Computational Visual Media
https://doi.org/10.1007/s41095-019-0159-7 Vol. 5, No. 4, December 2019, 337–345
Research Article

# Reconstructing piecewise planar scenes with multi-view regularization

**Weijie Xi**[1], **Xuejin Chen**[1] (✉)



**Abstract** Reconstruction of man-made scenes from multi-view images is an important problem in computer vision and computer graphics. Observing that man-made scenes are usually composed of planar surfaces, we encode plane shape prior in reconstructing man-made scenes. Recent approaches for single-view reconstruction employ multi-branch neural networks to simultaneously segment planes and recover 3D plane parameters. However, the scale of available annotated data heavily limits the generalizability and accuracy of these supervised methods. In this paper, we propose multi-view regularization to enhance the capability of piecewise planar reconstruction during the training phase, without demanding extra annotated data. Our multi-view regularization enables the consistency among multiple views by making the feature embedding more robust against view change and lighting variations. Thus, the neural network trained by multi-view regularization performs better on a wide range of views and lightings in the test phase. Based on more consistent prediction results, we merge the recovered models from multiple views to reconstruct scenes. Our approach achieves state-of-the-art reconstruction performance compared to previous approaches on the public ScanNet dataset.

**Keywords** scene modeling; multi-view; regularization; neural network

## 1 Introduction

Multi-view reconstruction has been extensively studied in computer graphics and computer vision for decades. Studying this problem can be beneficial for many practical applications, such as indoor navigation, augmented reality, and human–robot interaction. However, reconstruction of indoor scenes is non-trivial due to object clutters, occlusions, large variety of object appearances, etc.

Traditional multi-view reconstruction methods build dense correspondence between pixels and recover the 3D positions of points. Hand-crafted similarity metrics (e.g., normalized cross-correlation and semi-global matching [1, 2]) are employed to compute point-wise correspondences among multiple views. Without integrating contextual information or global geometric prior, these methods have trouble in reconstructing complete and comprise models for scenes with repetitive patterns or many textureless regions.

With the rapid development of deep learning techniques, many deep neural network-based (DNN-based) methods have been proposed [3–6]. These methods usually compute the pixel-wise plane-sweep cost volume [7] from multiple view images and use the 3D convolutional neural network (CNN) to infer the depth map directly. These methods work well on reconstructing an object where the images are taken under multiple views around the object. However, the computational cost of using 3D CNNs to infer depth map from cost volumes is huge.

Man-made scenes, especially indoor scenes, are usually composed of plane surfaces [8]. This shape prior provides more global geometric constraint for depth inference and helps on generating cleaner scene models. Recently, a series of deep learning-based methods have been proposed to reconstruct piece-wise planar scenes from image [9–12]. From a single RGB image, these methods simultaneously segment plane instances and estimate 3D plane parameters. The

1 University of Science and Technology of China, Hefei, 230026, China. E-mail: W. Xi, xiwj@mail.ustc.edu.cn; X. Chen, xjchen99@ustc.edu.cn (✉).

Manuscript received: 2019-12-17; accepted: 2019-12-24

planarity hypothesis has been proven very effective in reconstructing man-made scenes [8, 13].

Although these deep learning methods have achieved promising success in single-view reconstruction, inadequate consideration of multi-view consistency and the deficiency of densely annotated training data lead to poor generalizability of the inference feature, especially when the viewpoint changes significantly. As a result, over-segmentation and under-segmentation often appear in the predicted plane instance, which gives rise to the wrong reconstruction of the 3D scene. One possible way to solve this problem is to increase the amount of training data. However, it is nontrivial to obtain accurate depth maps and plane annotations.

In this paper, we propose a multi-view regularization method which enhances existing single-view reconstruction networks for piecewise planar scenes to obtain more robust feature representation. Our multi-view regularization mainly works in the training phase. From multiple images of the same scene, we enforce the network to embed the inference features of the same plane instance under different views as the same as possible while the features of different plane instances are as different as possible. In the test phase, our method is similar to the existing single-view planar reconstruction network. Each single-view image is fed to the network to obtain an instance segmentation map of the plane and corresponding plane parameters to infer the scene depth map. Then the prediction results of multiple images are merged to compose the 3D scene, as shown in Fig. 1. Compared with the existing methods, our method does not bring additional calculation while improving network performance. During the training phase, there is no increase in the annotation amount of the training data.

