Generative Adversarial Frontal View to Bird View Synthesis

Xinge Zhu[†] Zhichao Yin[§] Jianping Shi[§] Hongsheng Li[†] Dahua Lin[†]
[†]CUHK-SenseTime Joint Lab, CUHK
[§]SenseTime Research
{zhuxinge, yinzhichao, shijianping}@sensetime.com
hsli@ee.cuhk.edu.hk, dhlin@ie.cuhk.edu.hk

Abstract

Environment perception is an important task with great practical value and bird view is an essential part for creating panoramas of surrounding environment. Due to the large gap and severe deformation between the frontal view and bird view, generating a bird view image from a single frontal view is challenging. To tackle this problem, we propose the BridgeGAN, i.e., a novel generative model for bird view synthesis. First, an intermediate view, i.e., homography view, is introduced to bridge the large gap. Next, conditioned on the three views (frontal view, homography view and bird view) in our task, a multi-GAN based model is proposed to learn the challenging cross-view translation. Extensive experiments conducted on a synthetic dataset have demonstrated that the images generated by our model are much better than those generated by existing methods, with more consistent global appearance and sharper details. Ablation studies and discussions show its reliability and robustness in some challenging cases. Codes are available at https://github.com/WERush/BridgeGAN

1. Introduction

View synthesis is a long-standing problem in computer vision [5, 11, 25, 33, 35], which facilitates many applications including surrounding perception and virtual reality. In modern autonomous driving solution, the limited viewpoint of on-car cameras restricts the system from reliably understanding the environment, acquiring accurate global view for better policy making and path planning. Due to the large view gap (90 degrees in our task) and severe deformation, generating the bird view image from the front view is not even easy for human being. In this paper, we would like to push the envelop of synthesis between two drastically different views, although the challenging nature of this problem leaves room for further improvements. To our best knowledge, it is the first attempt to generate the bird view based on single frontal view image, which serves better perceptual understanding and sparks future researchers to explore information from multiple views for perception.

Figure 1. (**a**): The frontal view image. (**b**): The ground truth bird view image. (**c**): The intermediate homography view image. (**d**): The generated bird view image by our model.

With the 3D structure of the scene, bird view generation can be easily achieved by changing the view point and projection. However, when the only input is a frontal view image, it will be substantially more difficult. The same object has different appearances and sizes in the images of bird view and frontal view. Meanwhile, the semantic representation and basic color should be consistent between two views. Imagining a car in front of you, after transforming to bird view, it should be the same car but with completely different appearance and size. An example is shown in Fig. 1 (a) and (b). The large gap and severe deformation make it a challenging task and it is far from being solved.

Traditional methods for bird view generation are generally built on multi-sensor systems [22, 29, 36]. Recently, by regarding the view transformation as a view synthesis task, many 3D view synthesis methods obtained promising results by modeling the underlying 3D geometry [5, 33]. Image-based rendering models [4, 41], on the other hand, generate new views by re-using the pixels from source images. However, these methods can only transform already visible content, e.g., they cannot render the top view of a car from the input frontal view or side-view images.

Given the large gap between the frontal view and the bird view images, they can be naturally regarded as two different domains. With the vigorous study of generative adversarial

network [6], many powerful cross-domain image translation systems [12, 34, 38, 42, 43] have been proposed, which can generate images with plausible appearance. The representative work pix2pix [10] utilizes a conditional adversarial network, converting an image from one representation of a given scene to another, e.g. semantic labels to images, edge-map to photograph. However, these models could only perform translations for the aligned images in color or texture level, e.g. from zebra to horse or from greyscale to color. Translating images across two domains with a large gap in between (e.g. the images captured from different viewpoints) remains a challenging task even with the latest GAN-based techniques.

In this paper, we propose the **BridgeGAN** model, a novel bird view generation model from single frontal view images. To bridge the large gap between the two views, we incorporate the homography view as the intermediate view, with a homography matrix [1] to perform the perspective mapping, as shown in Fig. 1 (c). It serves as a bridge to connect two views. Hence, this is where the 'Bridge' comes from. The homography view serves to decrease the gap between frontal view and bird view, but it produces undesired distortions. Conditioned on the three views in our task, a multi-GAN based model is proposed to learn the cross-view translation. We extend the cycle-consistency loss [42] to a dual cycle-consistency loss for matching the three views and constraining the cross-domain translation to be a one-to-one correspondence. Furthermore, a cross-view feature consistency loss is designed to make all three views have a shared feature representation in low (e.g. color) and high (e.g. content) level. The final result is shown in Fig. 1 (d). Experimental results demonstrate that, by generating images consistent in terms of global structure and details, our method results in significantly better performance compared to the baselines. Ablation studies verify the effect of each components, and discussions show the reliability of our model in the case of challenging scenario.

The main contributions are summarized as follows:

(1) To our best knowledge, we are the first to address the novel problem of generating bird view image based on a single frontal view image, which enables better perceptual understanding and multiple views perception.

(2) We propose the BridgeGAN, i.e., a novel generative model for bird view synthesis, in which the homography view is first introduced to bridge the gap and then a multi-GAN based model is proposed to perform cross-view transformation.

(3) Extensive experiments demonstrate that the proposed model generates significantly better results compared with baselines, which is able to preserve the global appearance and the details of objects. More discussions also verify its reliability and robustness.

