

Weakly supervised learning of indoor geometry by dual warping

Pulak Purkait Ujwal Bonde Christopher Zach
Toshiba Research Europe, Cambridge, U.K.

{pulak.cv, ujwal.bonde, christopher.m.zach}@gmail.com

Abstract

A major element of depth perception and 3D understanding is the ability to predict the 3D layout of a scene and its contained objects for a novel pose. Indoor environments are particularly suitable for novel view prediction, since the set of objects in such environments is relatively restricted. In this work we address the task of 3D prediction especially for indoor scenes by leveraging only weak supervision. In the literature 3D scene prediction is usually solved via a 3D voxel grid. However, such methods are limited to estimating rather coarse 3D voxel grids, since predicting entire voxel spaces has large computational costs. Hence, our method operates in image-space rather than in voxel space, and the task of 3D estimation essentially becomes a depth image completion problem. We propose a novel approach to easily generate training data containing depth maps with realistic occlusions, and subsequently train a network for completing those occluded regions. Using multiple publicly available dataset [18, 12] we benchmark our method against existing approaches and are able to obtain superior performance. We further demonstrate the flexibility of our method by presenting results for new view synthesis of RGB-D images.

1. Introduction

Scene completion has drawn a lot of attention recently from the computer vision [19], robotics [9] as well as the neuroscience community [5]. Most of these works are driven by the assumption that 3D scene completion is important for 3D scene understanding which in turn is useful for tasks such as robot navigation. Towards this end, recent works have addressed 3D scene completion by semantic voxel filling [18, 1]. However, these approaches are limited as follows: (i) semantic labeling of 3D voxels will generally produce coarse labelings in order to be computationally feasible, and (ii) labeling of voxels in the object interior might be redundant as one is mostly interested in object surfaces. Therefore, we focus on predicting detailed surfaces rather than semantic voxels. Moreover, methods

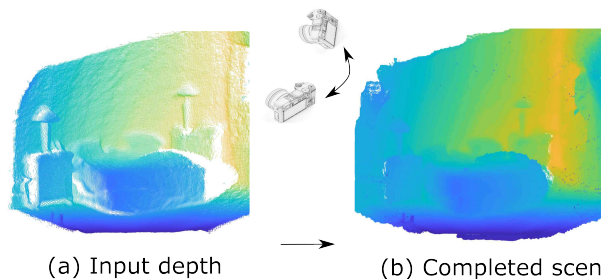


Figure 1. Given an input depth image proposed network can generate multiple depth images which can be further combined for 3D scene completion.

predicting entire voxel grids rely on large labeled datasets that are expensive to create and label [18]. In contrast, our system does not require any additional labeled data and only relies on calibrated depth images which are easy to acquire using existing RGB-D sensors [16]. An instance of the predicted output is displayed in Fig. 1. In summary our contributions are as follows:

- We propose a network architecture to predict depth at arbitrary viewpoints given a single depth image.
- The network is trained solely using unlabeled depth images without relying on additional supervision signals.

2. Literature Review

In computer vision literature the problem of scene completion was addressed by some of the earliest work into human perception understanding [10], where it is conjectured that human perception relies on its ability to complete scenes. With the availability of cheap sensors [16] and computational resources a renewed interest is seen in this field. In [18, 2] the authors predict the complete 3D scene from a single depth image using a network to learn shape prior of indoor objects. Given a new scene they predict the voxelised volume with semantic labels for each voxel along with its occupancy probability. Learning these prior requires a large synthetic dataset [18, 1] or the need to manually label real world data [12], both of which are expensive procedures. In comparison our method solely relies