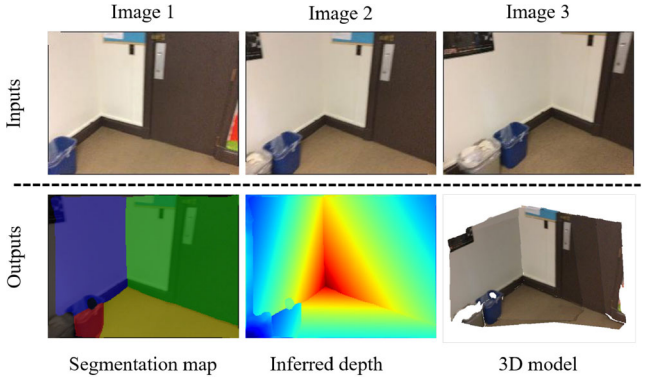



**Fig. 1** Given multiple images under different viewpoints, we reconstruct the scene by recovering plane segments and the depth map under the planar shape constraint for each image. Based on our multi-view regularization, the recovered planes and depths are more consistent and can be composed to reconstruct the scene more completely.

In summary, our contributions are the following:

- We propose multi-view regularization to enhance the feature extraction capability of existing single-view planar reconstruction networks, especially against to viewpoint changes and lighting variance.
- Our method does not require extra annotated training data and does not bring additional computational cost during the test phase.
- Our method achieves state-of-the-art planar reconstruction performance on the public dataset ScanNet [14].

## 2 Related work

**Multi-view reconstruction**. Conventional methods [1, 2, 15, 16] adopt plane-sweep algorithm to construct matching confidence, then optimize the disparity of reference frustum from matching confidence. Although these methods show successful results under ideal Lambertian scenarios [3], they suffer from poor generalizability to common scenes. Incomplete and incorrect reconstruction usually occur due to low texture, object occlusion, or transparency. Recently, deep learning techniques widely success in image recognition and detection, and many deep learning-based methods [3–6] have been proposed for 3D reconstruction. These methods construct a pixel-wise cost volume according to extracted 2D features, then use 3D CNN to directly regress the disparity of reference frustum. While these methods achieve successful performance in object reconstruction benchmark [17, 18], they are computationally expensive.

**Planar scene modeling**. For 3D plane modeling tasks, traditional methods extract 3D geometric cues and apply global geometric constraints in modeling process. Delage et al. [19] first extract line segments from the input image, and then use Markov Random Field (MRF) model to label superpixels according to the predefined plane classes. But this needs to be established under the assumption of the Manhattan world. Barinova et al. [20] first extract geometric primitives from an input 2D image, and then use the Conditional Random Field (CRF) model to

label the extracted geometric primitives. Similarly, assumption that the scene is composed of flat ground and vertical walls should be made. Although these methods can effectively extract and model the plane in the scene, they all need strong prior assumptions which severely restricts the generalizability of the algorithm.

**Learning-based methods**. In the literature, a series of learning-based methods have been proposed for plane scene modeling. Early work [21] introduces a learning-based framework to infer the depth map from a single RGB image. With the improvement of deep learning technology, more and more methods based on Convolutional Neural Networks (CNNs) have been proposed. However, most of these techniques simply produce the scene depth map without extracting plane information such as plane instance and plane parameters.

Recently, several CNN-based methods have been proposed to extract 3D plane structure. Among them, PlaneNet [9] and Yang and Zhou [10] use the semantic segmentation framework to extract plane instances in the scene, and directly infer plane parameters by a CNN. The difference between them is the way of supervision of the plane parameters. PlaneNet [9] uses traditional plane fitting methods to generate annotations for plane parameters, while Yang and Zhou [10] translate the plane parameter prediction problem into plane depth prediction of the scene and use the ground truth depth values to supervise the plane parameters. However, a drawback of these methods is that only a fixed number of planes can be inferred each time because of the widely used semantic segmentation network architecture. This problem was later solved by PlaneRCNN [11] and Yu et al. [12]. Based on the framework of Mask-RCNN, PlaneRCNN [11] first detects the plane in the image and then performs instance segmentation on the detected mask. Yu et al. [12] address this problem by mean-shift clustering of the extracted feature maps to perform instance segmentation for the planes in the scene.

## 3 Methodology

Our multi-view regularization method is proposed to enhance piece-wise planar scene reconstruction from a single RGB image. Combining the recovered 3D planes from multiple images, we could obtain a more complete model for the scene. The network framework adopted by our method is the single-view plane modeling network [12] as the backbone for feature inference. We first quickly introduce the network architecture and then introduce our multi-view regularization method. Figure 2 illustrates the framework of our proposed method.