2. Related Work

Bird-Eye View There are few works in literature that aim to tackle the problem of perspective transformation. Most of these methods are geometry based. More specifically, Lin *et al.* [15] introduced a fitting parameters searching algorithm to estimate a perspective matrix for image coordinate transformation. Similarly, in [30], an inverse perspective matrix was used to perform the view transformation. However, the largest problem of such kind of methods is the distortion, especially the region in a distance. Another group of methods are vision based. [18, 36] achieved a bird-eye view by stitching images from a four to six fish-eye lens cameras system. Sung *et al.* [29] proposed a camera parameter optimization algorithm to establish surround view from multi-camera images. However, in most cases, a multi-camera system is not available in the vehicle and the common source is the frontal view image. Moreover, such methods cannot create new view invisible from existing cameras.

View Synthesizing A large body of view synthesizing works are geometry based. With a huge amount of multi-view images, 3D stereo algorithms [5] are applicable to reconstruct the 3D scene and then be utilized to synthesize novel views. Ji *et al.* [11] proposed to synthesize middle view images by using two rectified view images. Yan *et al.* [33] proposed a perspective transformer network to learn the projection transformation after reconstructing the 3D volume of the object. However, most of these methods are trained with 3D supervision and all view pairs can be generated via a graphics engine. In our setting, the training data is limited in both views and numbers and the 3D supervision is also unavailable.

Synthesis between Remote Sensing Image and Ground-level Image Existing works [35, 25] explored to predict the semantic segmentation of ground level image from its remote sensing image, and then apply the semantic layout to synthesize the ground image. However, the remote sensing image and ground-level image have the much higher viewpoint and lack of textures, which are significantly different from our on-car camera image and bird view image. Besides, these methods also require the segmentation mask to supply the synthesis, which is not feasible in our setting. Furthermore, due to the little or no overlap between remote sensing image and ground image, this view synthesis desires the model to imagine a large unseen region while our setting has more proper inference with more overlap.

Image Generation Recently, image generation has been a heated topic with the emerge of Generative Adversarial Networks (GANs) [6]. A GAN consists of two modules, a generator G and a discriminator D. These two parts G and D play a minimax game. G is trained to generate images to confuse the D, and D is trained to distinguish between real

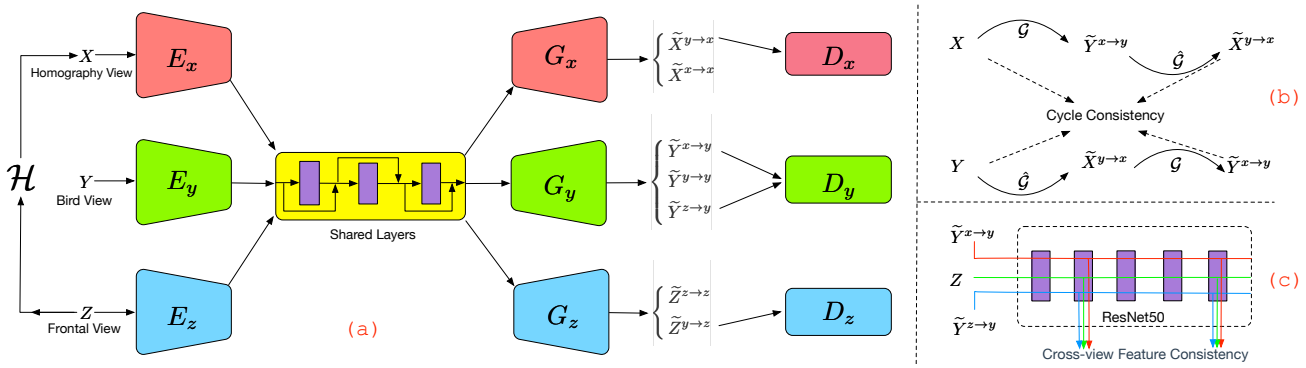

Figure 2. (a) Our model contains three GANs representing three domains, i.e. homography view(X), Bird view(Y) and Frontal view(Z). Each of them consists of three modules: Encoder(E), Generator(G) and Discriminator(D). Since all three domains have a shared semantic meaning, we employ shared layers for learning unified intermediate representations. Gs learn the cross-domain translations and reconstruction mappings based on the intermediate representation. Ds are adversarial discriminators for the respective domains, distinguishing between real and fake samples. (b) Single cycle-consistency loss includes two components, i.e. forward cycle: $X \xrightarrow{\mathcal{G}} \widetilde{Y}^{x\to y} \xrightarrow{\hat{\mathcal{G}}} \widetilde{X}^{y\to x} \approx X$ and backward cycle: $Y \xrightarrow{\hat{\mathcal{G}}} \widetilde{X}^{y\to x} \xrightarrow{\mathcal{G}} \widetilde{Y}^{x\to y} \approx Y$. (c) To enforce the consistent cross-view features, we introduce the loss network (ResNet-50) for regularizing the inconsistent problem.

and fake samples.

Various methods were developed to generate images based on GANs. Conditional generative adversarial nets (CGANs) [20] used the labels as the conditional information to both generator and discriminator and generated digit images of specified label. Pix2pix model [10] was built on the CGANs to learn the mapping with full supervision. Similar ideas have been applied to numerous tasks such as generating images from sketches or from text [24, 28, 37]. Many researchers also tried to attempt multiple GANs for learning a mapping from input domain to target domain. CoupledGANs [17] used a weight-sharing strategy to learn a common representation across domains. CycleGAN [42] and DiscoGAN [12] introduced an inverse mapping and a cycle consistency loss to constrain the mapping between two domains. DualGAN [34] and UNIT [16] also employed another GAN to learn to invert the image translation task. Our model is also multi-GAN based. But unlike these image translation approaches with two GANs, our model applies three GANs with two consistency constraints to strengthen the generation capability due to the large gap between the frontal view and the bird view image.

