Semantic View Synthesis

Hsin-Ping Huang[1], Hung-Yu Tseng[2], Hsin-Ying Lee[2], Jia-Bin Huang[3]

[1]UT Austin [2]University of California, Merced [3]Virginia Tech

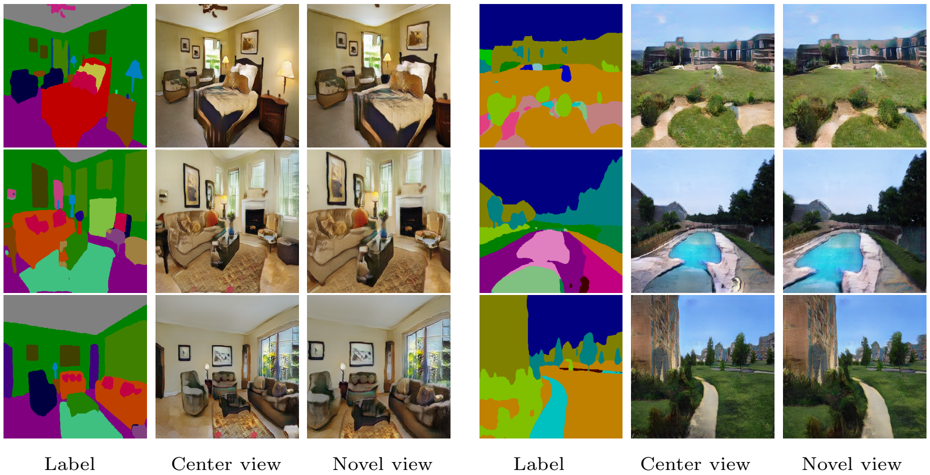


Fig. 1: **Semantic view synthesis.** We introduce a new visual synthesis problem, semantic view synthesis — synthesizing a *photorealistic* image that supports *free-viewpoint rendering* given a single semantic label map. To achieve such visual effects, we build a two-step inference pipeline upon recent advances in semantic view synthesis and novel view synthesis. We show that our model learns to generate scene representations for rendering geometrically consistent and semantically meaningful novel views. We demonstrate the efficacy of our method using a wide variety of indoor (*left*) and outdoor (*right*) scenes.

Abstract. We tackle a new problem of semantic view synthesis — generating free-viewpoint rendering of a synthesized scene using a semantic label map as input. We build upon recent advances in semantic image synthesis and view synthesis for handling photographic image content generation and view extrapolation. Direct application of existing image/view synthesis methods, however, results in severe ghosting/blurry artifacts. To address the drawbacks, we propose a two-step approach. First, we focus on synthesizing the color and depth of the visible surface of the 3D scene. We then use the synthesized color and depth to impose explicit constraints on the multiple-plane image (MPI) representation prediction process. Our method produces sharp contents at the original view and geometrically consistent renderings across novel viewpoints.

The experiments on numerous indoor and outdoor images show favorable results against several strong baselines and validate the effectiveness of our approach.

1 Introduction

Visual content creation using generative models has been gaining increasing attention. Driving by the advances in generative models, recent work has demonstrated impressive performance on a wide range of tasks, including image generation from various contexts (e.g., noises [12,24], images [22,56,26,20,1], text [51,43], and audio [28]), view interpolation and extrapolation [8,15,55,41,44], and image editing [2,5,42]. These algorithms greatly help unleash human imagination and support creative processes. In this paper, we introduce a new form of visual content creation task by integrating (1) semantic image synthesis and (2) novel view synthesis.

Semantic image synthesis [3,37,46,35] is a specific form of image-to-image translation task that aims to generate photorealistic images from semantic label maps. Such an application is intuitive as users can easily draw and refine the semantic map on a digital canvas and then use the algorithm to synthesize *2D images* with plausible appearances. As these algorithms produce only 2D outputs, it is challenging for users to manipulate the viewpoints of the synthesized image in a geometrically consistent manner.

View synthesis, on the other hand, takes a sparse set of real images (captured at different viewpoints) as inputs and synthesizes novel views of the same scene [7,15,55,41,44]. This is achieved by explicitly or implicitly modeling the *3D structure* of the scene. However, these methods are applicable only to real images.

In this paper, we propose to tackle a new problem: *semantic view synthesis* — generating free-viewpoint rendering of a synthesized scene using a semantic label map as input (Figure 1). Compared to the existing semantic image synthesis task, the semantic view synthesis problem offers two unique advantages (Figure 2). First, it allows the users to easily manipulate the viewpoints of the synthesized image with minimal effort. Second, it supports temporally and geometrically consistent rendering of 3D fly-through effects.

To enable this new application, we develop a two-step method, drawing inspirations from the recent advances in semantic image synthesis and view synthesis algorithms. First, given the input semantic label map, we leverage a state-of-the-art image synthesis model, SPADE [35], to generate a photorealistic color image and the corresponding disparity map. The synthesized color/disparity images capture the appearance and structure of the *visible surface* of the scene. Second, to handle the dis-occluded contents (which become visible at novel views), we infer a multiplane images (MPI) representation [55] using the synthesized color/disparity as constraints. The resulting output of our method is an MPI representation that naturally supports view synthesis at any viewpoints. We conduct extensive quantitative and visual comparisons on three datasets (ADE20K [53], ADE20k-outdoor [37], and NYUv2 [33]) covering various indoor and outdoor