### 3.1 Single-view planar modeling network

For the convenience of presentation, we denote SVPNet as the single-view plane network [12]. Taking a single color image as input, SVPNet infers a depth map from 3D plane recovery, and the blue block in Fig. 2 shows the

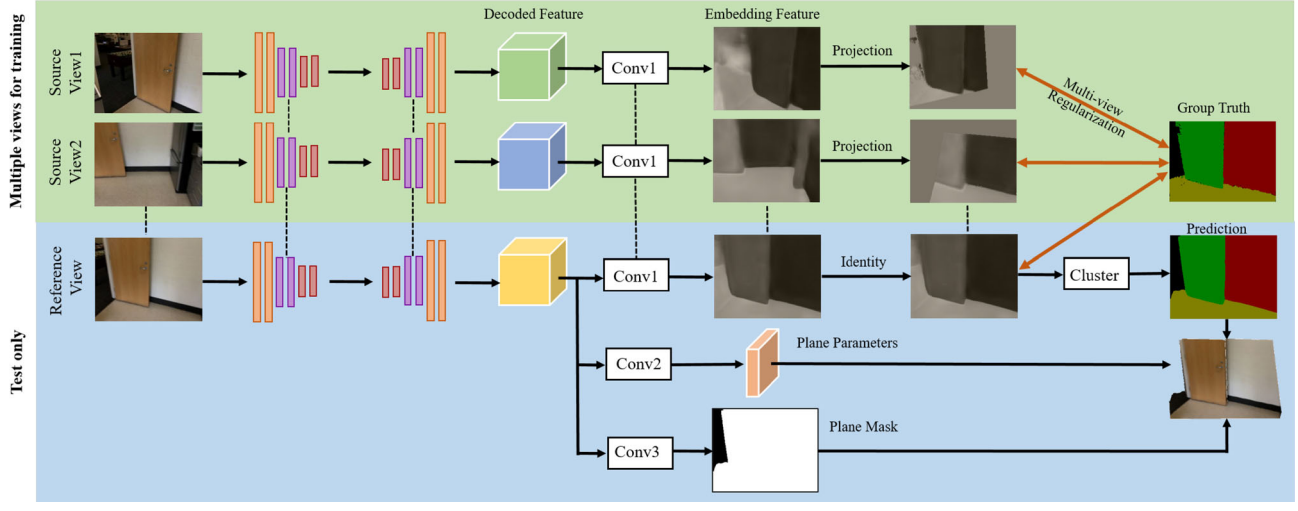



**Fig. 2** The architecture of the proposed multiview regularization. At the training phase, the input of the network contains a reference image with annotation and $K$ images under different views. The embedding features of the $K$ source images are projected from their original views onto the reference view. At the test phase, only each color image is fed to the reconstruction network for 3D modeling separately and then composed in 3D.

inference pipeline. It first exploits an encoder–decoder framework to extract features from the color image, then through three convolution layers to produce a 2-channel plane embedding feature map, pixel-wise plane parameters, and a binary semantic segmentation mask which indicate plane or not, respectively. The extracted plane embedding feature maps are clustered to obtain plane instance by a modified mean shift clustering algorithm. Concurrently, the scene depth map of plane regions is inferred from extracted plane parameters. Combining the instance-wise plane segmentation mask and inferred depth map, a piecewise 3D planar scene is modeled.

### 3.2 Multi-view regularization

The purpose of our multi-view regularization design is enforcing the consistency of plane feature embedding from different views against view change and lighting variance during the training phase. To enhance single-view planar scene modeling performance, we take multiple images under different views of the same scene as input for training. Among these images, one image is labeled with ground truth plane instance segmentation labels $S$, plane parameters $\{P_i\}_{i=1}^{C}$, and scene depth $D$. The image with annotation we called reference image $I^{\mathrm{r}}$ and the other $K$ images are called source-view images. All of these images are fed into $K+1$ share weight sub-networks which include encoder–decoder block and embedding layer which to extract plane embedding feature $X^{\mathrm{r}}, X_i^{\mathrm{s}}, \cdots, X_K^{\mathrm{s}}$ for each input color image.

To establish the pixel-wise correspondence between the reference image and $K$ source-view images, the embedding feature maps in the source-views $X_1^{\mathrm{s}}, \cdots, X_k^{\mathrm{s}}$ are projected into the reference view according to their camera parameters and ground truth depth map. Therefore, we obtain $K+1$ projected embedding feature maps from the $K$ source views and reference view, denote as $\{X_i^p\}_{i=1}^{K+1}$.