3. Methodology

Bird view synthesis can be viewed as a cross-domain image generation task, in which the source domain is frontal view and target domain is bird view. Unlike these traditional cross-domain image translation tasks that are performed between two aligned images [12, 42], such as from zebras to horses and photos to paintings, bird view synthesis is more challenging because the large gap and severe deformation exist between frontal view and bird view. To tackle this challenge, we introduce an intermediate view, homography view, to bridge the gap in our task, then a multi-GAN based model is proposed to realize the cross-domain translation.

3.1. Framework Overview

Our full BridgeGAN is illustrated in Fig. 2 (a). There are three domains: X for homography view, Y for bird view and Z for frontal view. The BridgeGAN learns by synthesizing one view from another via GAN, enabling the network to learn the intermediate representation shared between different views and the reconstruction ability upon the intermediate representation.

More specifically, the BridgeGAN model consists of three GANs for the three domain representations and translations as a multi-GAN system. Each GAN contains an Encoder (E) to transform image to an intermediate representation, and a Generator (G) to transform the intermediate representation to a new image. In addition, there exists a discriminator (D) to distinguish between generated images and real images, which drives the generative image towards real one.

During training, the bird view domain is chosen as a pivot and two cross-domain translations exist in the pipeline, i.e. between homography view (X) and bird view (Y), and between frontal view (Z) and bird view (Y). We introduce shared layers to enforce the GANs to share some higher-layer parameters for a consistent intermediate representation between views. After training, the proposed framework generates bird view images by two steps: first doing a homography estimation from Z to X and then performing cross-domain translation from homography view (X) to bird view (Y).

Mappings	I_x	$\mathcal{G}$	$\mathcal{F}$	$\hat{\mathcal{G}}$	$\hat{\mathcal{F}}$
Subnetworks	$\{E_x, G_x\}$	$\{E_x, G_y\}$	$\{E_z, G_y\}$	$\{E_y, G_x\}$	$\{E_y, G_z\}$

Table 1. The relation between mappings and subnetworks in our model. For the mapping $\mathcal{G} : X \rightarrow Y$, it is composed of two subnetworks, *i.e.* Encoder E_x and Generator G_y. In addition, I_y and I_z have the same formation as I_x.

3.2. Multi-GAN Learning

We design the multi-GAN system to learn the cross-domain translations for bird view generation from both frontal view and homography view.

View Transformation. Our model builds upon seven view mappings between all three domains in our multi-GAN system, including cross-domain mappings $\mathcal{G} : X \rightarrow Y$; $\mathcal{F} : Z \rightarrow Y$, their inverse mappings $\hat{\mathcal{G}} : Y \rightarrow X$; $\hat{\mathcal{F}} : Y \rightarrow Z$ and three identity mappings I_x, I_y, I_z. The cross domain mapping captures the view transformation ability whereas the identity mapping ensures the reconstruction consistency. Specified, the homography view and frontal view are complementary parts to realize the bird view translation, where the homography view decreases the gap from frontal view to bird view and the frontal view provides more global context that homography view lacks of in turn. We do not include mapping between frontal view Z and homography view X since it is a fixed perspective transformation. Each view transformation is represented by a GAN. Table 1 summarizes the mappings and corresponding implementations in our model.

Full Objective. Our objective contains three terms: *adversarial loss* $\mathcal{L}_{\text{GAN}}$ for matching the distribution of generated image to the distribution of target image, *dual cycle-consistency loss* $\mathcal{L}_{\text{cyc}}$ for constraining the cross-domain mappings to be a one-to-one correspondence and to be well covered on bi-directions (bijective mappings), and *cross-view feature consistency loss* $\mathcal{L}_{\text{cfc}}$ for encouraging generated bird view image and frontal view image to keep the feature representations consistent in low and high level, such as color and content. Our full objective is:

$$\mathcal{L} = \mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{cyc}} + \mathcal{L}_{\text{cfc}}. \tag{1}$$

Adversarial Loss. Inheriting from GAN [6], we apply the adversarial losses to these cross-domain mappings, i.e. $\mathcal{G}$, $\mathcal{F}$, $\hat{\mathcal{G}}$ and $\hat{\mathcal{F}}$. This objective enforces our model to learn mapping from its input domain to target domain.

For the mapping $\mathcal{G} : X \rightarrow Y$, we express the objective as:

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(\mathcal{G}, D_y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log(D_y(y)] \\ + \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(1 - D_y(\mathcal{G}(x)))],\end{aligned} \tag{2}$$

where $\mathcal{G}$ aims to generate the image that looks similar to image from target domain Y, while D_y tries to identify the real and fake samples. Similarly, we introduce this objective for other cross-domain mappings, i.e. $\mathcal{F}$, $\hat{\mathcal{G}}$ and $\hat{\mathcal{F}}$. Hence, we get the objectives, $\mathcal{L}_{\text{GAN}}(\mathcal{F}, D_y, Z, Y)$, $\mathcal{L}_{\text{GAN}}(\hat{\mathcal{G}}, D_x, Y, X)$ and $\mathcal{L}_{\text{GAN}}(\hat{\mathcal{F}}, D_z, Y, Z)$, respectively. The total adversarial loss is the sums of cross-domain mappings, which is expressed as:

$$\begin{aligned}\mathcal{L}_{\text{GAN}} = (\mathcal{L}_{\text{GAN}}(\mathcal{G}, D_y, X, Y) + \mathcal{L}_{\text{GAN}}(\hat{\mathcal{G}}, D_x, Y, X)) \\ + (\mathcal{L}_{\text{GAN}}(\mathcal{F}, D_y, Z, Y) + \mathcal{L}_{\text{GAN}}(\hat{\mathcal{F}}, D_z, Y, Z)).\end{aligned} \tag{3}$$