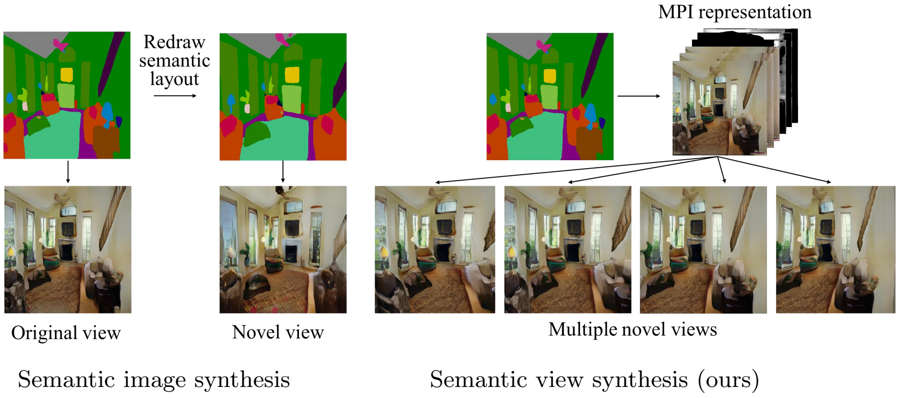


Fig. 2: **Application.** The new problem of semantic view synthesis offers two advantages over the existing semantic image synthesis task. (a) **Faster editing of viewpoints.** (*Left*) To refine the viewpoint of a synthesized image, the users would have to *redraw* the semantic layout of the scene and apply the image synthesis algorithm on the new semantic layout again to produce the desired view. (*Right*) Taking a single semantic layout as input, our method produces an MPI representation that naturally supports fast, free-form novel view rendering. (b) **Consistent rendering over viewpoints.** (*Left*) As novel view images are *independently* generated, the synthesized contents may not be consistent. (*Right*) Our semantic view synthesis, in contrast, enables 3D fly-through effects with plausible motion parallax.

scenes. Our results demonstrate clear improvement over several strong baseline methods and alternative designs.

In summary, we make the following contributions:

- We introduce a new *semantic view synthesis* task that aims to synthesize images of free-viewpoint from semantic masks.
- We propose a novel two-step training and inference pipeline: (1) color and disparity image synthesis for the visible surface and (2) MPI prediction with explicit constraints from the first step. (Section 3)
- We build several baseline approaches for this new problem and validate the efficacy of our proposed framework on a wide variety of indoor and outdoor scenes. (Section 4)

2 Related Work

Monocular depth prediction aims to estimate the depth of a scene from a *single-view* RGB image. It is a challenging problem due to the difficulty of obtaining explicit 3D cue from the single-view RGB image without additional information (e.g., stereo pair). To conquer the problem, several supervised learning schemes [9,19,25,48] utilize the ground-truth depth notation in the RGB-D dataset and train fully-convolutional networks (e.g., [31]) to capture the image prior. However, these approaches require large and diverse annotated data for the training. Numerous self-supervised approaches [10,54,59,49,11] have been proposed to avoid the labor-intensive annotating process. For instances, training with stereo videos [10], monocular videos [11], incorporating the information of camera poses or optical flow [49,54,59,58]. Nevertheless, these supervised and unsupervised methods often train their models using data from specific domains (e.g., driving scenes from the KITTI dataset) and therefore have difficulty in generalizing to diverse scenes in the wild. On the other hand, a line of approaches uses multi-view internet photos [30], MannequinChallenge [29] or 3D movies [38,45] as the source of data. In particular, training with mixed datasets from different sources achieves strong generality on unseen scenes. Our work leverage the pre-trained single-view depth estimation model from MiDaS [38] to obtain (pseudo) ground truth of depth/disparity maps for images in our training dataset.

Novel view synthesis aims to generate novel views based on single or multiple images. Earlier learning-based approaches [8,23] take multiple posed images as input and produce the target views by blending the warped input images. Such approaches, however, only *interpolate* among the given viewpoints and do not handle dis-occlusions. Recent advances explore generating novel view through a 3D scene representations, such as multi-plane images [7,55,41,32,44], layered depth images [6], mesh representations [14,40], and point clouds [47]. The multi-plane image representation [7,32,41,44,55] is a set of RGBA layers at discrete disparity levels. The novel views are rendered by homographic projection and alpha blending of the MPI layers. The layered depth image approach [6] represents 3D images as a foreground RGBD image and a background RGBD image. To generate the novel views, the RGB image is warped by the depth image, then composite by a predicted visibility mask. This approach requires supervision of the background image and only works for synthetic scenes. 3D photography [14,40] focuses on generating 3D effects for real-world photos; they represent 3D images as a multi-layer 3D mesh. These methods generate scene representation at the reference (original) viewpoint. The novel view images can be rendered by projecting the scene representation to the desired viewpoint.

Our work also produces an MPI representation as our output for supporting novel view synthesis. Our problem setting, however, differs significantly from prior MPI-based methods. Prior methods often require (at least) two images as inputs, which consist of the appearance of visible surfaces, cues of scene depth, and some content of the occluded background. In contrast, the input to our method is one semantic label map. Our experimental results show that direct application of prior MPI-based methods leads to severe blurry ghosting artifacts when rendered at novel views. Our two-step approach substantially reduces these artifacts via imposing explicit constraints on the MPI representation during training and testing time.

Image-to-image translation aims to learn the mapping between two image domains [1,20,22,27,56,57]. These techniques demonstrate a wide range of applications such as image inpainting, image super-resolution, domain adaptation [4,18], and semantic image synthesis [46,35]. In particular, semantic image synthesis learns to generate photo-realistic images conditioned on semantic label maps. Pix2pix [22] adopts a U-Net architecture to synthesize low-resolution images from a semantic map. To operate in high-resolution settings, Pix2pixHD [46] introduces the multi-scale generator and discriminator network structure to enhance the quality of the generated images. SPADE [35] further improves Pix2pixHD with the spatially-adaptive normalization layers. Different from the semantic image synthesis frameworks, we aim to synthesize 3D representation of a scene from a *single-view* semantic segmentation layout.