In order to formulate the multi-view regularization term, the discriminative loss function is extended from Ref. [22] to multi-view feature maps. The single-view discriminative loss function is defined as

$$\mathcal{L}_{\mathrm{sv}} = \mathcal{L}_{\mathrm{dist}} + \mathcal{L}_{\mathrm{var}} \tag{1}$$

where the variance term $L_{\mathrm{var}}$ pulls each embedding to the mean embedding of the corresponding plane instance, and the distance term $L_{\mathrm{dist}}$ pushes the embedding centers of each plane instances away from each other. The definition of them are

$$\mathcal{L}_{\mathrm{var}}(\{\mu_c\}^C, X, S) = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{i=1}^{N_c} f\left(\parallel \mu_c - x_i \parallel - \delta_{\mathrm{v}}, 0\right) \tag{2}$$

$$\mathcal{L}_{\mathrm{dist}}(\{\mu_c\}^C) = \frac{1}{Z_C}\sum_{\substack{c_A=1 \\ c_A \neq c_B}}^{C}\sum_{c_B=1}^{C} f\left(2\delta_{\mathrm{d}} - \parallel \mu_{c_A} - \mu_{c_B} \parallel, 0\right) \tag{3}$$

where $Z_C = C(C-1)$ is a normalization constant acting as the average of the push distance between different plane instances, $f(x) = \max(x, 0)$, $C$ is the number of the ground truth plane instances, $N_c$ is the number of pixels in the plane cluster $c$, $\mu_c$ is the mean embedding of the plane cluster $c$, and $\{\mu_c\}^C$ is the overall set of mean embedding, $x_i$ is the $i^{\mathrm{th}}$ embedding vector on plane cluster $c$ of the embedding feature $X$, $\parallel \cdot \parallel$ is the L2 distance term, $\delta_{\mathrm{v}}$ and $\delta_{\mathrm{d}}$ are the two margins for $L_{\mathrm{var}}$ and $L_{\mathrm{dist}}$ respectively.

The single-view loss is used to constrain the compactness of the embedding space, while our multi-view regularization term tends to enforce the consistency of the embedding vectors of the same plane instance under different views. For each source-view image, we get the projected embedding feature map $X_i^{\mathrm{p}}$ and the corresponding plane instance label $S_{\mathrm{r}}$ which is the annotation under reference view image. Combining these projected embedding feature maps with the embedding feature map of the reference view, we denote the as $\{F_k\}^{K+1} = \{X^{\mathrm{r}}, X_1^{\mathrm{p}}, \ldots, X_K^{\mathrm{p}}\}$. For each plane instance that can be extracted from $S_{\mathrm{r}}$, we compute the multi-view regularization term. Specifically, we calculate the mean embedding $\mu_c$ of all the pixels in each plane instance among all the $K+1$ embedding feature maps $\{F_k\}^{K+1}$. With the shared mean embedding $\mu_c$ among all $K+1$ views, our multi-view regularization term is given by

$$\mathcal{L}_{\mathrm{mv}} = \mathcal{L}_{\mathrm{dist}}(\{\mu_c\}^C) + \frac{1}{K+1}\sum_{k=1}^{K+1}\mathcal{L}_{\mathrm{var}}(\{\mu_c\}^C, F_k, S_{\mathrm{r}}) \tag{4}$$

Since the ultimate goal is to recover the plane instances and their geometry, we combine the semantic segmentation loss $\mathcal{L}_{\mathrm{s}}$, plane parameter loss $\mathcal{L}_{\mathrm{n}}$ which are defined in Ref. [12] with our multi-view regularization term $\mathcal{L}_{\mathrm{mv}}$ to train our designed model. Therefore, the overall loss function of the network is written as

$$\mathcal{L} = \mathcal{L}_{\mathrm{s}} + \mathcal{L}_{\mathrm{n}} + \mathcal{L}_{\mathrm{mv}} \tag{5}$$

All these loss function terms are differentiable, so

the entire network can be trained in an end-to-end manner.

## 4 Experiments

We first evaluate our multi-view regularization term on the task of single view reconstruction both quantitatively and qualitatively. Then we show a series of reconstruction results by merging the inferred geometry from multiple views. We implement our method with PyTorch [23]. We use SGD optimizer [24] with a learning rate of $10^{-4}$ and a weight decay of $10^{-5}$, and the batch size is set to 16.

We construct the dataset for training and test based on the ScanNet [14]. Following previous single-view piecewise planar scene modeling works [9, 12], we first construct a dataset containing 48,214 reference images for training. For each reference image, the ground truth annotation with plane instance segmentation map and plane parameters are obtained by fitting planes on the reconstructed mesh in the scenes of ScanNet. To train the network with the multi-view regularization, we select two images which are 20 frames before and after each reference image in the image sequence with corresponding camera intrinsic and extrinsic parameters. At the test phase of single view reconstruction, the trained network takes only one color image as input. For fairness, we evaluate our model on the test set which consists of 760 images, the same as Refs. [9, 12].