Furthermore, these generators are tasked to not only fool the discriminators but also to be near the ground truth image in the pixel level. For pixel-level loss of these generators, we use L1 distance rather L2 as L1 encourages less blurring.

Dual Cycle-consistency Loss. To reduce the space of the possible mapping functions and enforce the mappings between two domains to be a one-to-one bijective mapping, a cycle-consistency loss [12, 34, 42] was introduced. We extend the cycle-consistency loss to a dual cycle-consistency loss for matching the three domains in our task. Each cycle consistency serves a cross-domain mapping and its inverse mapping, which measures how well the origin input is reconstructed after a sequence of two generations.

The first cycle consistency is for the mapping $\mathcal{G} : X \rightarrow Y$ and its inverse mapping $\hat{\mathcal{G}} : Y \rightarrow X$. As illustrated in Fig. 2 (b), we depict the pipeline of forward cycle consistency and backward cycle consistency, i.e. $X \xrightarrow{\mathcal{G}} \widetilde{Y}^{x \rightarrow y} \xrightarrow{\hat{\mathcal{G}}} \widetilde{X}^{y \rightarrow x} \approx X$ and $Y \xrightarrow{\hat{\mathcal{G}}} \widetilde{X}^{y \rightarrow x} \xrightarrow{\mathcal{G}} \widetilde{Y}^{x \rightarrow y} \approx Y$. Thus the first cycle consistency objective can be given by:

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(\mathcal{G}, \hat{\mathcal{G}}) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\| \hat{\mathcal{G}}(\mathcal{G}(x)) - x \|_1] \\ + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\| \mathcal{G}(\hat{\mathcal{G}}(y)) - y \|_1],\end{aligned} \tag{4}$$

where the L1 norm is applied to measure the distance between $\hat{\mathcal{G}}(\mathcal{G}(x))$ and x.

Similarly, for another cross-domain mapping $\mathcal{F} : Z \rightarrow Y$ and its inverse mapping $\hat{\mathcal{F}}$, the second cycle consistency objective is:

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(\mathcal{F}, \hat{\mathcal{F}}) = \mathbb{E}_{z \sim p_{\text{data}}(z)}[\| \hat{\mathcal{F}}(\mathcal{F}(z)) - z \|_1] \\ + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\| \mathcal{F}(\hat{\mathcal{F}}(y)) - y \|_1].\end{aligned} \tag{5}$$

Finally, we express the dual cycle consistency objective function as:

$$\mathcal{L}_{\text{cyc}} = \lambda_1(\mathcal{L}_{\text{cyc}}(\mathcal{G}, \hat{\mathcal{G}}) + \mathcal{L}_{\text{cyc}}(\mathcal{F}, \hat{\mathcal{F}})). \tag{6}$$

Cross-view Feature Consistency Loss. Considering two images from frontal view and bird view, the semantic contents and colors are consistent among them. We postulate

there exists a cross-view feature consistency among three different views in our task. However, the adversarial loss could make the generated image look realistic but its contents may not be closely relevant to the source, and the dual cycle-consistency loss enforces a one-to-one constraint on cross-domain mappings rather than a joint consistency on low- and high-level features among three domains. Therefore, we design another consistency loss to enforce the constraint of consistent cross-view feature representations among all three views.

Since the generated bird view image is not aligned with frontal view, direct pixel-wise loss (L1, L2, etc.) is not suitable. We introduce a loss network ϕ pretrained for image classification to satisfy this constraint. Instead of minimizing the pixel-level distance, we encourage the generated bird view image and real frontal view image to have similar low- and high-level feature representations extracted from loss network ϕ. Feature maps extracted from early layer model the low-level features and feature maps from higher layer serve as high-level representations. The loss function is the normalized L2 distance:

$$\begin{aligned}\mathcal{L}_{\phi}(\widetilde{Y}, Z) = &\frac{1}{C_j H_j W_j} \parallel \phi_j(\widetilde{Y}) - \phi_j(Z) \parallel_2^2 \\ &+ \frac{1}{C_k H_k W_k} \parallel \phi_k(\widetilde{Y}) - \phi_k(Z) \parallel_2^2,\end{aligned} \tag{7}$$

where $\widetilde{Y}$ and Z are generated bird view and frontal view image, respectively. j and k denote the jth and kth layer in the loss network. C, H and W are the shape of feature maps. The cross-view feature consistency can be given by:

$$\begin{aligned}\mathcal{L}_{\text{cfc}} = \mathcal{L}_{\text{cfc}}(\mathcal{G}, \mathcal{F}, Z, \phi) = &\mathbb{E}_{x \sim p_{\text{data}}(x)} \mathcal{L}_{\phi}(\mathcal{G}(x), Z) \\ &+ \mathbb{E}_{z \sim p_{\text{data}}(z)} \mathcal{L}_{\phi}(\mathcal{F}(z), Z),\end{aligned} \tag{8}$$

where $\mathcal{G}(x)$ and $\mathcal{F}(z)$ are the generated bird view images from homography view and frontal view, respectively. As shown in Fig. 2 (c), in our experiments, ϕ is the resnet-50 [9] pretrained on the ImageNet [2] dataset. We utilize the feature maps from $res2c$ layer as the low-level feature and $res5c$ as the high-level representation.