Cross-modal distillation transfers the knowledge between different modalities. Existing works [13,17] use learned representation from a large labeled dataset of the source modality as a supervised signal to train tasks of target modality with limited data. For example, the method in [13] utilize ImageNet-pretrained model to train new representations for optical flow and depth images. To address the problem of collecting a large indoor/outdoor dataset of semantic map to depth image pairs, our work also incorporates the idea of cross-modal distillation. Specifically, We transfer the knowledge of monocular depth prediction model (predicting depth maps from images) and semantic segmentation (predicting semantic layouts from images) to our *semantic depth synthesis* (predicting depth from semantic layouts). To this end, we present a two-branch version of a SPADE network [35] to predict both color and depth from a single semantic map.

3 Method

3.1 Overview

Our goal is to learn to synthesize novel-view color images from a given a semantic label map. As shown in Figure 3, our scene representation generation process consists of (1) image and disparity generation module and (2) MPI prediction module. With the generated MPI, we can project and blend the MPI to produce the desired target views. In this section, we first describe the data preparation in Section 3.2. We then detail the training procedure of scene representation generation including image and disparity generation and the MPI prediction in Section 3.3. Finally, we introduce the novel view synthesis procedure at test time in Section 3.4.

3.2 Data preparation

We build a dataset from the RealEstate10K dataset [55], which consists of 80,000 indoor/outdoor YouTube video clips with camera poses for each frame. To extract training pairs of the semantic layout and the corresponding disparity map, we adopt the idea of cross-modal distillation (Figure 5a). Specifically, we apply PSPNet [52] (pretrained on the ADE20K [53]) to obtain segmentation map annotation. Similar, we apply the pre-trained MiDaS [38] monocular depth estimation network to estimate the corresponding disparity map. Since MiDaS predicts the *relative* disparity with unknown scale/shift, we use the absolute depth prediction from DPSNet [21] to estimate the scale and shift for each training image. The

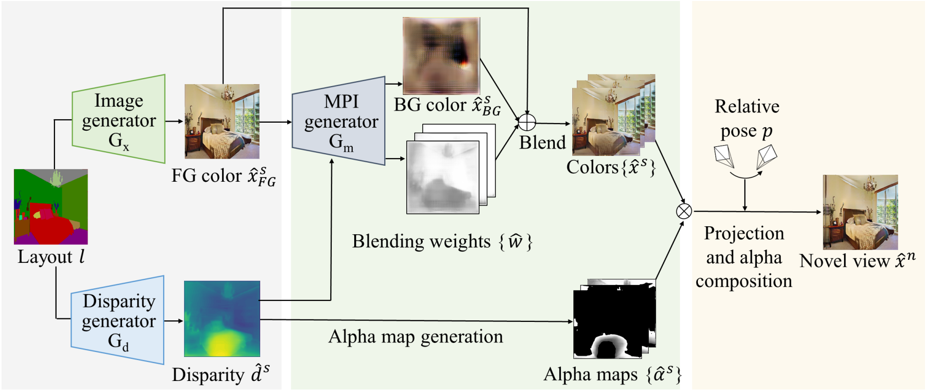


Image generator G_x
FG color $\hat{x}^s_{FG}$
Layout l
Disparity generator G_d
Disparity $\hat{d}^s$
MPI generator G_m
BG color $\hat{x}^s_{BG}$
Blending weights $\{\hat{w}\}$
Blend
Colors $\{\hat{x}^s\}$
Alpha map generation
Alpha maps $\{\hat{\alpha}^s\}$
Relative pose p
Projection and alpha composition
Novel view $\hat{x}^n$

(a) Image/disparity generation (Sec. 3.3)

(b) MPI prediction (Sec. 3.3)

(c) View synthesis (Sec. 3.4)

Fig. 3: **Method overview.** Our method first produces an MPI-based scene representation via a two-step approach (a)(b). (a) Our first step focuses on synthesizing the color and disparity image from the given semantic label map as the *visible surface*. Here, we present a Y-shaped network with partially shared color/depth decoder architecture to ensure consistency between the synthesized color and depth maps. (b) We then infer the MPI representation that captures the color and structure of both the visible surface and the dis-occluded surfaces. With only one single RGB image as input, it is challenging to learn MPI with high-quality view renderings. This is because the network needs to predict both the appearances at multiple depth levels as well as the alpha (transparency) maps. To address this issue, we directly generate the alpha maps using the synthesized depth map from step (a), and we use the synthesized depth map for modulating the activations in normalization layers [35] in our MPI generator. Such an approach imposes effective constraints and results in improved MPI prediction. (c) Given target camera poses, we can then project and blend the generated MPI representation for rendering images at novel views.

relative disparity images are then transformed into absolute disparity images that serve as the (pseudo) ground-truth images for training. We collect training pairs from each frame in the RealEstate10K dataset. While existing Habitat [39] framework also provides semantic layouts, disparity maps and multi-view images with camera poses, we did not use it as the dataset contains indoor scenes only.

3.3 Scene Representation Generation

We adopt a two-step prediction strategy due to the difficulty of predicting MPI representation in one step. First, our image and disparity generator takes the semantic layout l as input and learns to synthesize the corresponding color image $\hat{x}^s_{FG}$ and disparity image $\hat{d}^s$ of the visible surface. Second, the MPI generator uses the synthesized color image and the disparity as input and predicts an MPI representation $\hat{m}^s$ of the scene.