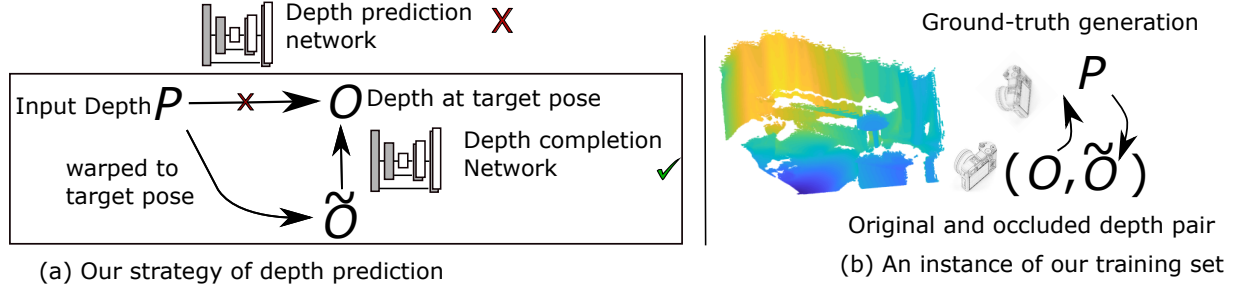


Figure 2. We avoid predicting depth at the target pose, i.e., $P \rightarrow \text{depth prediction network} \rightarrow O$, by the following depth completion strategy: $P \rightarrow \text{warp to target pose} \rightarrow \tilde{O} \rightarrow \text{depth completion network} \rightarrow O$.

on (synthetically) transforming calibrated depth images to generate training data. Moreover the final prediction in the above-mentioned works is a coarse grid of voxels whereas our system outputs a full-resolution depth image for a desired viewpoint.

Our approach shares similarities with [22] and [13], which also use displacement fields to solve an image completion/inpainting task. However, these methods are restricted to single or few objects and its unclear how they would generalize to natural scenes.

We model the scene completion task as an instance of depth image inpainting (e.g. [7]). Using low rank approximations is a popular framework for image inpainting [6]. In [20] the authors extend this approach from RGB to depth images using additional regularization to the gradients. Similar to our work they do not require additional labeled data. However the noise patterns used in image inpainting are often either random and unstructured (e.g. salt and pepper noise) or structured but artificial (e.g. text superimposition). In contrast this work explicitly considers naturally occurring structured missing regions, i.e. occlusions generated by warping images based on a given depth map.

The authors of [14, 23] use RGB information to complete a sparse depth image. This is restrictive as we can only complete viewpoints that have corresponding RGB images. On the other hand as we do not rely on RGB images we are able to synthesize arbitrary viewpoints within a reasonably distance from the observed depth map.

3. Dataset generation via dual warping

Given a depth image at a particular pose our target is to generate depth views at arbitrary locations. A natural way to accomplish this would be to generate a pair of depth images of the same scene from different view-points and use them as ground truth. However, for this we would require the complete 3D model of a scene from which synthetic pairwise views could be generated, i.e. two different depth images of the same scene with the ground truth poses. These 3D models are rarely available or time-consuming to acquire.

In this work we exploit a novel strategy to generate training data solely from given depth images. Let P be a given depth image and O be the depth image at the target pose that we want to estimate. The estimation of the depth image at a novel view-point $P \rightarrow O$ consists of is modelled via two stages: (i) a geometric warping step and (ii) filling of the occluded regions. The former is very straight-forward and can be computed efficiently. Thus, we pose the novel view generation as a depth completion problem (Fig. 2). A convolutional network is trained for this task.

A strongly supervised approach requires training data consisting of depth map pairs (O, \tilde{O}_P) , where \tilde{O}_P is the depth map P warped to the pose of O , and the task of the DNN is to fill in missing depth values in \tilde{O}_P to match O . It therefore requires acquisition of multiple (at least two) depth maps for each scene, and a strongly supervised method is consequently not applicable on e.g. unordered collections of unrelated depth maps. Thus, we replace the strongly supervised task by a weakly supervised one, which—as a by-product—turns out to be also less challenging in terms of problem difficulty (see below). Let \tilde{O} be the depth map obtained by warping \tilde{P}_O (i.e. O warped to the pose of P) back to the pose of O , then the training data consists of pairs (O, \tilde{O}) . Since a given depth map is warped twice we call it “dual warping” (see Fig. 2). It only requires independent depth images $\{O\}$ and a method to generate realistic nearby poses (corresponding to the poses of depth maps $\{P\}$, if they were supplied). Thus, we state our first strategy for training data generation below:

Strategy 1. *The occlusion $O \setminus \tilde{O}$ generated by warping forth and back serves us the ground truth occluded and complete image pair (\tilde{O}, O) .*

