant B* indicate that adding the FM loss substantially improves the performances. Although adding VGG loss also improves the results, its effect is less pronounced compared to the feature matching loss, as observed from evaluation results of *Variant C*.

4.4.3 Generalization Ability

Given the non-existence of a publicly available dataset containing paired real-world surround-view fisheye and corresponding BEV images, all prior experiments were conducted on Parallel Domain's synthetic dataset. When transitioning to real-world images, the current implementation of BevGAN shows poor generalization capabilities (see Figure 6). Since that direct training or fine-tuning on a real dataset is not feasible, we chose to retrain our model using the following data augmentations to improve its generalization capabilities.

- Image-level data augmentations : image dropout with a probability $p \in [0.0, 0.01]$, sharpening, change in brightness and hue channels.
- Roll rotations of one degree per camera.
- BEV Flip Left/Right and rotations of fixed degrees (90° , 180° and 270°) on yaw axis with a probability of 0.8.

Table 5 provides a comparative evaluation between BevGAN and *BevGAN*†, the variant trained using previously cited augmentations. Despite that BevGAN outperforms *BevGAN*† showing a difference of $0.64db$ in PSNR and 20% in MS-SSIM, *BevGAN*† demonstrates superior generalization capabilities, as highlighted in Figure 6. Additional zero-shot inferences of IPM and *BevGAN*† on our internal real dataset are illustrated in Figure 7. The generated images demonstrate that *BevGAN*† accurately locates vehicles and parking slots in the scene. Nevertheless, we report some distortions and a drop in textural quality as the captured distributions depend on the synthetic dataset on which the model was origi-



(a) GT. (b) IPM. (c) PanoGAN. (d) Pix2PixHD. (e) BevGAN.

Figure 5: Qualitative comparison between different methods on Parallel Domain dataset. From left to right : Ground Truth, IPM, Multi-input PanoGAN, Multi-input Pix2PixHD and BevGAN (ours). The proposed BevGAN generates more realistic results with finer details on all scenes settings compared to other state-of-the-art methods.

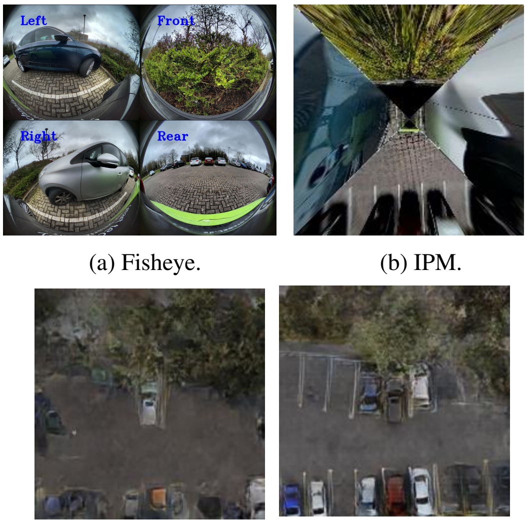
nally trained. That being said, *BevGAN*† struggles to reproduce objects not encountered during the training phase, such as buildings and road markings, which are absent from Parallel Domain Dataset.

Table 5: Quantitative evaluation of BevGAN against *BevGAN*† on Parallel Domain Dataset.

Method	PSNR (↑)	MS-SSIM (↑)	LPIPS (↓)	SD (↓)	FID (↓)
BevGAN	23.34	0.81	0.17	0.18	52.7
BevGAN†	22.7	0.76	0.23	0.2	70

5 CONCLUSION

In this work, we introduce BevGAN : a novel geometry-guided cGANs model designed to generate a BEV image from a set of four surround-view fisheye images. BevGAN integrates two key elements : a cross-view transformer-based generator and multi-scale discriminators. The generator operates across multiple scales and leverages fisheye-camera-aware positional embeddings to generate high-quality BEV



(a) Fisheye. (b) IPM.

(c) BevGAN. (d) *BevGAN*†.

Figure 6: Example of zero-shot inference of BevGAN on real-world fisheye images.

images that faithfully represent the scene.

Experiments conducted on a synthetic dataset demonstrate that the proposed BevGAN outperforms leading methods used in parking assistance and vision monitoring systems, as well as state-of-the-art GAN approaches for cross-view generation. Moreover, experiments show that with just a few added augmentation strategies, BevGAN can effectively generalize to images acquired from real cameras.