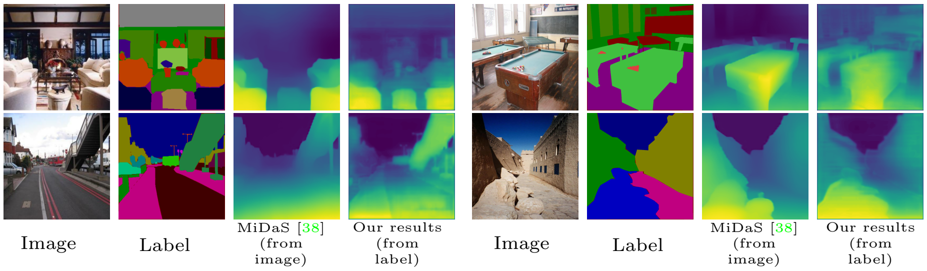
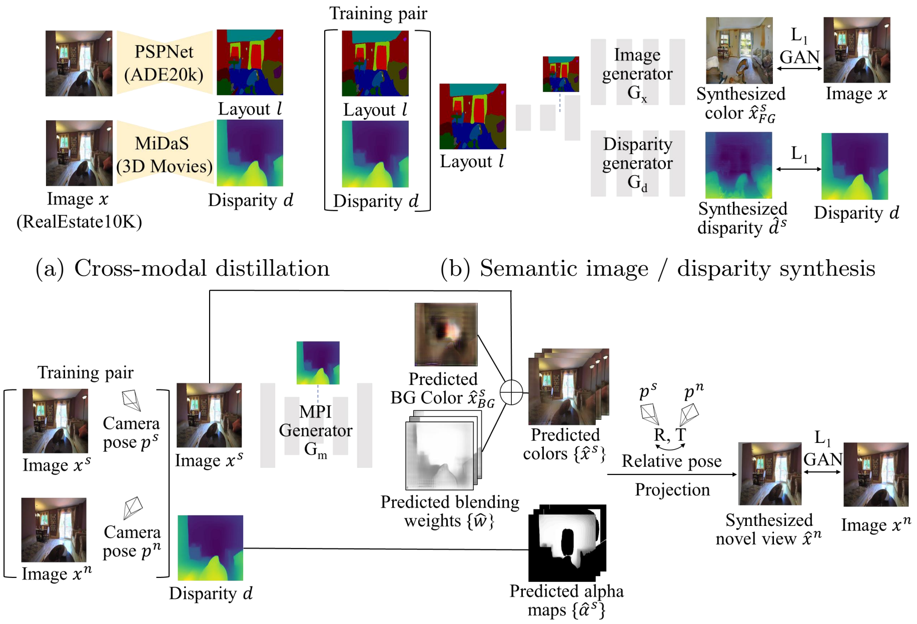


Fig. 4: **Sample results of depth synthesis.** Comparing the prediction from MiDaS [38] (computed from color images), our model produces plausible depth images based on semantic label maps.

Image and disparity generation. Image and disparity generator aims to synthesize the color $\hat{x}^s_{FG}$ and disparity image $\hat{d}^s$ of *visible surface* of the scene (Figure 5b). To this end, we modify the SPADE [35] model into two-stream generators (with the color generator G_x and the disparity generator G_d). The two-stream generators G_x and G_d share the first three SPADE-style ResNet blocks. Using the training pairs of semantic layout l and disparity image d, we use the losses in SPADE [35] for training the color stream and an $\ell 1$ reconstruction loss for training the disparity stream. Figure 4 shows sample results of disparity prediction from a semantic label map.

MPI prediction. For simple scenes (e.g., there is no apparent occluded region in the input image), using a single image with the associated disparity map will suffice for modeling the 3D scene. However, synthesizing novel-view images with only color and disparity map inevitably induce visible artifacts, particularly in the dis-occluded regions, thereby failing to render general scenes where multiple depth layers exist. We therefore use an MPI representation [55] for handling the depth-complex scenarios. An MPI [55] $m = \{(x_k, \alpha_k)\}_{k=1}^{K}$ is a collection of RGBA images, where K is the number of depth planes. Each layer k is an image plane placed at a fixed depth with respect to a virtual reference camera. The color images x_k at each depth plane indicate the visible view, while the alpha image α_k represents the visibility, which has a range between 0 and 1.

However, we find that predicting the MPI using only a single color image results in poor visual quality. The primary reason is that without depth cues (e.g., stereo pair in [55]), it is challenging to predict accurate alpha (transparency) maps for compositing multi-plane images. To tackle this issue, we directly compute and constrain the alpha images from the synthesized disparity map $\hat{d}^s$. Since the synthesized disparity map $\hat{d}^s$ provides a strong prior for the scene visibility at different depth layers, we transform it into the alpha images $\{\hat{\alpha}^s_k\}$ in our MPI representation (Figure 6). Specifically, we first transform the disparity image into a *one-hot representation* with K disparity channels, according to the inverse depth. Then, we apply a half Gaussian blur along the disparity channel, which produces blurring effect only *behind* the predicted disparity and has a



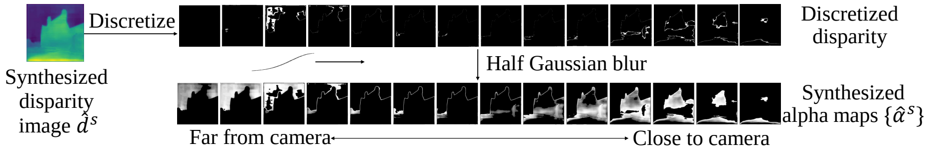
(a) Cross-modal distillation (b) Semantic image / disparity synthesis

(c) MPI prediction

Fig. 5: **Training pipeline.** Our model training process consists of the following steps. (a) **Cross-modal distillation**: We generate *pseudo* training pairs for training the semantic image/depth generation by applying the pre-trained depth estimation model [38] and semantic segmentation PSPNet [52] on training images. (b) **Semantic depth and image synthesis**: Using the generated training pairs from the cross-modal distillation step, we use a two-stream (color and disparity) SPADE network to generate the visible surface. We train the color stream using the losses in [35] and the disparity stream with an $\ell 1$ loss (based on the normalized disparity values). (c) **MPI prediction**: We use training pairs of source/target images with relative pose annotations (provided by [55]). We train the MPI generator to produce colors at multiple depth levels and use $\ell 1$ and GAN loss to enforce the consistency between the projected image and the target image. Note that the MPI generator does *not* need to predict the alpha (transparency) maps.

peak value at the predicted disparity. The blurred one-hot disparity images are then used as the alpha images in our MPI representation.