Note that some parts of the depth map O become occluded during the “dual warping”. Further, $O_y = \tilde{O}_y$ for all pixels y with visible depth $\tilde{O}_y > 0$, and e.g. $\tilde{O}_y = 0$ for occluded pixels. To this end, one can raise a fundamental question: why does one require a complex strategy 1 to train a depth completion network. A straight-forward choice would be following one:

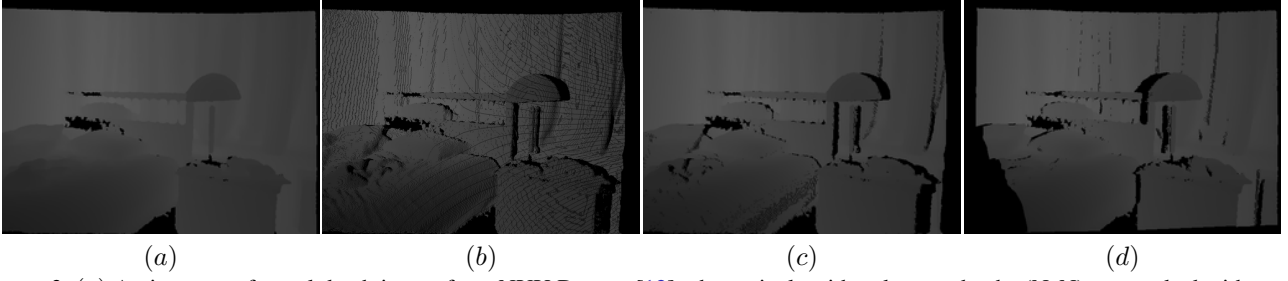


Figure 3. (a) An instance of a real depth image from NYU Dataset [12] where pixels with unknown depths (NaN) are marked with zero. (b) Depth after warped forth and back from the original depth to a random location and orientation. (c) Same as (b) but projected with upscaling with resolution factor 2 to reduce the aliasing affect. (d) An additional instance of (c) with opposite view-points. Note that the pairs $((c), (a))$ and $((d), (a))$ serve as the ground truths for our depth completion network.

Strategy 2. *Removing random regions at arbitrary pixel locations in the depth images—the occluded and the original depth image pair (\tilde{O}, O) can serve as the ground truth for depth image completion.*

However, we argue that there is a clear shift in domains between the training data (where random missing regions are presented to the network) and test data (where missing regions are occurring due to occlusions). We claim (and experimentally verify) that strategy 1 brings the training distribution closer to the test distribution, and its properties are further discussed in Sec. 3.2. In Sec. 5.2 we also validate that strategy 2 performs inferior to our dual warping strategy 1 for dataset generation.

3.1. Warping procedure

Let (x, y) be the original pixel coordinates of the depth image and $K = [f, 0, \bar{x}; 0, f, \bar{y}; 0, 0, 1]$ be the camera matrix where f is the focal length and (\bar{x}, \bar{y}) are the principle point of the camera. Further, let s_{xy} be the depth at pixel (x, y) . The depth at the corresponding pixel (x', y') at the relative pose (R, T) can be written as

$$s_{x'y'}\mathbf{x}' = K(RK^{-1}s_{xy}\mathbf{x} + T) \quad (1)$$

where \mathbf{x} and \mathbf{x}' are the homogeneous pixel co-ordinates at (x, y) and (x', y') respectively. We utilize (1) for forward and backward warping. Hence, our warping procedure essentially corresponds to rendering of 3D point clouds with a z-buffer test enabled for hidden surface removal. In the following section we postulate and empirically validate that the above strategy will not introduce any additional occlusion (up to aliasing effects due to point instead of mesh rendering). Further, the aliasing effect is addressed by upscaling the depth image by a factor of 2 before warping to the target pose. Note that the camera matrices are modified accordingly (i.e. $[2f, 0, 2\bar{x}; 0, 2f, 2\bar{y}; 0, 0, 1]$) while warping with the higher resolutions. The effectiveness of our warping strategy is demonstrated by the example shown in Fig. 3.