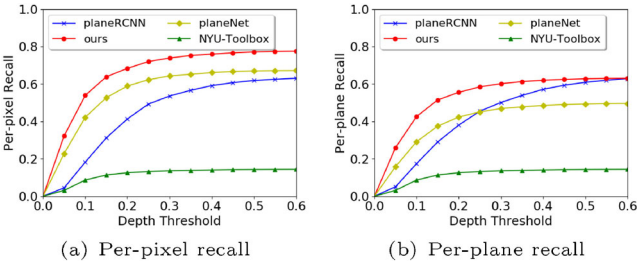
### 4.1 Quantitative evaluation

We first compare our model with the baseline model SVPNet [12]. We evaluate the planar scene modeling performance with the per-plane recall and per-pixel recall metrics which are the percentage of correct prediction of the planes and pixels, respectively. The predicted plane is considered a correct prediction when plane instance segmentation intersection-over-union (IOU) score with the ground truth segmentation map is larger than 0.5 and the mean difference between the inferred depth and ground truth depth map is less than a threshold $\sigma_{\mathrm{d}}$, which ranges from 0.05 to 0.6 m. The experimental comparison results are shown in Table 1. We can see that our model achieves better planar scene modeling performance than the baseline model, especially when the depth threshold is small. Considering that our method does not bring any additional network parameters and computational cost compared to the baseline model at the test phase, we can conclude that the proposed multi-view regularization improves the capability of the single-view inference network.

Furthermore, we compare our method with PlaneRCNN [11], PlaneNet [9], and NYU-Toolbox [25]. Figure 3 shows the comparison results the test set. Our method achieves the best performance among these methods, which ulteriorly demonstrates the effectiveness of our method.

### 4.2 Qualitative evaluation

Figure 4 shows a group of results of plane instance segmentation, depth estimation, and reconstructed models using our method and the baseline model [12] on the ScanNet dataset [14]. Comparing (b) and (c) in Fig. 4, we can observe that our method enhances the ability of the network of detecting and




(a) Per-pixel recall (b) Per-plane recall

**Fig. 3** Comparison between the enhanced SVPNet by our multi-view regularization term and three existing methods for single-view planar scene reconstruction on the ScanNet test set.

**Table 1** Per-pixel and per-plane recalls on the ScanNet test dataset compared with the baseline SVPNet [12]

| | Depth threshold $\sigma_{\mathrm{d}}$ (m) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 | 0.55 | 0.60 |
| | Per-pixel recall (%) | | | | | | | | | | | |
| Baseline | 30.65 | 51.26 | 62.54 | 68.10 | 71.62 | 73.79 | 74.85 | **76.03** | 76.53 | 76.78 | 77.01 | 77.31 |
| Ours | **31.85** | **53.09** | **63.47** | **68.44** | **72.12** | **74.25** | **75.14** | 75.89 | **76.86** | **77.13** | **77.47** | **77.53** |
| | Per-plane recall (%) | | | | | | | | | | | |
| Baseline | 23.07 | 39.58 | 49.22 | 54.16 | 57.46 | 59.50 | 60.61 | **61.53** | 61.88 | 62.14 | 62.40 | 62.56 |
| Ours | **25.45** | **42.01** | **50.60** | **55.04** | **58.01** | **59.90** | **60.81** | 61.47 | **62.06** | **62.30** | **62.56** | **62.60** |