The alpha images generated by this simple process has three desired properties. First, the pixels at the predicted disparity level are fully visible, resulting in sharp contents at the center view. Second, the blurred alpha images allow the MPI generator to predict the BG colors and blending weights for handling dis-occluded regions at novel views. Third, as the alpha images are generated


Discretize
Discretized disparity
Half Gaussian blur
Synthesized disparity image $\hat{d}^s$
Synthesized alpha maps $\{\hat{\alpha}^s\}$
Far from camera
Close to camera

(a) Discretization (b) Gaussian blur

Fig. 6: **Alpha images.** (a) **Discretization**: we first transform the disparity image to a one-hot encoding image. (b) **Gaussian blur**: we then apply a half Gaussian blur along the disparity channel, and use the result as our alpha images. The alpha images shown here are 14 out of total 128 planes.

in a deterministic manner, the MPI generator can focus only on predicting the color images at multiple planes.

To predict the color images, $\{\hat{x}^s_k\}$ in the MPI representation, we use a SPADE-based [35] MPI generator G_m that takes the color image of the visible surface $\hat{x}^s_{FG}$ as main input, and uses the disparity image $\hat{d}^s$ for modulating the activations in normalization layers. The MPI generator synthesizes a background color image $\hat{x}^s_{BG}$ and a set of blending weights $\{\hat{w}_k\}$. The color images $\{\hat{x}^s_k\}$ are calculated as the weighted sum of the foreground $\hat{x}^s_{FG}$ and the background $\hat{x}^s_{BG}$:

$$\hat{x}^s_k = \hat{w}_k \odot \hat{x}^s_{FG} + (1 - \hat{w}_k) \odot \hat{x}^s_{BG} \tag{1}$$

We refer the reader to Zhou et al. [55] for more details on synthesizing novel view images using an MPI representation.

Training MPI generator. Figure 5c illustrates the training process of MPI prediction. We use the data sampling strategy in [55] to sample the training image pair $(x^s, x^n) = (x^s_{FG}, x^n)$ (note that x^s is equivalent to x^s_{FG}) with corresponding camera poses (p^s, p^n) , as well as the disparity image d^s, where the notation s and n indicate the *source* and *novel* view, respectively. Our MPI generator predicts the color images $\{\hat{x}^s_k\}$ from the source color image x^s_{FG}. We transform the disparity image d^s into alpha images $\{\alpha^s_k\}$.

With the predicted MPI representation $\hat{m}^s = (\{\hat{x}^s_k\}, \{\alpha^s_k\})$, we can use the warped multi-plane images according to the relative pose p^{n-s} between the source pose p^s and novel pose p^n. Given the warped MPIs, we then use the over-composited approach [36] to composite the novel view $\hat{x}^n$. We train the MPI generator using an $\ell 1$ loss and a GAN loss of weight 0.01 between the generated and the ground-truth color image at the novel view x^n.

3.4 Novel view synthesis

Similar to the training process, at test time, we follow the two-step approach for generating an MPI. First, we generate color $\hat{x}^s_{FG}$ and disparity image $\hat{d}^s$ from input semantic layout l. We then use both color $\hat{x}^s_{FG}$ and disparity image $\hat{d}^s$ to predict the MPI representation $\hat{m}^s = (\{\hat{x}^s_k\}, \{\hat{\alpha}^s_k\})$. Given a relative camera pose, we can warp and over-composite the predicted MPI and obtain the novel view image $\hat{x}^n$.

4 Experimental Results

4.1 Experimental setup

Datasets. We validate our method on three datasets.

- **ADE20K** [53] is a dataset of diverse indoor and outdoor scenes. It consists of 2,000 testing images with 150 semantic classes.
- **ADE20K-outdoor** [37] is a subset of outdoor scenes in ADE20K dataset. It consists of 1,035 testing images with 150 semantic classes.
- **NYU** [33] is an indoor dataset. It consists of 249 testing images with 13 semantic classes.

Implementation details. We implement our system in PyTorch and use the Adam optimizer with $\beta_1 = 0$, $\beta_2 = 0.9$ for all network training. All the experiments are conducted on an NVIDIA GTX 1080. The color module, the disparity module and the MPI module are trained for 600k/300k/300k iterations respectively. We use a batch size of one with a learning rate of 0.0002. We use $K = 128$ image planes for our MPI representations. We set the disparity of each alpha map equally distributed from 0.01m to 1m, according to the inverse depth. The Gaussian blur we use for the alpha images has a peak 1, window 31, and the σ value of 10. We set the size of the target synthesized images as 384×384 for all the models. Our source code and the pre-trained models are available on the project website.

Baselines. We compare our methods with four baseline methods.

- (a) **Direct (U-Net)** synthesizes the multi-plane images directly from the semantic layout using a fully-convolutional encoder-decoder architecture [55].
- (b) **Direct (SPADE)** also synthesizes the multi-plane images directly from the semantic layout, but uses a generator with spatially-adaptive normalization [35].
- (c) **Cascade (MPI)** first synthesizes a color image from the semantic layout using SPADE [35], then apply an MPI predictor using the synthesized image as input. Here, we modify the original MPI generation model in [55] so that it takes a single image as input.
- (d) **Cascade (KB)** first synthesizes a color image from the semantic layout using SPADE [35], then apply a recent single-image view synthesis method (3D Ken Burns [34]).

Training and testing details of the baseline models can be found in the supplementary material.

