DPNet: Depth and Pose Net for Novel View Synthesis via Depth Map Estimation

Ge Zhu, Yu Liu(✉), and Yumei Wang

Institute of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
{zhuge2020110241,liuy,ymwang}@bupt.edu.cn

Abstract. Novel view synthesis is regarded as one of the efficient ways to realize stereoscopic vision, which paves the way to virtual reality. Image-based rendering (IBR) is one of the view synthesis strategies, which warps pixels from source views to target views in order to protect low-level details. However, IBR methods predict the pixels correspondence in an unsupervised way and have limits in getting accurate pixels. In this paper, we propose Depth and Pose Net (DPNet) for novel view synthesis via depth map estimation. We introduce two nearby views as implicit supervision to improve the pixels correspondence accuracy. Besides, the depth net firstly predicts the source depth map and then the pose net transforms the source depth map to the target depth map which is used to calculate pixels correspondence. Experimental results show that DPNet generates accurate depth maps and thus synthesizes novel views with higher quality than state-of-the-art methods on the synthetic object and real scene datasets.

Keywords: view synthesis · IBR · depth map · pixels correspondence · DPNet

1 Introduction

Novel view synthesis (NVS) solves the problem of generating new view images of a scene or an object in the condition that one or more input views are given, which can be used to generate all possible viewpoints of real-world scenes in virtual reality (VR). As shown in Fig. 1, given one image of the chair, we generate a new image of the same chair from a novel viewpoint (GT represents the real target view image). However, traditional NVS methods often rely on parallax plots or depth maps, which have the limitations of high computational costs. With the development of neural networks, supervision-based learning methods can synthesize high-quality novel views. NVS can be used in different application areas, including image editing, and animating still photographs.

F. Gao et al. (Eds.): ChinaCom 2022, LNICST 500, pp. 294–308, 2023.
https://doi.org/10.1007/978-3-031-34790-0_23

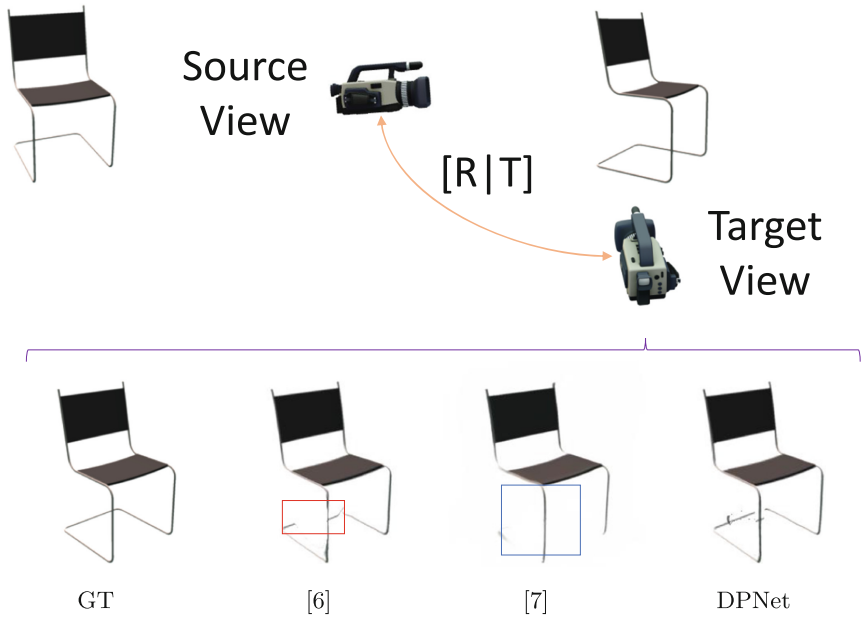


Fig. 1. With an input image, a novel view of the same object or scene is synthesized. It shows the results of DPNet compared to other two methods and all of them synthesis the target view by predicting target depth map. □ and □ prove that DPNet produces more clear result than other two depth guided view synthesis methods (Color figure online)

The NVS methods that have been approached in the last few years fall into two main types: 3D geometry-based methods and IBR methods. For 3D geometry-based methods, comprehensive 3D understanding is important so that the first step is to get the approximate underlying 3D structure. Some methods estimate the underlying 3D geometry in form of voxels [1] and mesh [2], and then put the corresponding camera transformations to the pixels of the 3D structure to produce the final output [3]. However, they not only require a commitment of time and resources, but also produce holes where lack of a prior information. In such conditions, hole-filling algorithms are needed but sometimes these algorithms are not effective [4].