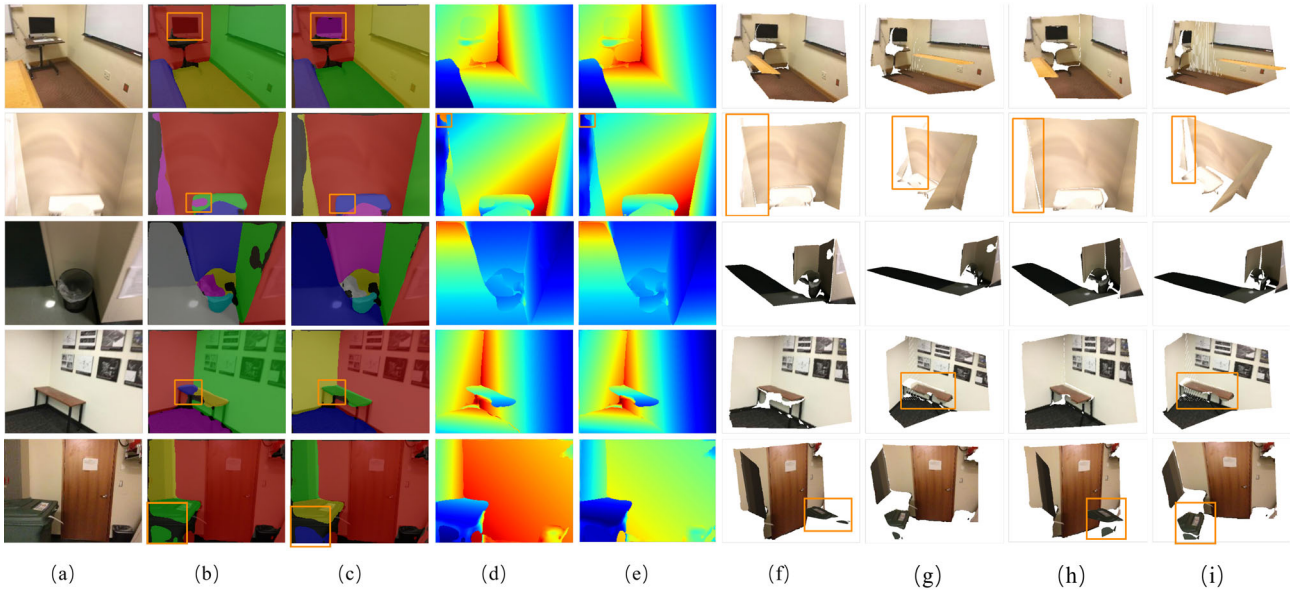



**Fig. 4** Comparison between the single-view reconstruction results of our method and the baseline SVPNet [12]. (a) Input images. (b) The plane clustering results by SVPNet and (c) our method. (d) The inferred depth map by SVPNet and (e) our method. (f, g) Two novel views of the recovered models using SVPNet and (h, i) our method.

recovering small plane instances and improves the model robustness to avoid over-segmentation of entire plane into multiple pieces. From Figs. 4(d)–4(i), we conclude that our method also enhances the ability of the network to recover correct 3D plane parameters because the three branches of the network share the same feature extractor and they can be mutually improved by joint training.

### 4.3 Multiple view consistency evaluation

The goal we designed the multi-view regularization is to enhance the feature consistency of multiple views, thereby improving the robustness of the network against view change and lighting variance. We validate the multi-view consistency of our method from two perspectives: plane segmentation consistency and 3D planar model correctness.

**Plane segmentation consistency**. We randomly select three images from different views in the same scene and test our enhanced single-view reconstruction network on each image independently. Figure 5 displays the plane instance segmentation results obtained by our method and three state-of-the-art approaches [9, 11, 12]. It shows that our method achieves the best result in maintaining the segmentation consistency among different views. Compared with the existing methods, our method can effectively reduce over-segmentation and under-segmentation as highlighted by the orange boxes. Moreover, the plane boundaries generated by our methods are more clear and consistent.

**Geometry correctness**. For the task of modeling piecewise planar scenes, the ultimate goal is to get a complete and correct 3D model of the scene. Given multiple images taken under different views in the same scene, we first run our enhanced network to reconstruct the plane instances for each image. Then we transform the reconstructed mesh from each image to the reference view according to their associated camera parameters and merge the meshes to obtain a more complete reconstruction result of the scene. Figure 6 shows the reconstructed results of two scenes using our method and the baseline model [12]. For each scene, we select three frames under adjacent viewpoints for reconstruction. Generally, our method reconstructs more view-consistent results. In the first scene, the red book flys away in the reconstructed model using SVPNet because it is segmented to the wall region by mistake. The desktop area is stretched a lot using SVPNet [12]. In comparison, the shape of the desktop is well recovered and consist among different views. In the second scene, the wall region reconstructed from different views using our method is much more consistent than that reconstructed using SVPNet [12]. These examples illustrate that our multi-view regularization can enhance the robustness of the single-view reconstruction network against view

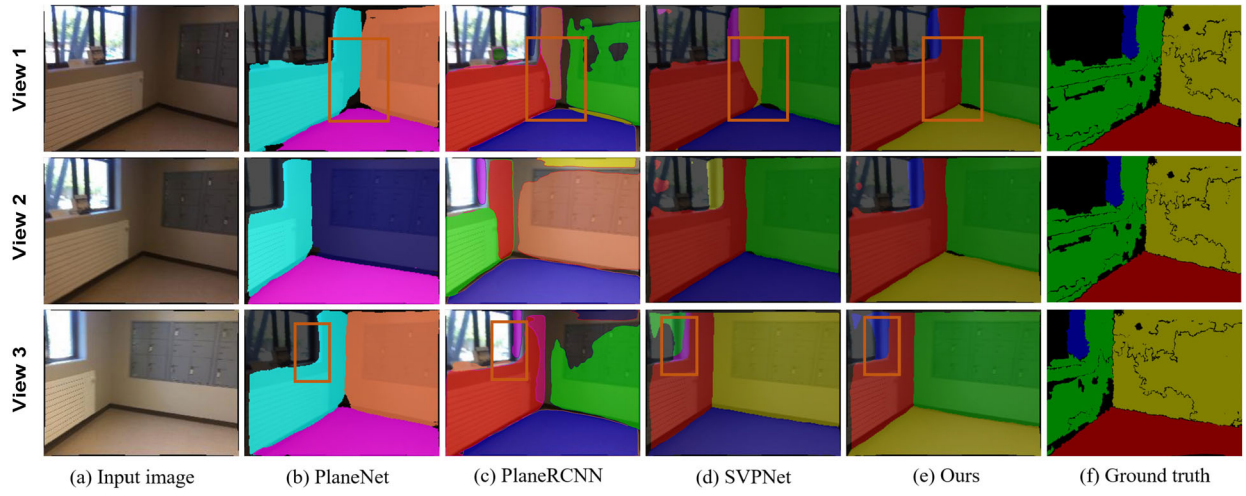



**Fig. 5** Comparison of the plane segmentation results of multiple images under different views using our method and other approaches.