3.3. Iterative Optimization

For the training of GANs, it can be viewed as playing a minimax game until finding a saddle point. In the proposed framework, these encoders and generators are in one team versus another team composed of adversarial discriminators. Our full optimization objective is:

$$\min_{E_x, E_y, E_z, G_x, G_y, G_z;} \max_{D_x, D_y, D_z} \mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{cyc}} + \mathcal{L}_{\text{cfc}}. \tag{9}$$

We apply an alternative optimization approach, which iteratively updates the network blocks in the following order: (1) Update discriminators: the network is first optimized by maximizing the discriminators' accuracy with encoders and generators fixed. (2) Update encoders and generators by minimizing the loss with discriminators fixed.

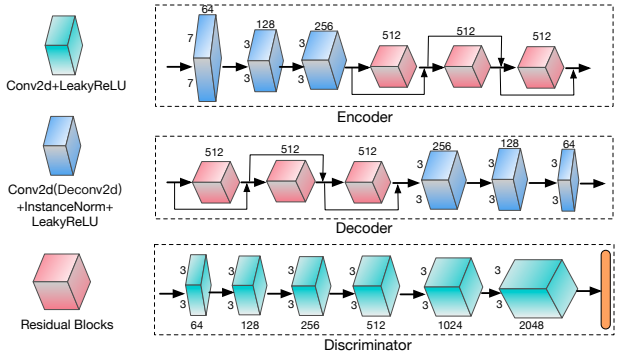

Figure 3. Detailed architectures of subnetworks in our proposed framework, including encoder, decoder and discriminator. Basic components are shown in the left column. Note that the convolutional layers would be replaced by deconvolutional layers [19] in the decoder and the shape of each layer is labelled.

3.4. Implementation Details

As illustrated in Fig. 3, we elaborate the structure of encoder, decoder and discriminator in our model. There are three basic components shown in the left column. Inheriting from [27], we utilize the LeakyReLU and instance normalization [31] to improve training stability. The encoder and decoder are both composed of six blocks and detailed architectures are shown in the figure. For the residual block, we use the same architecture as resnet [9].

Shared Layers. As shown in Fig. 2 (a), shared layers are composed of three stacked residual blocks [9] between encoder and decoder networks. The features from three domains are handled by shared layers for enforcing a unified intermediate representation.

Homography Estimation. We follow the traditional homography estimation pipeline [1] which is composed of two stages: corner estimation and robust homography estimation. For the stage I, we utilize the ORB [26] features as the descriptor. For the stage II, a parameter searching algorithm RANSAC [3] is used as the estimator. Then a direct linear transform (DLT) algorithm [8] is applied to get the homography matrix $\mathcal{H}$. We can obtain the homography view by using the equation $X = \mathcal{H}Z$. More details are introduced in Section 4.2.

4. Experiments

4.1. Dataset and Evaluation Metrics

Dataset. In [23], they developed a framework to collect data from Grand Theft Auto V video game, in which the game camera automatically toggles between frontal and bird view at each time step. In this way, they gathered in-

formation about the road scene from both views. We download the dataset from their official website[1]. As shown in the Fig. 1 (a), we remain the region right in front of the vehicle, which is the common part between the frontal view and the bird view. After data processing, we obtain a training set with 40,000 pairs of images and a testing set with 4,000 pairs.

Evaluation Metrics. We use two traditional metrics PSNR, SSIM [32], and a neural network based metric LPIPS [39]. PSNR relies on low-level differences. SSIM mainly reflects the perceived change in the structural information. LPIPS uses pretrained deep models to evaluate the similarity, which highly agrees well with humans. Specifically, we use the pretrained AlexNet in the LPIPS[2] metrics.

4.2. Training Setup

For the homography estimation, we utilize multi-scale ORB features as the descriptor and select the top 30 scoring matches as the input to the RANSAC estimator. After obtaining the homography matrix, we perform the homography estimation based on the cropped frontal view image to get the homography view image. All three view images are resized to 320x192 as the inputs.

We apply the PatchGAN [14] to the discriminators, in which the discriminator tries to classify whether overlapping image patches are real or fake. Similar to EBGAN [40], we add the gaussian noise to the shared layers and generator. During training, Adam [13] optimization is applied with β_1=0.5 and β_2=0.999. We train the model on a single Titan X GPU with learning rate=0.0001. Each mini-batch contains three frontal view images, three homography view images and three bird view images. For the weighted factor λ_1, we apply $\lambda_1 = 10$, which is chosen by using a cross-validation method.

4.3. Experimental Results

Baseline methods ***Pix2pix*** [10]. This method proposes a conditional GAN for image translation. Without hand-engineering loss functions, the conditional adversarial networks could also achieve reasonable results.

CycleGAN [42]. Like our method, CycleGAN is also a multi-GAN model. It consists of two GANs representing two domains and a cycle-consistency loss is used to regularize the cross-domain mapping.