Unlike 3D geometry-based methods, IBR methods generate novel images based on input images. The pixels from source views can be reprojected to the target view, low-level details such as colors and textures are well-protected. Zhou *et al.* [5] directly estimates the appearance flow and get final pixels value of target views from input views, and Chen *et al.* [6] predicts the target depth map to obtain the pixel-to-pixel correspondences with 3D warping. Hou *et al.* [7] also predicts the depth map of target view, but warps feature maps to generate the final target view image rather than directly warping pixels from source view. These IBR methods all achieve great view synthesis quality.

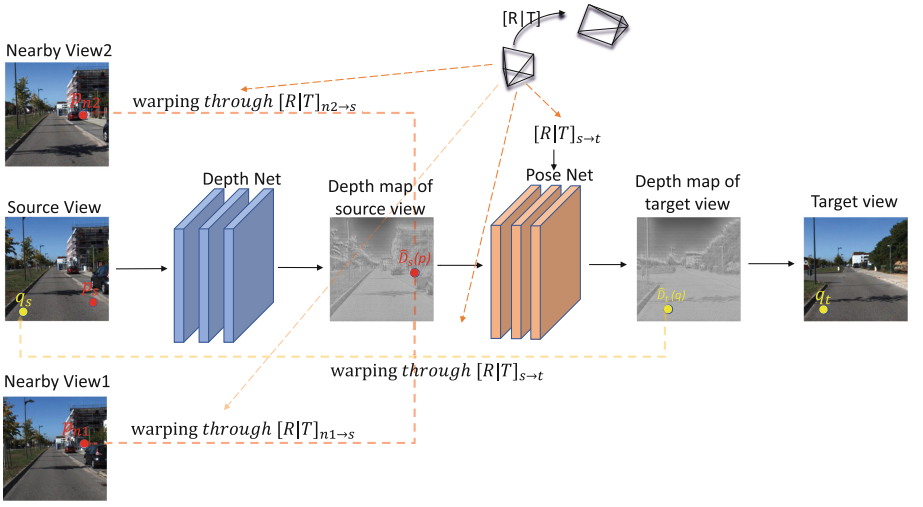


Fig. 2. Overview of the view synthesis pipeline. There are two main components: the depth net and pose net. The depth net takes only the source view as input and generates the depth map $\hat{D}_s$. Moreover, $\hat{D}_s$ and two nearby views are used to reconstruct source view that the L1 loss between generated source view and real source view can be alleviated to train the depth net. Then the pose net extracts the feature points of $\hat{D}_s$ to produce the depth map of target view $\hat{D}_t$ Finally, the $\hat{D}_t$ is used to warp pixels of source to the target view with bilinear interpolation.

In this paper, we design a reasonable frameworks to improve the view synthesis quality. Motivated by the advantage of IBR method, we take the method that it warps pixels from source view to target view with the help of target depth map. More specifically, we propose the DPNet consisting of a depth net and a pose net as shown in Fig. 2. DPNet firstly predicts the source depth map and subsequently deduces the target depth map rather than directly producing target depth map. In such way, the short-connection structure can be introduced into pose net. So multi-level feature maps extracted from source depth map can be transferred to target depth map to improve the depth map accuracy. To further improve the accuracy of predicted depth map of source view, two nearby view images are reprojected to source view through predicted depth map of source view. Then the camera transformation and the predicted source depth map are put into pose net to generate the target depth map. Subsequently, the generated target depth map is used to calculate the dense correspondences between the source view and the target view via perspective projection. Finally, the final output image is synthesized via pixel warping.

To get clear and continuous synthesis results, four specially designed loss functions are used to train the DPNet. The supervision loss is used to improve the depth map estimation accuracy. And the L1 reconstruction loss and VGG perceptual loss are used to generate realistic images. Moreover, the edge smoothness loss can make the final target depth map more continuous in edge. Detailed experiments are conducted on real scene [8] and synthetic object [9] datasets, the depth estimation accuracy and image quality are evaluated qualitatively and quantitatively. The experiment results demonstrate that DPNet actually improves the depth estimation accuracy and image quality.

2 Related Work

Study of novel view synthesis has a long history in computer vision and graphics. These researches differ based on whether they use pure images or 3D geometry structure and on whether a single view image or multiple view images are put into neural network. Recently, neural radiance fields and generative models are the new directions.

2.1 Geometric View Synthesis

If multiple images of a scene are provided, with the help of COLMAP [10,11], a 3D geometry scaffold can be constructed. Riegler *et al.* [3] firstly ran structure-from-motion [10] to get camera intrinsic and camera poses, then ran multi-view stereo [11] on the posed images to obtain per-image depth maps, and finally fused these maps into a point cloud. Similarly, Penner *et al.* [12] warped the extracted source feature maps into the target view using the depth map which was derived from the 3D geometry scaffold. A confidence image and a color image for each input image are obtained through these warped feature maps. Then these confidence images and color images were aggregated to get a final output. More recently, deep learning techniques created a new level of possibility and flexibility. Lombardi *et al.* [13] learned an implicit voxel representation of an object given many training views and generated a new view of that object when tested.