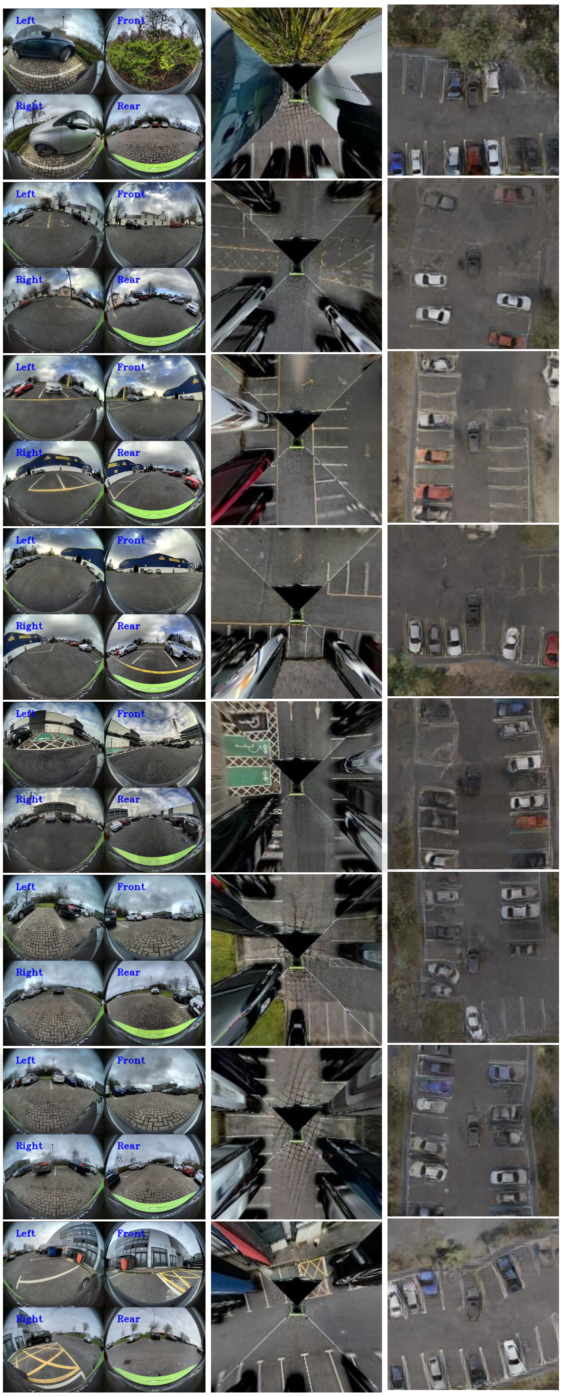
BevGAN introduces promising performances for practical integration into advanced around-view systems for real-world vehicles. Notably, our method expands the covered area around the ego-vehicle to a range of $\pm 12m$, which is a significant improvement compared to the $\pm 5m$ coverage range offered by current systems.

Future works will focus on exploring novel view representations like Tri-Perspective View (TPV) representation (Huang et al., 2023) (Gieruc et al., 2024) for a robust description of the 3D scene. Additionally, steps will be taken to collect a more diverse dataset encompassing a larger range of driving and parking scenarios.

REFERENCES

Bartoccioni, F., Zablocki, E., Bursuc, A., Perez, P., Cord, M., and Alahari, K. (2022). Lara: Latents and rays for multi-camera bird's-eye-view semantic segmentation. In *6th Annual Conference on Robot Learning*.

Business Research Insights (n.d.). Business research insights. Accessed: Febuary 2024.



(a) Fisheye images. (b) IPM. (c) *BevGAN*†.

Figure 7: Qualitative comparison between IPM and augmented BevGAN on real-world fisheye images.

Can, Y. B., Liniger, A., Unal, O., Paudel, D. P., and Gool, L. V. (2020). Understanding bird's-eye view of road semantics using an onboard camera. *IEEE Robotics and Automation Letters*, 7:3302–3309.

Do, H. and Choi, J. Y. (2020). Context-based parking slot detection with a realistic dataset. *IEEE Access*, 8:171551–171559.

Gieruc, T., Kästingschäfer, M., Bernhard, S., and Salzmann, M. (2024). 6img-to-3d: Few-image large-scale outdoor driving scene reconstruction.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Commun. ACM*, 63(11):139–144.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6629–6640, Red Hook, NY, USA. Curran Associates Inc.

Huang, Y., Zheng, W., Zhang, Y., Zhou, J., and Lu, J. (2023). Tri-perspective view for vision-based 3d semantic occupancy prediction.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *CVPR*.

Jain, V., Wu, Q., Grover, S., Sidana, K., Chaudhary, D.-G., Myint, S., and Hua, Q. (2021). Generating bird's eye view from egocentric rgb videos. *Wireless Communications and Mobile Computing*, 2021:1–11.

Jiachen, L., Zheyuan, Z., Xiatian, Z., Hang, X., and Li, Z. (2022). Learning ego 3d representation as ray tracing.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 694–711, Cham. Springer International Publishing.

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. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, Los Alamitos, CA, USA. IEEE Computer Society.

Lee, Y., Kim, M., Ahn, J., and Park, J. (2023). Accurate visual simultaneous localization and mapping (slam) against around view monitor (avm) distortion error using weighted generalized iterative closest point (gicp). *Sensors*, 23(18).