4.2 Quantitative evaluation

We use the Fréchet Inception Distance (FID) [16] to measure the distance between the distribution of generated images and real images. We use ADE20K images as real images. For measuring the realism of novel view synthesis, we

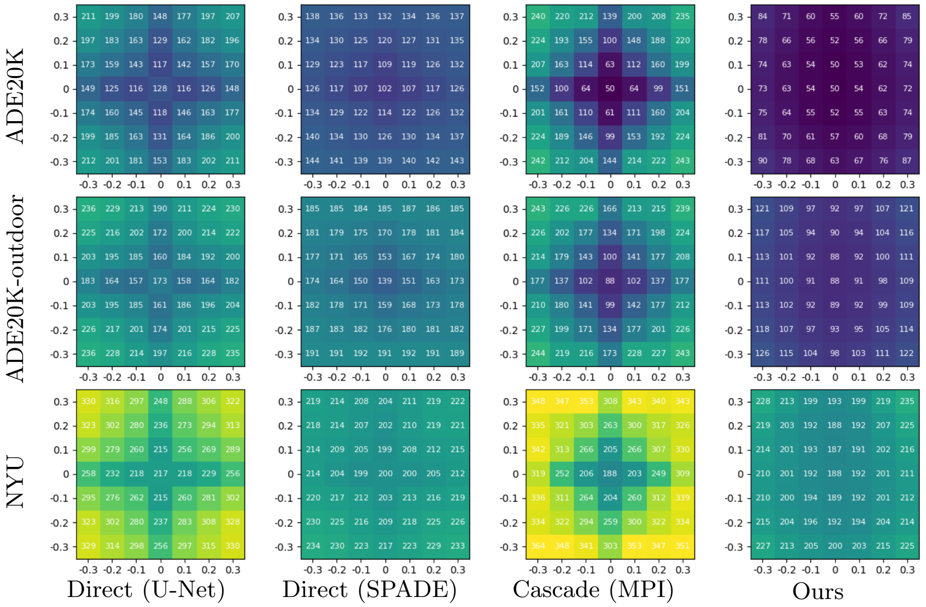


Fig. 7: **Quantitative evaluation.** We compare the results of three alternative approaches for semantic view synthesis and our model on the ADE20K, ADE20K-outdoor, and NYU datasets. Each table shows the FID score of generated novel view images at 7 × 7 grids of target viewpoints. A lower FID score is better. Using Cascade (MPI) and Direct (U-Net) MPI architectures is unable to produce sharp, photorealistic contents (therefore high FID scores). The Direct (SPADE) method can synthesize detailed contents at the center view due to the use of SPADE [35]. However, its performance degrades rapidly when the camera viewpoints move away from the center view. Our two-step generation preserves the detailed content in the front layer while maintaining photorealism under novel views. We were not able to include Cascade (KB) due to different camera movements.

evaluate the FID scores of generating novel views at 7 × 7-grid viewpoints on x-y planes with camera movement from −0.3 meter to 0.3 meter across both axes. The center view with camera movement $(0, 0)$ shows the performance of semantic image synthesis. As shown in Figure 7, all the baselines, and our model produce the lowest FID score at the center view, and the FID score gradually increases when the camera movement becomes larger. The trend is similar across different datasets. We discuss the results based on the ADE20K dataset below.

Results at the center view. Comparing methods directly synthesizing MPIs from layouts, Direct (SPADE) performs better than Direct (U-Net) (102 vs. 128) due to the use of the SPADE architecture. Comparing methods that both employ the SPADE generator, Cascade (MPI) performs better than Direct (SPADE) (50

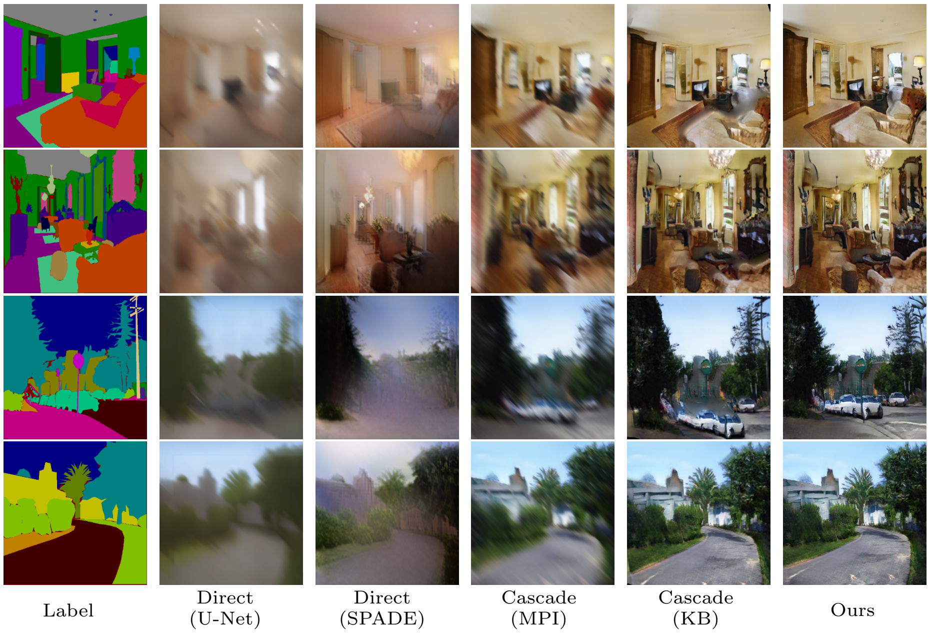


Fig. 8: **Visual comparisons.** We compare the generated novel view images of four other baselines and our model among ADE20K and ADE20K-outdoor datasets. The left column shows the input label at the center viewpoint.

vs. 102), suggesting the difficulty of directly predicting MPI from semantic layout. Our method achieves the same FID score 50 when compared with Cascade (MPI) at the center view as the input (synthesized color image) is the same.

Results at the novel views. When evaluating the results at a novel view (e.g., $(0.3, 0.3)$ meters away from the center), we observe that while the Cascade (MPI) method performs well at the center view, it produces significantly inferior to the methods that directly predict MPI. In contrast, our method produces lowest FID scores among the competing baselines.