2.2 3D from Single Image

Inference about 3D shapes can serve as an implicit step in view synthesis. Given the serious inadequacy of recovering 3D shapes from a single image, recent work deployed neural networks for this task. They could be categorized by their output representation into mesh, point cloud, and voxel. With a single image as input, Tatarchenko *et al.* [14] predicted many unseen views and their depth maps from input, and these views were fused into a 3D point cloud which was later optimized to obtain a mesh. In [4], the features extracted from single input and the depth map estimated from the same input were used to create a point cloud carrying

features. Many works explore using a DNN to predict 3D object shapes [15] or the depth map of a scene given an image [16]. These works focus on the quality of the 3D predictions as opposed to the view-synthesis task.

2.3 Image-Based Rendering Methods

Recently, many deep neural networks are developed to learn the image-to-image mapping between source view and target view [5–7,17,18]. Zhou *et al.* [5] directly estimated the appearance flow map in order to warp pixels of source view to their position of target view, Sun *et al.* [17] further refined the output by fusing multiple views with confidence map. With the help of predicted target depth map, Chen *et al.* [6] directly warped pixels of source view to target view and Hou *et al.* [7] warped the multi-level feature map extracted from source view to synthesize the final output. To improve the quality of synthesized image, Park *et al.* [18] used two consecutive encoder-decoder networks, firstly predicting a disocclusion aware flow and then refining the transformed image with a completion network. And in this paper, the target depth map couldn't be predicted from inputs directly, instead, the source depth map is firstly estimated by depth net and then the source depth net is transformed to target depth map through pose net.

2.4 Generative Models and Neural Radiance Fields

View synthesis can also be thought as an image generation task, and it has a lot to do with the field of generative modeling of images [19,20]. In [21], explicit pose control was allowed, they also used voxel. Although these methods can be used for view synthesis, the resulting view lacks consistency and has no control over the objects to be synthesized. The neural radiation field [22] produced impressive results by training an multi-layer perception (MLP) to map 3D rays to occupancy and color. Images are synthesized from this representation by volume rendering. This approach has been extended to an unlimited collection of outdoor scenes and crowdsourced images.

3 The Proposed Method for Novel View Synthesis

Figure 2 shows an overview of DPNet, it consists of two subnets: the depth net Ψ_D and the pose net Ψ_P. The depth net estimates the depth map of source view firstly. For the depth net, we use the skip-connection structure with four downsampling and upsampling layers to give a final prediction of the same spatial resolution with the input. This is followed by a sigmoid layer and a renormalization step, so the depth of prediction falls within the minimum and maximum values for each dataset. The predicted depth map is used to warp two nearby view images to source view, and L1 distance between generated source view and real source view is used to train the depth net. As for pose net, the given transformation matrix is applied on the 3D feature points extracted from the

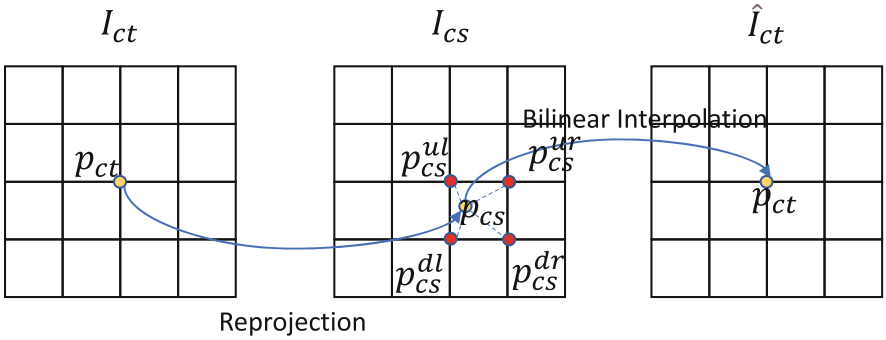


Fig. 3. Illustration of the pixels warping process from source view to target view. For each pixel point p_{ct} in the target view, it is firstly reprojected onto the source view based on the predicted depth map and camera pose transformation, and then the pixels value in target view are obtained by bilinear interpolation.

predicted source depth map to obtain the 3D feature points of the target depth map. Later, when the transformed 3D feature points are given, the depth map of the target view is predicted. Then the estimated depth map is used to find dense correspondences between target and source views. Finally, the source image is warped into the target image via bilinear interpolation.