DiscoGAN [12]. DiscoGAN is another multi-GAN based model which is proposed to discover the cross-domain relations.

CoGAN [17]. This method consists of two GANs, and a strategy of weights sharing is introduced. Unlike our method sharing the representation in the higher layer, CoGAN shares the weights on its first few layers.

[1]http://imagelab.ing.unimore.it/scene-awareness

[2]https://github.com/richzhang/PerceptualSimilarity

Homo Since bird view generation is a perspective transformation in geometry, the homography estimation is also a viable solution. We term it as Homo and regard it as a baseline method.

Comparison against baselines Since pairs of images are provided in the dataset, we train all baseline methods in a **supervised** manner for a fairness, which means that during the training of baselines, **paired images (homography view and corresponding bird-view images) are fed into those baselines**. We use the public implementations of these baseline methods.

We compare our results with five baseline methods on the GTAV dataset. The evaluation scores for the proposed method and compared methods are reported in Table 4.3. We can see that DiscoGAN and CoGAN get the worse scores. CycleGAN and Pix2pix have similar scores and CycleGAN achieves a marginally higher results, while all these methods have worse scores compared to Homo. Our proposed model achieves the best results against all other baselines. The better score indicates that our proposed model generates images which are more realistic and reasonable, and further more similar to the ground truth images in the global structure.

In order to make intuitive comparisons, we display some representative examples generated by our proposed model and baselines in Fig. 4. In each example, there are eight images, which always appear in the same order; from left to right: frontal view, ground truth, our proposed model, Homo, CycleGAN, DiscoGAN, CoGAN and Pix2pix. It can be seen that the samples generated by the DiscoGAN and CoGAN are blurry, and the structure of vehicles in the image generated by CoGAN is not correct. CycleGAN generates images with more details, but also presents severe artifacts and incorrect position of object. Pix2pix maintains the basic structure, but severely lacks of details and suffers from the mode collapse problem (e.g. images generated by Pix2pix in Fig. 4 (4) and (5)). Moreover, CycleGAN generally remains the homograpy view and suffers from distortion (e.g. images generated by CycleGAN in Fig. 4 (3) and (5)). Homo could produce reasonable images in the global appearance while it also has a severe distortion, especially for the object (e.g. the vehicle) in a distance. With the global context from frontal view and two consistent constraints, our proposed model produces more realistic images, which keep the global structure consistent with the ground truth and possess richer details and correct color. For instance, in the fourth example (Fig. 4 (3)), the image generated by our model has more sharper details (e.g. consistent color with the ground truth) and reasonable appearance of vehicle.