4.3 Visual comparisons

Figure 8 compares the generated novel view images of four baselines and our model. Two-step methods, Cascade (MPI), Cascade (3D Ken Burns) and Ours, produce images with sharper contents. Direct (U-Net) and Direct (SPADE) tend to produce blurry and less plausible contents. In particular, the results of Cascade (MPI) suffer from blurry due to the difficulty of generating alpha images when no depth cues (e.g., multiple images, plane sweep volume) are available. The Cascade (KB) inpaints the dis-occluded region at only *one* novel viewpoint. Such a method supports 3D Ken Burns effect with a simple camera trajectory such as zooming in, but not free-viewpoint rendering.

Table 1: **Ablation study.** (a) FID scores under different numbers of depth layers. (b) FID scores of replacing the MPI prediction with per-frame background inpainting. We use NYU dataset for this experiment.

(a) Number of depth layers.

	Camera movement		
	(0, 0)	(0.1, 0.1)	(0.2, 0.2)
128	**188.83**	**191.04**	207.70
64	190.70	193.71	210.06
32	190.60	194.73	**205.59**

(b) Handling dis-occlusion.

	Camera movement		
	(0, 0)	(0.1, 0.1)	(0.2, 0.2)
Ours	**188.83**	**191.04**	**207.70**
Diffusion	190.63	192.55	210.16
GatedConv	190.67	192.83	210.00

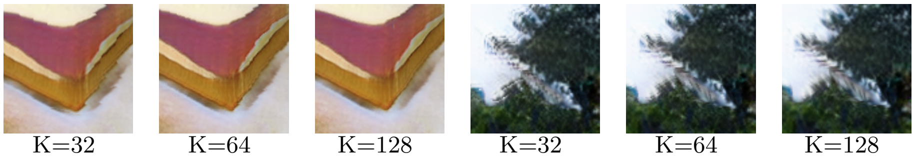

K=32 K=64 K=128 K=32 K=64 K=128

Fig. 9: **Number of depth layers.** Increasing the number of depth levels improves the rendered quality.

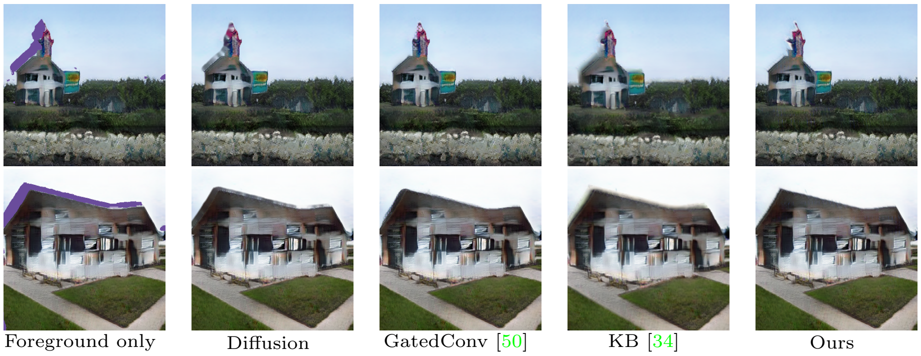

Foreground only Diffusion GatedConv [50] KB [34] Ours

Fig. 10: **Disocclusion handling.** The purple regions (left) are the dis-occluded region. *Diffusion* and *GatedConv* produce artifacts. The 3D Ken Burns method [34] generates blurry and unnatural dis-occluded contents. Our model hallucinates visually appealing results.

4.4 Ablation study

Number of depth layers. Table 1a shows the results of having a different number of depth layers in our MPI. At $(0.2, 0.2)$, the model with $K = 32$ achieves better FID. At $(0, 0)$ and $(0.1, 0.1)$, the model with $K = 128$ achieves better FID. We conclude that more MPI planes lead to slightly blurrier results for large camera movement. Figure 9 illustrates that the novel view synthesized with 32 depth layers show more artifacts than 64 or 128 depth layers.

Background inpainting. We explore alternative methods for handling the dis-occluded regions when rendering at novel views. We use the standard backward warping to project the synthesized color image using disparity image to render the novel views. We then inpaint the missing pixels using either simple diffusion (implemented in OpenCV) or a learning-based image inpainting model (GatedConv [50]).

Table 1b shows that our method achieves lower FID scores at three viewpoints. Note that as all the novel view images are processed independently, *Diffusion* and *GatedConv* approaches do not retain the consistency across different viewpoints. We refer the readers to the supplementary materials for video results. Figure 10 shows that while our method produces slightly blurry foreground (due to the over-composition of multi-plane images), our MPI representation hallucinates plausible dis-occluded regions.

4.5 User study

We conducted a perceptual user study to quantify the user preference over the proposed method and the six baseline approaches. For each test during the study, we present two novel view videos of the same scene generated by two different methods with circular camera motion (in randomized order). We then ask the participant to select his/her preferred result. There are 120 videos (60 pairwise comparisons) generated from the layouts in ADE20K, ADE20K-outdoor, and NYU datasets used. We conduct the study with 47 participants (2820 binary votes). The results shown in Figure 11 validate that the proposed method synthesizes more realistic novel view videos compared to the baseline approaches.

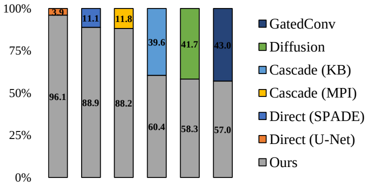


Fig. 11: **User study.** We show the user preference between the proposed method and baselines.

5 Conclusions

We have introduced a new problem called semantic view synthesis. The problem aims to generate a photorealistic image from a given semantic label map that supports novel view rendering. The new form of visual content creation offers significantly more immersive experience than the conventional 2D image synthesis task. This is technically achieved by carefully integrating techniques from semantic image synthesis and view synthesis. Our core idea is to model the 3D scene by first modeling the visible surface then further inferring the full 3D scene representation. We conduct an extensive experimental evaluation to validate our model design and show favorable results over several baseline methods.