3.1 Pixels Warping

The reprojection process and the bilinear interpolation process are shown in Fig. 3. For the reprojection process, the per-pixel correspondence C is obtained from the target depth map D_{ct} by converting from a depth map to 3D coordinates [X, Y, Z] and perspective projections:

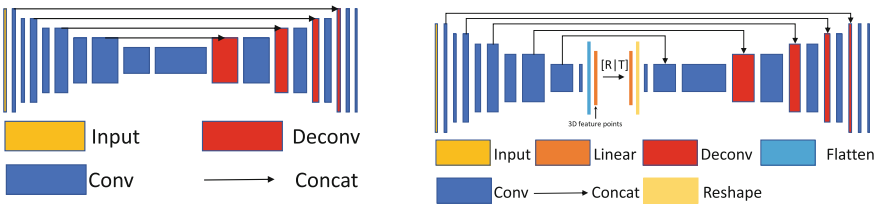
$$[X, Y, Z]^T = D_{ct}(x_{ct}, y_{ct})K^{-1}[x_{ct}, y_{ct}, 1]^T, \tag{1a}$$

$$[x_{cs}, y_{cs}, 1]^T \sim T_{ct\to cs}[X, Y, Z, 1]^T, \tag{1b}$$

where each pixel (x_{ct}, y_{ct}) in the target view corresponds to the pixel position (x_{cs}, y_{cs}) in the source view. Moreover, K is the camera intrinsic matrix and $T_{ct\to cs}$ represents the transformation matrix from target view to source view. For the bilinear interpolation process, with the obtained per-pixel correspondences $C_{ct\to cs}$, the pixels in the correspondences source view can be warped to the correspondences target view:

$$\begin{aligned} I_{ct}(x_{ct}, y_{ct}) = \sum_{x_{cs}} \sum_{y_{cs}} & max(0, 1 - |x_{cs} - C_{ct\to cs}(x_{ct}, y_{ct})|) \\ & max(0, 1 - |y_{cs} - C_{ct\to cs}(x_{ct}, y_{ct})|) I_{cs}(x_{cs}, y_{cs}). \end{aligned} \tag{2}$$

Introducing the intermediate step of predicting depth map enforces the network to adhere to geometric constraints, resolving ambiguous correspondences. This process is substituted by $I_{ct} = PW(I_{cs}, D_{ct}, T_{ct\to cs})$.



(a) depth net (b) pose net

Fig. 4. Network architecture of the depth and pose modules. The width and height of each blue/red rectangular block respectively represent the output channel and spatial dimension of the feature map at the corresponding layer, and each decrease/increase in width and height size represents a change by the factor of 2 (the last conv layer is the output, it does not obey the rules). For depth net, it consists of 4 downsampling lawyers and 4 upsampling lawyers with the skip-connection structure. For pose net, inspired by [6], we also extract the latent code (3D feature points) to inject the camera transformation and predict the target depth map. (Color figure online)

3.2 Depth Map Estimation

The depth net takes a single input image to get the source depth map $D_s = \Psi_D(I_s)$. Moreover, two additional nearby view images I_{n1} and I_{n2} plus their camera transformation $T_{s\to n1}$ and $T_{s\to n2}$ are introduced to warp their pixels to source view to improve the depth estimation accuracy. $\hat{I}_{s1} = PW(I_{n1}, D_s, T_{s\to n1})$ and

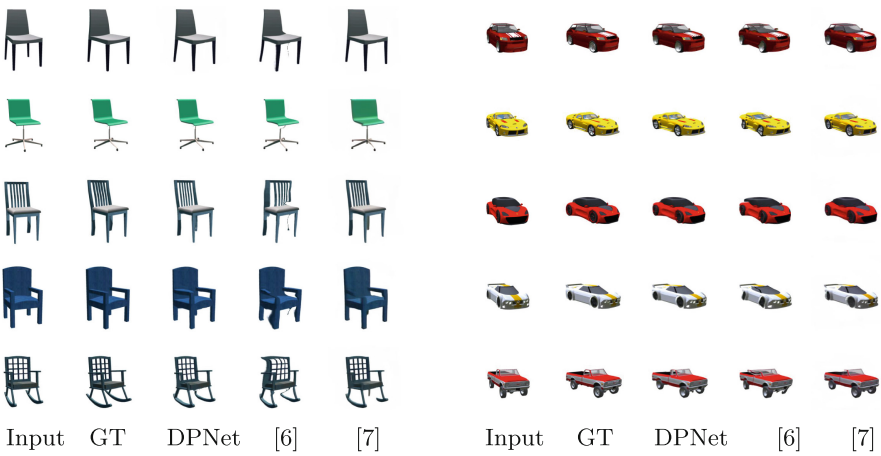


Fig. 5. Results on ShapeNet chair and car datasets. DPNet generates more structure-consistent predictions than [6] (for example, it can't generate a distorted leg in line 5); on the other hand, the generated images of the DPNet are more clear than [7] that it can rebuild rich low-level details (for example, it generates more clear chair surface in line 2).