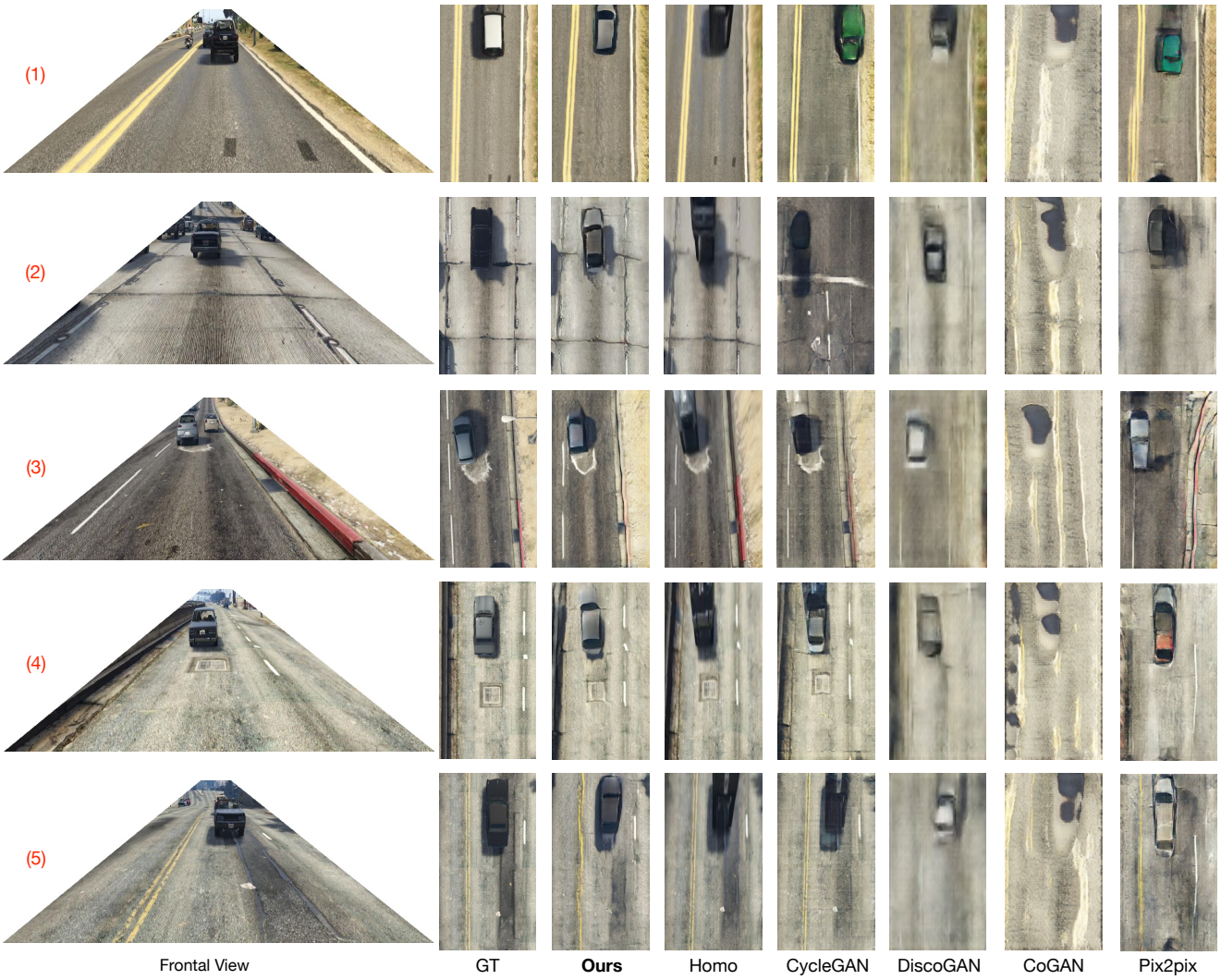

Figure 4. Example results by our proposed method and baselines. Best viewed in color.

Method	DiscoGAN	CycleGAN	CoGAN	Pix2pix	Homo	Ours
SSIM ↑	0.5342	0.5568	0.5453	0.5486	0.5715	**0.5961**
PSNR ↑	4.8432	4.8754	4.8602	4.8703	4.9234	**5.0056**
LPIPS ↓	0.3741	0.2902	0.3588	0.3083	0.2709	**0.2427**

Table 2. The results of our proposed method and baseline methods.

4.4. Ablation Study

To verify the contributions of different components in our model, we design several variants to perform an ablation study. The details of these variants are shown as follows:

- **Ours without domain** Z: We remove the frontal view Z. This variant only contains two GANs representing the left two domains.
- **Ours without domain** X: We remove the intermediate view X. This variant performs cross-view translation directly from the frontal view to bird view.
- **Ours without dual cycle-consistency**: We remove the dual cycle-consistency loss from the total loss.

Variant	Ours w/o Z	Ours w/o X	Ours w/o cyc loss	Ours w/o cfc loss	Ours
SSIM ↑	0.5726	0.5116	0.5701	0.5842	**0.5961**
LPIPS ↓	0.2634	0.3225	0.2679	0.2581	**0.2427**

Table 3. Results of ablation study. Cyc loss denotes the dual cycle-consistency loss, and cfc is the cross-view feature consistency loss.

- **Ours without Cross-domain feature consistency**: We remove the cross-domain feature consistency loss from the total loss.

Table 4.4 reports the results of ablation study. It can be found that removing the loss function, i.e. dual cycle-consistency or cross-view feature consistency, degrades the results of model, so does removing the domain X or domain Z. In particular, removing the domain X greatly worsens the quality of the generated images, which indicates the huge gap between frontal view and bird view and the importance of the intermediate view X. We therefore con-

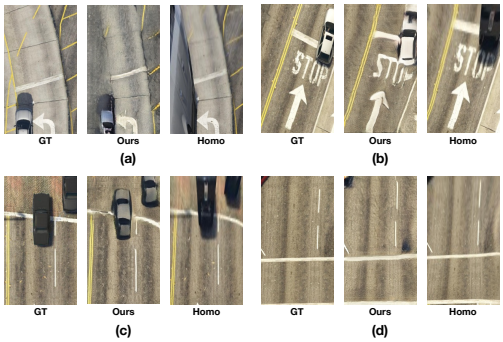

Figure 5. Some challenging examples. (a): The car is right in front of the camera. (b): The car is on the curved road. (c): There are multiple vehicles. (d): There is no vehicle. In these cases, the homography estimation fails and suffers severe distortions.

clude that all three views play a critical role in our model, and two consistency constraints are also important parts of improving quality of generated images.

4.5. Discussion

Could the model work well when the homography view fails? We can find that the proposed intermediate view bridges the large gap. Hence, could the model work well when the homography view fails or suffers severe distortions? In order to further verify the capability of our proposed model, we show some challenging samples in Fig. 5. In Fig. 5(a) and (b), the vehicles are right in front of the camera and on the curved road, respectively, where the homography view fails and suffers from severe distortions. The proposed model successfully generates the reasonable bird view with the failed homography estimation.

Does the model synthesize the cars that do not exist? In the normal GAN model, it might generate diverse outputs even non-exist things given the same input. However, our proposed model would not synthesize a non-exist car because the homography view is introduced as an input source which could provide initial location beliefs on the cars, even for the homography estimation with severe distortions it also supplies the stable beginning position of vehicles. As shown in Fig. 5(c) and Fig. 5(d), the proposed model successfully synthesizes a multi-car bird view image and a no-car bird view image under the guidance of the homography view, respectively. It demonstrates that our proposed model is reliable and would not hallucinate vehicles that do not exist.

4.6. User Study

Automatic evaluation measures, such as SSIM, could reduce the cost of time and human effort. However, these methods are not always reliable. In [7, 21], they found that some generated images are not more realistic but score very highly, which indicates that these automatic measures are not fully correlated with human perception. Thus, the user study is applied to supplement the evaluation, in which the generated images are evaluated by human observers.

Model	DiscoGAN	CycleGAN	CoGAN	Pix2pix	Homo	Ours
Score1	0.0249	0.1622	0.0134	0.0875	0.2730	**0.4390**
Score2	0.0173	0.1736	0.0251	0.0517	0.3299	**0.4024**

Table 4. Results of two types of user study experiments. Score1 is the result for the first experiment and the Score2 is for the second.