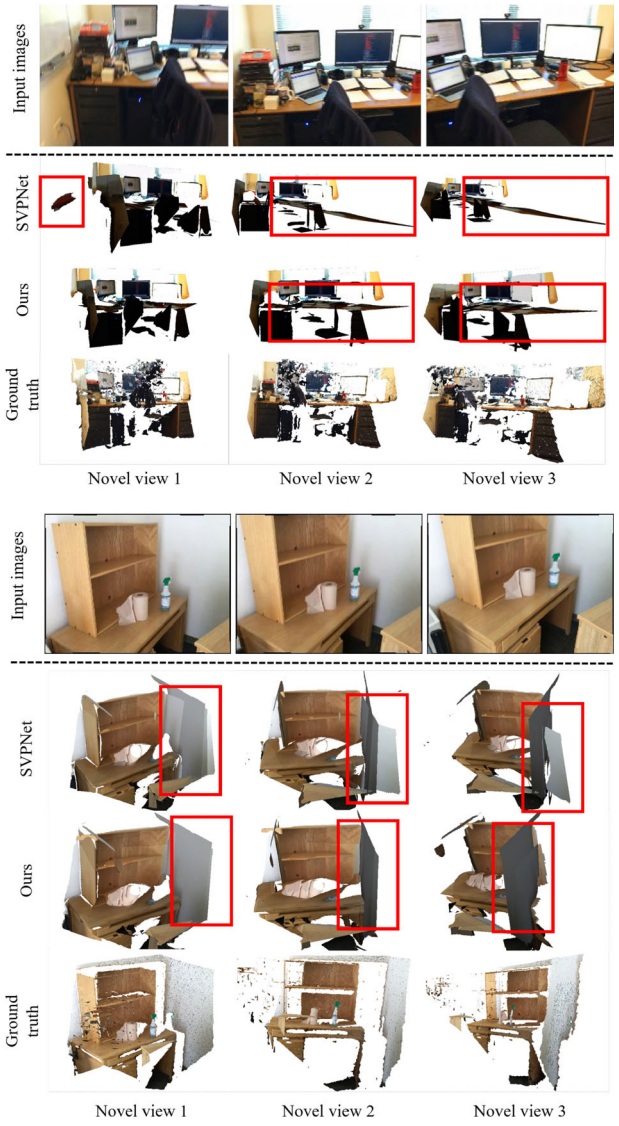



**Fig. 6** The reconstruction results of two scenes from multiple images using our method and SVPNet [12].

change and lighting variation, and therefore a more accurate 3D scene model can be obtained.

## 5 Conclusions

In this paper, we propose a novel method to enhance a single-view reconstruction network by multi-view regularization for modeling piecewise 3D planar scenes. Our method enforces the consistency of the embedding features during the training phase, thereby enhances the robustness of the network against view change and lighting variation. Our method achieves state-of-the-art performance on the task of indoor scene reconstruction on the public ScanNet dataset. We believe that our proposed multi-view regularization can be flexibly integrated with many single-view inference networks without bringing extra computational cost.

## Acknowledgements

This work was supported by the National Key R&D Program of China under Grant 2017YFB1002202, the National Natural Science Foundation of China (NSFC) under Grant 61632006, as well as the Fundamental Research Funds for the Central Universities under Grants WK3490000003 and WK2100100030.

## References

[1] Gallup, D.; Frahm, J.-M.; Mordohai, P.; Yang, Q.; Pollefeys, M. Real-time plane-sweeping stereo with multiple sweeping directions. In: Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition, 1–8, 2007.

[2] Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 30, No. 2, 328–341, 2008.

[3] Yao, Y.; Luo, Z. X.; Li, S. W.; Fang, T.; Quan, L. MVSNet: Depth inference for unstructured multi-view stereo. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11212*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer International Publishing, 785–801, 2018.

[4] Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; Quan, L. Recurrent MVSNet for high-resolution multiview stereo depth inference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5525–5534, 2019.

[5] Chen, R.; Han, S.; Xu, J.; Su, H. Point-based multiview stereo network. In: Proceedings of the IEEE International Conference on Computer Vision, 1538–1547, 2019.