$\hat{I}_{s2} = PW(I_{n2}, D_s, T_{s\to n2})$. And the two L1 distance between I_s and $\hat{I}_{s1}$ and between I_s and $\hat{I}_{s2}$ are the important part of final training loss function. The specific structure is described in Fig. 4(a), it consists of four downsampling layers and four upsampling layers, and the skip-connection structure transfers multi-level feature maps to create more stable predictions.

3.3 Depth Map Transformation and Target View Generation

In pose net, transformation matrix are applied to latent code to predict depth map of the target view, and the pose network is used to learn compact latent representations that are transformation equivariant. Given the source depth map, the 3D feature points z_s extracted from predicted source depth map can be regarded as a set of points $z_s \in R^{n\times 3}$. Then the 3D feature points are multiplied with the given transformation $T_{s\to t} = [R|t]_{s\to t}$ to get the transformed 3D feature points for the target view:

$$\widetilde{z}_s = T_{s\to t} \cdot \dot{z}_s \tag{3}$$

where $\dot{z}_s$ is the homogeneous representation of z_s. Then the target depth map D_t is created through $\widetilde{z}_s$. With the generated target depth map D_t and corresponding camera transformation $T_{s\to t}$, the target view image is synthesized $\hat{I}_t = PW(I_s, D_t, T_{t\to s})$. Because the input to pose net is a source depth map and not a source view image, the skip-connection structure can be introduced to transfer multi-level feature maps to make the pose net more effective (as shown in Fig. 4(b)).

3.4 Training Loss Functions

The framework can be trained in an end-to-end manner. For each input sample, a single source image, two nearby view images, one target view image and their relative transformation are provided. The depth net and the pose net are optimized jointly. To train the depth net in an implicit supervised manner, the supervision loss is used to improve the depth map estimation accuracy. For pixels regression, the L1 reconstruction loss and VGG perceptual loss are used to

Table 1. Results on ShapeNet objects. DPNet performs better than [6,7] for both chair and car objects, showing that it can deal with complex shape of chairs and rich colors and textures of cars (↓ suggests the smaller the better, ↑ suggests the larger the better).

METHODS	CHAIR		CAR	
	L1↓	SSIM↑	L1↓	SSIM↑
Chen *et al.* [6]	0.0559	0.9224	0.0338	0.9424
Hou *et al.* [7]	0.0583	0.9237	0.0346	0.9392
DPNet	**0.0413**	**0.9381**	**0.0295**	**0.9491**

generate realistic images. Moreover, the edge smoothness loss can make the final target depth map more continuous in edge.

L1 Reconstruction Loss. The L1 reconstruction loss is the L1 loss between the predicted target view $\hat{I}_t$ and the ground truth I_t. Described as:

$$L_{recon} = \|\hat{I}_t - I_t\| \tag{4}$$

To minimize this reconstruction loss, the network learns to produce realistic new views by predicting the necessary depth maps.

Supervision Loss. The supervision loss consists of two parts, both of them are the L1 distance between the ground truth source view I_s and the generated source view $\hat{I}_s$:

$$L_{sup} = \|\hat{I}_{s1} - I_s\| + \|\hat{I}_{s2} - I_s\| \tag{5}$$

To minimize this supervision loss, the depth net learns to produce more accurate source depth map.

VGG Perceptual Loss. In addition to the L1 reconstruction loss, we also employ VGG perceptual loss to obtain realistic synthesis results. The pre-trained VGG16 network is used to extract features from the generated fake results and ground-truth images, and the perceptual loss is the sum of feature distances (L1 distance) calculated from multiple layers.

Edge Smoothness Loss. The edge smoothness loss encourages local smoothing of the predicted depth map. The loss is weighted because depth discontinuities usually occur at the edges of the image:

$$L_{edge} = \frac{1}{N}\sum_{i,j}|\partial x \widetilde{D}_t^{ij}|e^{-\|\partial x I_t^{ij}\|} + |\partial y \widetilde{D}_t^{ij}|e^{-\|\partial y I_t^{ij}\|} \tag{6}$$

where $\widetilde{D}_t$ is the predicted depth map of the target view and I_t is the ground-truth target view.