Loukkal, A., Grandvalet, Y., Drummond, T., and Li, Y. (2021). Driving among flatmobiles: Bird-eye-view occupancy grids from a monocular camera for holistic trajectory planning. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 51–60.

Mallot, H., Bülthoff, H., J.J., L., and S, B. (1991). Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biological cybernetics*, 64:177–85.

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., and Smolley, S. P. (2017). Least squares generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, Los Alamitos, CA, USA. IEEE Computer Society.

May, Y., Wangy, T., Baiy, X., Yang, H., Hou, Y., Wang, Y., Qiao, Y., Yang, R., Manocha, D., and Zhu, X. (2022). Vision-centric bev perception: A survey.

Musabini, A., Bozbayir, E., Marcasuzaa, H., and Ramírez, O. A. I. (2021). Park4u mate: Context-aware digital assistant for personalized autonomous parking. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 724–731. IEEE.

Musabini, A., Novikov, I., Soula, S., Leonet, C., Wang, L., Benmokhtar, R., Burger, F., Boulay, T., and Perrotton, X. (2024). Enhanced parking perception by multi-task fisheye cross-view transformers.

Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., and Theobalt, C. (2023). Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*.

Parallel Domain Plateform (n.d.). Parallel domain. Accessed: Febuary 2024.

Pham, T., Maghoumi, M., Jiang, W., Jujjavarapu, B. S. S., Sajjadi, M., Liu, X., Lin, H.-C., Chen, B.-J., Truong, G., Fang, C., et al. (2024). Nvautonet: Fast and accurate 360deg 3d visual perception for self driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7376–7385.

Regmi, K. and Borji, A. (2019). Cross-view image synthesis using geometry-guided conditional gans. *Computer Vision and Image Understanding*, 187:102788.

Reiher, L., Lampe, B., and Eckstein, L. (2020). A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–7.

Ren, B., Tang, H., and Sebe, N. (2021). Cascaded cross mlp-mixer gans for cross-view image translation. In *British Machine Vision Conference*.

Ren, B., Tang, H., Wang, Y., Li, X., Wang, W., and Sebe, N. (2023). Pi-trans: Parallel-convmlp and implicit-transformation based gan for cross-view image translation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Samani, E., Tao, F., Dasari, H., Ding, S., and Banerjee, A. (2023). F2bev: Bird's eye view generation from surround-view fisheye camera images for automated driving. *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9367–9374.

Smith, L. and Topin, N. (2018). Super-convergence: very fast training of neural networks using large learning rates. In *Defense + Commercial Sensing*.

Tan, M. and Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training.

Tang, H., Xu, D., Sebe, N., Wang, Y., Corso, J. J., and Yan, Y. (2019). Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. *2019 IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition (CVPR), pages 2412–2421.

Tang, H., Xu, D., Yan, Y., Torr, P. H. S., and Sebe, N. (2020a). Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7867–7876.

Tang, W., Li, G., Bao, X., Nian, F., and Li, T. (2020b). Mscgan: Multi-scale conditional generative adversarial networks for person image generation. In *2020 Chinese Control And Decision Conference (CCDC)*, pages 1440–1445.

TrueCar (n.d.). Truecar. Accessed: Febuary 2024.

Wang, L., Musabini, A., Leonet, C., Benmokhtar, R., Breheret, A., Yedes, C., Bürger, F., Boulay, T., and Perrotton, X. (2023). Holistic parking slot detection with polygon-shaped representations. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5797–5803. IEEE.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, Z., Simoncelli, E., and Bovik, A. (2003). Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398 – 1402 Vol.2.

Wu, S., Tang, H., Jing, X.-Y., Zhao, H., Qian, J., Sebe, N., and Yan, Y. (2023). Cross-view panorama image synthesis. *IEEE Transactions on Multimedia*, 25:3546–3559.

Yang, C., Chen, Y., Tian, H., Tao, C., Zhu, X., Zhang, Z., Huang, G., Li, H., Qiao, Y., Lu, L., et al. (2023). Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839.

Yang, W., Li, Q., Liu, W., Yu, Y., Ma, Y., He, S., and Pan, J. (2021). Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15531–15540.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*.

Zhang, L., Huang, J., Li, X., and Xiong, L. (2018). Vision-based parking-slot detection: A dcnn-based approach and a large-scale benchmark dataset. *IEEE Transactions on Image Processing*, 27(11):5350–5364.

Zhou, B. and Krahenbuhl, P. (2022). Cross-view transformers for real-time map-view semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13750–13759, Los Alamitos, CA, USA. IEEE Computer Society.

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

Zhu, X., Yin, Z., Shi, J., Li, H., and Lin, D. (2018). Generative adversarial frontal view to bird view synthesis. In *2018 International Conference on 3D Vision (3DV)*, pages 454–463.

Zou, J., Xiao, J., Zhu, Z., Huang, J., Huang, G., Du, D., and Wang, X. (2023). Hft: Lifting perspective representations via hybrid feature transformation for bev perception. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7046–7053.