We conducted two types of user study experiments on this dataset for quantitative evaluation. In the first experiment, we showed 110 groups of images. Each group showed one ground truth image, and six images generated by baseline methods and our proposed model based on the same input image. Users were asked to choose an image that is most similar to the ground truth and has the best quality in terms of appearance and detail. In the second experiment, we showed 110 groups of images which were randomly sampled from the whole testing set, and users were asked to choose an image that has the best quality for each group. A total of 25 users participated in this user study.

We calculate the percentages of each model whose generated image is selected as the best image. The results of two experiments are shown in Table 4.6. It can be observed that our proposed model achieves the best performances in both experiments with 16% and 7% higher scores than baselines respectively, which indicates that our model could generate more realistic and reasonable images which better satisfy the human standard.

5. Conclusion

In this paper, we propose a novel cross-view translation model, i.e., BridgeGAN, to address the new problem of bird view synthesis from a single frontal view image. It can provide better perceptual understanding and enable future researchers with multiple views perception attempt. Specifically, we first introduce an intermediate view to bridge the huge gap between frontal view and bird view. Then the multi-GAN model is proposed to perform the cross-view translation. Two constraints are introduced to our proposed model to ensure one-to-one cross-domain translations and consistent cross-view feature representations. With extensive experiments, our model typically preserves the scene structure, global appearance and details of objects in the bird view. Quantitative evaluations and user study demonstrate the superiority of the proposed model. Ablation studies verify the importance of each components. More discussions and examples show its reliability even in the challenging cases.

References

[1] S. Baker, A. Datta, and T. Kanade. Parameterizing homographies. *Technical Report CMU-RI-TR-06-11*, 2006. 2, 5

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255. IEEE, 2009. 5

[3] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 5

[4] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *IEEE CVPR*, pages 5515–5524, 2016. 1

[5] Y. Furukawa, C. Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1, 2

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 2, 4

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 8

[8] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 5

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 5

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 2, 3, 6

[11] D. Ji, J. Kwon, M. McFarland, and S. Savarese. Deep view morphing. *arXiv preprint arXiv:1703.02168*, 2017. 1, 2

[12] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017. 2, 3, 4, 6

[13] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, pages 702–716. Springer, 2016. 6

[15] C.-C. Lin and M.-S. Wang. A vision based top-view transformation model for a vehicle parking assistant. *Sensors*, 12(4):4431–4446, 2012. 2

[16] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NIPS*, pages 700–708, 2017. 3

[17] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *NIPS*, pages 469–477, 2016. 3, 6

[18] Y. Liu and B. Zhang. Photometric alignment for surround view camera system. In *ICIP*, pages 1827–1831. IEEE, 2014. 2

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE CVPR*, pages 3431–3440, 2015. 5

[20] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[21] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE CVPR*, pages 427–436, 2015. 8

[22] F. Nielsen. Surround video: a multihead camera approach. *The visual computer*, 21(1):92–103, 2005. 1

[23] A. Palazzi, G. Borghi, D. Abati, S. Calderara, and R. Cucchiara. Learning to map vehicles into birds eye view. In *International Conference on Image Analysis and Processing*, pages 233–243. Springer, 2017. 5

[24] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016. 3

[25] K. Regmi and A. Borji. Cross-view image synthesis using conditional gans. *arXiv preprint arXiv:1803.03396*, 2018. 1, 2

[26] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *IEEE ICCV*, pages 2564–2571. IEEE, 2011. 5

[27] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NIPS*, pages 2234–2242, 2016. 5

[28] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays. Scribbler: Controlling deep image synthesis with sketch and color. *arXiv preprint arXiv:1612.00835*, 2016. 3

[29] K. Sung, J. Lee, J. An, and E. Chang. Development of image synthesis algorithm with multi-camera. In *Vehicular Technology Conference (VTC Spring), 2012 IEEE 75th*, pages 1–5. IEEE, 2012. 1, 2

[30] D. C. Tseng, T. W. Chao, and J. W. Chang. Image-based parking guiding using ackermann steering geometry. In *Applied Mechanics and Materials*, volume 437, pages 823–826. Trans Tech Publ, 2013. 2

[31] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5

[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[33] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *NIPS*, pages 1696–1704, 2016. 1, 2

[34] Z. Yi, H. Zhang, P. T. Gong, et al. Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv preprint arXiv:1704.02510*, 2017. 2, 3, 4

[35] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs. Predicting ground-level scene layout from aerial imagery. In *IEEE CVPR*, volume 3, 2017. 1, 2

[36] B. Zhang, V. Appia, I. Pekkucuksen, Y. Liu, A. Umit Batur, P. Shastry, S. Liu, S. Sivasankaran, and K. Chitnis. A surround view camera solution for embedded systems. In *IEEE CVPR Workshops*, pages 662–667, 2014. 1, 2

[37] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1612.03242*, 2016. 3

[38] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *ECCV*, pages 649–666. Springer, 2016. 2

[39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv preprint arXiv:1801.03924*, 2018. 6

[40] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016. 6

[41] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *ECCV*, pages 286–301. Springer, 2016. 1

[42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. 2, 3, 4, 6

[43] X. Zhu, H. Zhou, C. Yang, J. Shi, and D. Lin. Penalizing top performers: Conservative loss for semantic segmentation adaptation. In *ECCV*, 2018. 2