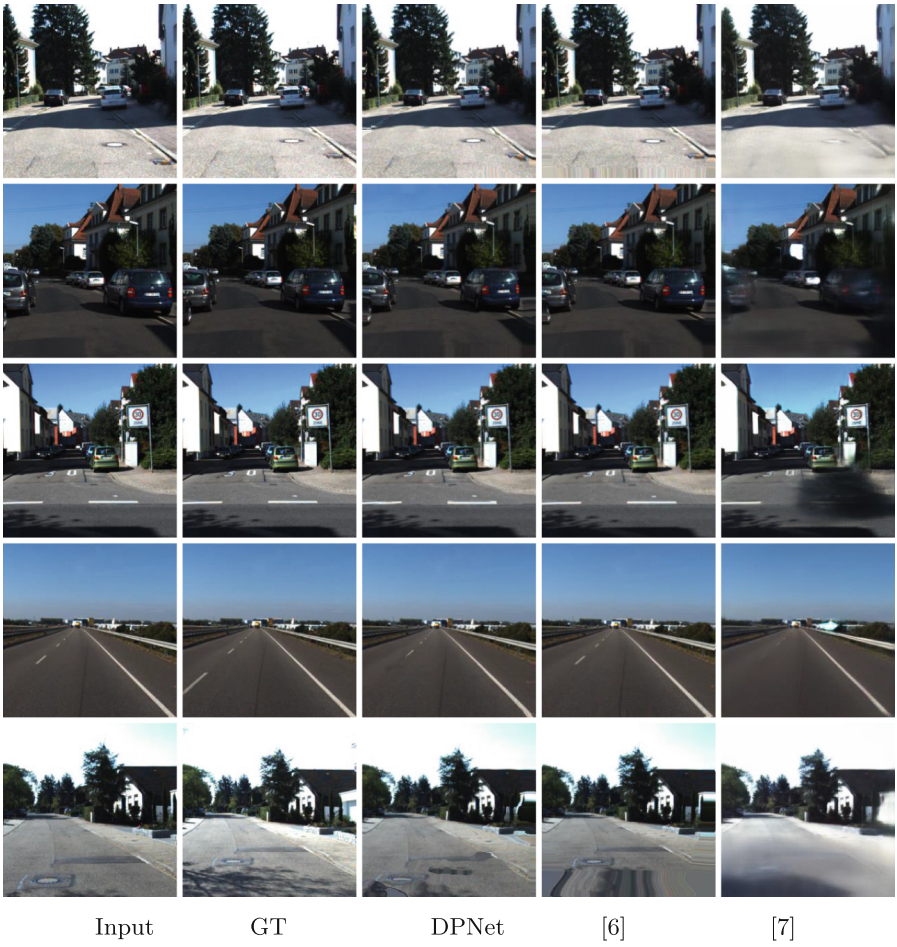


Fig. 6. Qualitative results of KITTI. DPNet produces clear and structurally consistent predictions, while the depth guided pixels warping [6] method produces distortion and the depth guided multi-level feature map warping method [7] produces blurry .

In summary, the final loss function of the joint training framework will be:

$$L = \lambda_r L_{recon} + \lambda_s L_{sup} + \lambda_v L_{vgg} + \lambda_e L_{edge} \tag{7}$$

where the λ_r, λ_s, λ_v, and λ_e are weights for different loss functions.

Table 2. Results on KITTI. DPNet achieves the best SSIM results, with L1 performance outperforming both Chen *et al.* [6] and Hou *et al.* [7]. (↓ suggests the smaller the better, ↑ suggests the larger the better).

METHODS	KITTI	
	L1↓	SSIM↑
Chen *et al.* [6]	0.1803	0.6751
Hou *et al.* [7]	0.1635	0.7253
DPNet	**0.1634**	**0.7273**

4 Experiment Results and Analysis

In this section, experiments are conduct on public datasets, ShapeNet dataset [8] and KITTI dataset [9]. DPNet is compared with state-of-the-art algorithms to evaluate the performance qualitatively and quantitatively. Further ablation studies verify the effectiveness of the different modules of DPNet.

4.1 Dataset and Experiment Setup

For datasets, two different types of datasets are used for experiment: ShapeNet dataset [8] is used for synthetic objects and KITTI dataset [9] is used for real-world scene. More specifically, cars and chairs in the ShapeNet dataset are selected. 3D understanding of datasets with complex structures and camera transformations are a great challenge (*e.g.* depth estimation) and datasets with rich textures will show whether these methods preserve fine-grained detail well. In these selected datasets, the chairs have more complex shapes and structures, but there will be more colorful patterns for the cars. For KITTI, the scene contains more objects, and translation is the primary transformation between frames, unlike ShapeNet, where rotation is the key transformation. In this case, there is less need for accurate depth estimation, and the ability to recover low-level detail is more important for performance.

ShapeNet. Rendered images are used with the dimension of 256×256 from 54 viewpoints (the azimuth from 0° to 360° with 20° increments, and the elevation of 0°, 10°, and 20°) for each object. The training and test pairs are two views with the azimuth difference within the range $[-40°, 40°]$. For ShapeNet chairs, there are 558 chair objects in the training set and 140 chair objects in the test set; For ShapeNet cars, there are 5,997 car objects in the training set and 1,500 car objects in the test set.

KITTI. There are 11 sequences and each sequence contains around 2,000 frames on average. The training pairs are restricted to be separated by at most 7 frames.