References

1. AlBahar, B., Huang, J.B.: Guided image-to-image translation with bi-directional feature transformation. In: ICCV (2019) 2, 4
2. Bau, D., Strobelt, H., Peebles, W., Wulff, J., Zhou, B., Zhu, J.Y., Torralba, A.: Semantic photo manipulation with a generative image prior. ACM Transactions on Graphics (TOG) **38**(4), 1–11 (2019) 2
3. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: ICCV (2017) 2
4. Chen, Y.C., Lin, Y.Y., Yang, M.H., Huang, J.B.: Crdoco: Pixel-level domain transfer with cross-domain consistency. In: CVPR (2019) 4
5. Cheng, Y.C., Lee, H.Y., Sun, M., Yang, M.H.: Controllable image synthesis via SegVAE. In: ECCV (2020) 2
6. Dhamo, H., Tateno, K., Laina, I., Navab, N., Tombari, F.: Peeking behind objects: Layered depth prediction from a single image. In: Pattern Recognition Letters (2018) 4
7. Flynn, J., Broxton, M., Debevec, P., DuVall, M., Fyffe, G., Overbeck, R., Snavely, N., Tucker, R.: DeepView: View synthesis with learned gradient descent. In: CVPR (2015) 2, 4
8. Flynn, J., Neulander, I., Philbin, J., Snavely, N.: DeepStereo: Learning to predict new views from the world's imagery. In: CVPR (2016) 2, 4
9. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: CVPR (2018) 3
10. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017) 4
11. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth prediction. In: ICCV (2019) 4
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014) 2
13. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: CVPR (2016) 5
14. Hedman, P., Kopf, J.: Instant 3D photography. In: SIGGRAPH (2018) 4
15. Hedman, P., Philip, J., Price, T., Frahm, J.M., Drettakis, G., Brostow, G.: Deep blending for free-viewpoint image-based rendering. ACM Transactions on Graphics (TOG) **37**(6), 1–15 (2018) 2
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: NIPS (2017) 10
17. Hoffman, J., Gupta, S., Darrell, T.: Learning with side information through modality hallucination. In: CVPR (2016) 5
18. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: ICML (2018) 4
19. Hu, J., Ozay, M., Zhang, Y., Okatani, T.: Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In: WACV (2019) 3
20. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018) 2, 4
21. Im, S., Jeon, H.G., Lin, S., Kweon, I.S.: DPSNet: End-to-end deep plane sweep stereo. In: ICLR (2019) 5
22. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017) 2, 4, 5

23. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. In: SIGGRAPH Asia (2016) 4
24. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: CoRR. vol. abs/1912.04958 (2019) 2
25. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 3DV (2016) 3
26. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M.K., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: ECCV (2018) 2
27. Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M., Yang, M.H.: Drit++: Diverse image-to-image translation via disentangled representations. IJCV pp. 1–16 (2020) 4
28. Lee, H.Y., Yang, X., Liu, M.Y., Wang, T.C., Lu, Y.D., Yang, M.H., Kautz, J.: Dancing to music. In: NeurIPS (2019) 2
29. Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., Freeman, W.T.: Learning the depths of moving people by watching frozen people. In: CVPR (2019) 4
30. Li, Z., Snavely, N.: MegaDepth: Learning single-view depth prediction from internet photos. In: CVPR (2018) 4
31. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015) 4
32. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. In: SIGGRAPH (2019) 4
33. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: ECCV (2012) 2, 10
34. Niklaus, S., Mai, L., Yang, J., Liu, F.: 3D Ken Burns effect from a single image. In: ACM Transactions on Graphics (2019) 10, 13
35. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019) 2, 4, 5, 6, 7, 8, 9, 10, 11
36. Porter, T., Duff, T.: Compositing digital images. In: SIGGRAPH (1984) 9
37. Qi, X., Chen, Q., Jia, J., Koltun, V.: Semi-parametric image synthesis. In: CVPR (2018) 2, 10
38. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. arXiv:1907.01341 (2019) 4, 5, 7, 8
39. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: ICCV (2019) 6
40. Shih, M.L., Su, S.Y., Kopf, J., Huang, J.B.: 3d photography using context-aware layered depth inpainting. In: CVPR (2020) 4
41. Srinivasan, P.P., Tucker, R., Barron, J.T., Ramamoorthi, R., Ng, R., Snavely, N.: Pushing the boundaries of view extrapolation with multiplane images. In: CVPR (2019) 2, 4
42. Tseng, H.Y., Fisher, M., Lu, J., Li, Y., Kim, V., Yang, M.H.: Modeling artistic workflows for image generation and editing. In: ECCV (2020) 2
43. Tseng, H.Y., Lee, H.Y., Jiang, L., Yang, W., Yang, M.H.: RetrieveGAN: Image synthesis via differentiable patch retrieval. In: ECCV (2020) 2
44. Tucker, R., Snavely, N.: Single-view view synthesis with multiplane images. In: CVPR (2020) 2, 4
45. Wang, C., Lucey, S., Perazzi, F., Wang, O.: Web stereo video supervision for depth prediction from dynamic scenes. In: 3DV (2019) 4

46. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: CVPR (2018) 2, 4, 5
47. Wiles, O., Gkioxari, G., Szeliski, R., Johnson, J.: SynSin: End-to-end view synthesis from a single image. In: CVPR (2020) 4
48. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. In: CVPR (2017) 3
49. Yin, Z., Shi, J.: GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In: CVPR (2018) 4
50. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: ICCV (2019) 13, 14
51. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017) 2
52. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017) 5, 8
53. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: CVPR (2017) 2, 5, 10
54. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017) 4
55. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. In: SIGGRAPH (2018) 2, 4, 5, 7, 8, 9, 10
56. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017) 2, 4
57. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: NIPS (2017) 4
58. Zou, Y., Ji, P., Tran, Q.H., Huang, J.B., Chandraker, M.: Learning monocular visual odometry via self-supervised long-term modeling. In: ECCV (2020) 4
59. Zou, Y., Luo, Z., Huang, J.B.: DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency. In: ECCV (2018) 4