[6] Luo, K.; Guan, T.; Ju, L.; Huang, H.; Luo, Y. PMVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo. In: Proceedings of the IEEE International Conference on Computer Vision, 10452–10461, 2019.

[7] Yang, R.; Pollefeys, M. Multi-resolution real-time stereo on commodity graphics hardware. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003.

[8] Monszpart, A.; Mellado, N.; Brostow, G. J.; Mitra, N. J. RAPter: Rebuilding man-made scenes with regular arrangements of planes. *ACM Transactions on Graphics* Vol. 34, No. 4, Article No. 103, 2015.

[9] Liu, C.; Yang, J.; Ceylan, D.; Yumer, E.; Furukawa, Y. PlaneNet: Piece-wise planar reconstruction from a single RGB image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2579–2588, 2018.

[10] Yang, F. T.; Zhou, Z. H. Recovering 3D planes from a single image via convolutional neural networks. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11214*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 87–103, 2018.

[11] Liu, C.; Kim, K.; Gu, J.; Furukawa, Y.; Kautz, J. PlaneRCNN: 3D plane detection and reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4450–4459, 2019.

[12] Yu, Z.; Zheng, J.; Lian, D.; Zhou, Z.; Gao, S. Single-image piece-wise planar 3D reconstruction via associative embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1029–1037, 2019.

[13] Zhang, Y. Z.; Xu, W. W.; Tong, Y. Y.; Zhou, K. Online structure analysis for real-time indoor scene reconstruction. *ACM Transactions on Graphics* Vol. 34, No. 5, Article No. 159, 2015.

[14] Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; Niessner, M. ScanNet: Richlyannotated 3D reconstructions of indoor scenes. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, 5828–5839, 2017.

[15] Furukawa, Y.; Ponce, J. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 32, No. 8, 1362–1376, 2010.

[16] Schönberger, J. L.; Zheng, E. L.; Frahm, J. M.; Pollefeys, M. Pixelwise view selection for unstructured multi-view stereo. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9907*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer International Publishing, 501–518, 2016.

[17] Jensen, R.; Dahl, A.; Vogiatzis, G.; Tola, E.; Aanaes, H. Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 406–413, 2014.

[18] Knapitsch, A.; Park, J.; Zhou, Q.-Y.; Koltun, V. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics* Vol. 36, No. 4, Article No. 78, 2017.

[19] Delage, E.; Lee, H.; Ng, A. Y. Automatic single-image 3d reconstructions of indoor manhattan world scenes. In: *Robotics Research. Springer Tracts in Advanced Robotics, Vol. 28*. Thrun, S.; Brooks, R.; Durrant-Whyte, H. Eds. Springer Berlin Heidelberg, 305–321, 2007.

[20] Barinova, O.; Konushin, V.; Yakubenko, A.; Lee, K.; Lim, H.; Konushin, A. Fast automatic single-view 3-d reconstruction of urban scenes. In: *Computer Vision – ECCV 2008. Lecture Notes in Computer Science, Vol. 5303*. Forsyth, D.; Torr, P.; Zisserman, A. Eds. Springer Berlin Heidelberg, 100–113, 2008.

[21] Saxena, A.; Chung, S. H.; Ng, A. Y. Learning depth from single monocular images. In: Proceedings of the 18th International Conference on Neural Information Processing Systems, 1161–1168, 2005.

[22] De Brabandere, B.; Neven, D.; Van Gool, L. Semantic instance segmentation for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 7–9, 2017.

[23] Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In: Proceedings of the 31st Conference on Neural Information Processing Systems, 2017.

[24] Zhang, T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Proceedings of the 21st International Conference on Machine Learning, 2004.

[25] Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from RGBD images. In: *Computer Vision – ECCV 2012. Lecture Notes in Computer Science, Vol. 7576*. Fitzgibbon, A.; Lazebnik, S.; Perona, P.; Sato, Y.; Schmid, C. Eds. Springer Berlin Heidelberg, 746–760, 2012.

**Weijie Xi** is a master candidate in the Department of Electronic Engineering and Information Science, University of Science and Technology of China. His research interests focus on geometry in computer vision. Weijie Xi obtained his B.S. degree from Chongqing University in 2018. He started his master in University of Science and Technology of China in 2018.

**Xuejin Chen** is an associate professor of the University of Science and Technology of China. She received her B.S. degree in 2003 and Ph.D. degree in 2008 from the University of Science and Technology of China (USTC). She conducted research as a postdoctoral scholar in the Computer Graphics Lab at Yale University from 2008 to 2010. She visited Stanford University from Feb. to Aug. 2017. Her research interests include 3D modeling, geometry processing, sketch-based content generation, and scene understanding.