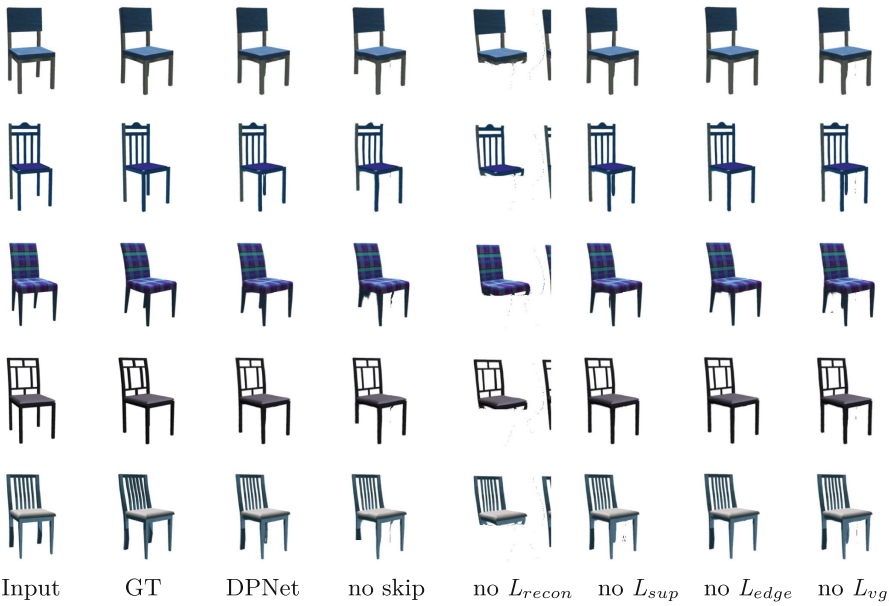


Fig. 7. Ablation studies results. We compare the performance of the full model with its variants. The results show that the lack of L_{recon} leads to incomplete objects (like the chair in the 5th column). The lack of skip-connection structure in depth net and pose net results in the chair leg shortage (like the chair in the 5th column). The L_{vgg} makes the results sharper. And the L_{sup} and the L_{edge} leads to more accurate depth map estimation that it makes the results more stable.

For experiment setup, the depth net and the pose net are jointly trained using the Adam solver same with [7] that $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and learning rate of 6×10^{-5}.

4.2 Evaluation Metrics and Evaluation Results

For evaluation metrics, Mean Absolute Error (L_1 error) and Structural SIMilarity (SSIM) Index are used as metrics to evaluate the synthetic results. For L1 metric, smaller is better; for the SSIM metric, larger is better. For image synthesis quality, DPNet is compared with two state-of-the-art depth map guided methods: one pixels warping method proposed by [6] and one multi-level feature map warping method proposed by [7]. For depth map estimation, DPNet is compared with one source depth map estimation method [4] and one target depth map estimation method [6]. Table 1 shows the results on test set of ShapeNet objects. DPNet performs best for both the chair and the car objects, showing that it can handle both the complex 3D structure of the chair and the rich texture of the car. Fig. 5 shows the qualitative results for all methods. The depth map guided pixels warping method [6] suffers from distortion and the depth map guided multi-level feature maps warping method [7] leads to blurry results. Two

Table 3. Results of ablation studies. All designed modules and loss functions help improve performance. (↓ suggests the smaller the better, ↑ suggests the larger the better).

METHODS	L1↓	SSIM↑
DPNet	**0.0414**	**0.9381**
no skip	0.0609	0.9199
no L_{recon}	0.1261	0.8786
no L_{sup}	0.0538	0.9295
no L_{edge}	0.0437	0.9360
no L_{vgg}	0.0423	0.9361

Table 4. Depth estimation results on ShapeNet chairs.

METHODS	L1-ALL	L1-REL	L1-INV	SC-INV
DPNet-source	**0.0576**	**0.0286**	**0.0145**	**0.0501**
Wiles *et al.* [4]	0.0699	0.0354	0.0184	0.0583
DPNet-target	**0.0598**	**0.0294**	**0.155**	**0.0516**
Hou *et al.* [7]	0.0610	0.0305	0.161	0.0523

nearby views are introduced to improve the accuracy of depth map estimation so that more impressive results are generated (*e.g.*, more complete chair leg and more detailed car roof are generated in line 5). All the methods are also evaluated on KITTI. Table 2 shows the quantitative results. DPNet performs better than [6] and obtains comparable results to [7]. Figure 6 shows the qualitative results, it can be seen that DPNet produces more clear predictions and better preserves the structure (check the bottom part of row 1, manhole cover in row 5). In a conclusion, DPNet can achieve high image synthesis quality.

4.3 Ablation Studies and Depth Estimation Results

To understand how the different modules of the framework work, we conduct an ablation study on ShapeNet chair as it is the most challenging dataset for 3D structures. Figure 7 and Table 3 show the performance of the different variants. No skip stands for removing the skip-connection structure in depth net and pose net. No L_{recon}, no L_{sup}, no L_{edge}, no L_{vgg} separately represents removing corresponding loss function from total loss function. The results show that the lack of L_{recon} leads to incomplete objects (like the chair in the 5th column). The lack of skip-connection structure in depth net and pose net leads to the chair leg shortage (like the chair in the 5th column). The L_{vgg} makes the results sharper. And the L_{sup} and the L_{edge} leads to more accurate depth map estimation that it makes the results more stable.

Moreover, to prove that more accurate depth maps are predicted, four metrics are used to evaluate depth map quality [7]. L1-all compute the mean absolute difference. L1-rel compute the mean absolute relative difference L1-rel $= \frac{1}{n}\sum_i |gt_i - pred_i|/gt_i$, and L1-inv metric is mean absolute difference in inverse depth L1-inv $= \frac{1}{n}\sum_i |gt_i^{-1} - pred_i^{-1}|$. Except L1 metrics, we also utilize sc-inv $= \left(\frac{1}{n}\sum z_i^2 - \frac{1}{n^2}\left(\sum z_i\right)^2\right)^{\frac{1}{2}}$, where $z_i = \lg\left(pred_i\right) - \lg\left(gt_i\right)$. The source depth map estimation is compared with [4] and the target depth map estimation is compared with [6]. Table 4 shows that our predicted depth is more accurate, which can explain why the DPNet can achieve better results than other methods.

5 Conclusion and Discussion

In this paper, DPNet is put forth to solve the novel view synthesis task. And it consists of two subnets: depth net and pose net. The depth net predicts the depth map of the source view from a single input view and two nearby view images are introduced to improve the accuracy of predicted depth map. Then the pose net is used for transformation between source depth map and target depth map. Moreover, the warping from source view pixels to target view pixels enables the preservation of low-level details, so more clear predictions are produced. Experimental results show that compared with above depth map guided warping methods, the performance of DPNet is better.

References

1. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: A unified approach for single and multi-view 3d object reconstruction. In: European Conference on Computer Vision, pp. 628–644(2016)
2. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3907–3916(2018)
3. Riegler, G., Koltun, V.: Stable view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12216–12225 (2021)
4. Wiles, O., Gkioxari, G., Szeliski, R., Johnson, J.: Synsin: End-to-end view synthesis from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7467–7477 (2020)
5. Zhou, T., Tulsiani, S., Sun, W., et al.: View synthesis by appearance flow. In: European Conference on Computer Vision, pp. 286–301 (2016)
6. Chen, X., Song, J., Hilliges, O.: Monocular neural image based rendering with continuous view control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4090–4100 (2019)
7. Hou, Y., Solin, A., Kannala, J.: Novel view synthesis via depth-guided skip connections. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3119–3128 (2021)
8. Chang, A., X., Funkhouser, T., Guibas, L., et al. Shapenet: An information-rich 3d model repository. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1512–3012 (2015)

9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361 (2012)
10. Schonberger, J.L., Frahm, M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4104–4113 (2016)
11. Schönberger, J.L., Zheng, E., Frahm, J.M., et al.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision, pp. 501–518 (2016)
12. Penner, E., Zhang, L.: Soft 3D reconstruction for view synthesis. In: ACM Transactions on Graphics, pp. 1–11 (2017)
13. Lombardi, S., Simon, T., Saragih, J., et al.: Neural volumes: Learning dynamic renderable volumes from images. arXiv preprint 2019)
14. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Multi-view 3d models from single images with a convolutional network. In: European Conference on Computer Vision, pp. 322–337 (2016)
15. Insafutdinov, E., Dosovitskiy, A.: Unsupervised learning of shape and pose with differentiable point clouds. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 2807–2817 (2018)
16. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2041–2050 (2018)
17. Sun, S.H., Huh, M., Liao, Y.H., et al.:Multi-view to novel view: Synthesizing novel views with self-learned confidence. In: Proceedings of the European Conference on Computer Vision, pp. 155–171(2018)
18. Park, E., Yang, J., Yumer, E., et al.:Transformation-grounded image generation network for novel 3d view synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3500–3509 (2017)
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets. Commun. ACM **63**(11), 139–144 (2020)
20. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. arXiv preprint (2018)
21. Nguyen-Phuoc, T., Li, C., Theis, L., et al.:Hologan: Unsupervised learning of 3d representations from natural images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7588–7597 (2019)
22. Niemeyer, M., Mescheder, L., Oechsle, M., et al.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3504–3515 (2020)