Each depth image is warped to a random pose and then warped back to the original pose. These random poses

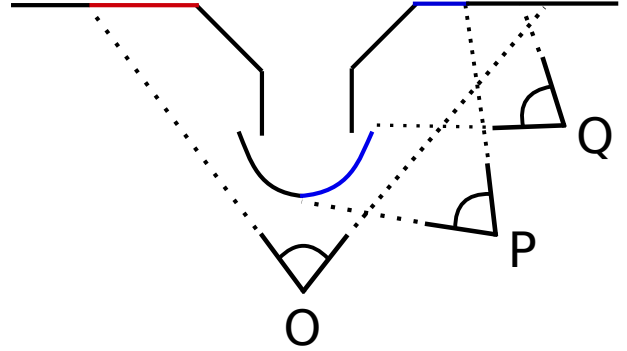


Figure 4. An example where \tilde{O}_P (warped forth and back to the pose P) and $P^{[O]}$ observe similar objects (marked by blue). In contrast, \tilde{O}_Q (warped forth and back to the pose Q) observe more objects than $Q^{[O]}$ (marked by red).

are generated on the horizontal plane where the translation and orientation increments are uniformly sampled within the range of $[-1m, 1m]$ and $[-15^\circ, 15^\circ]$. Our synthesized poses thus emulate essentially the lateral motion of a hand-held depth camera. The axis of the angular shift is chosen as vertical. Each warping generates a pair of original and occluded image. We generate 25 different original-occluded image pairs for each depth image. Thus the size of our ground truth dataset is $25 \times$ the original depth image dataset. Note that \tilde{O} and $P^{[O]}$ are depth images observe scene areas visible from both of the poses of O and P .

3.2. Analysis of the generated occlusion patterns

Although in strategy 1 we warp twice (to and from a random location), we (essentially) do not introduce any additional occlusions. In fact, the occluded region generated by the strategy is contained in the occluded region generated by warping an actual depth image P at the random pose to the original pose:

Lemma 1. *The occlusion $O \setminus \tilde{O} \subseteq O \setminus P^{[O]}$ where $P^{[O]}$ is the depth image generated by the warping the depth image*

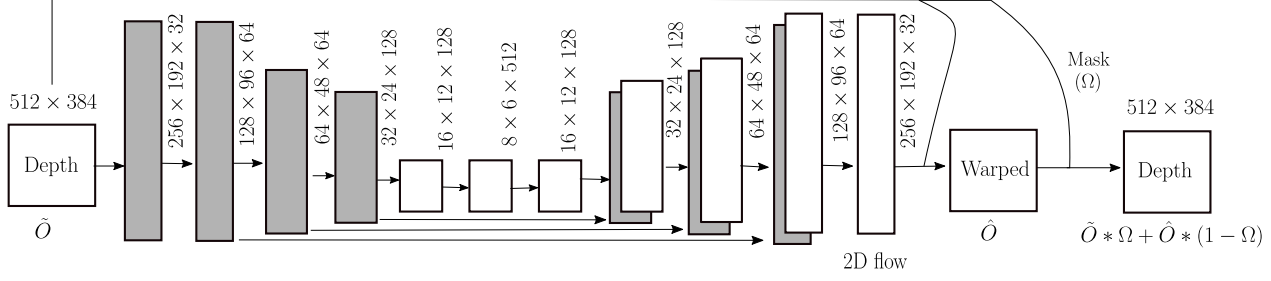


Figure 5. Architecture: The encoder and the decoder is marked by gray and white respectively. The network estimates 2D flow of the source and the target pixel locations of similar depth. The depth at the source locations are copied at the target pixel locations with the unknown depth. The depth at the known pixel locations are kept unaltered. The network is trained with the training image pairs (O, \tilde{O}) and minimize the \mathcal{L}_{ℓ_1} loss of the occluded part [see eq. (3)]. No that no direct supervision for the source and target flow is provided and let the network learn the flow directions for which the loss is minimum.

P to the original pose. Further, if the camera at P does not observe any additional objects, then $O \setminus \tilde{O} = O \setminus P^{[O]}$.

Proof. Let us denote the warping step of a depth map O into the pose of P yielding $O^{[P]}$ by $O \xrightarrow{P} O^{[P]}$, and let $\tilde{O} := (O^{[P]})^{[O]}$ be the result of dual warping $O \xrightarrow{P} O^{[P]} \xrightarrow{O} \tilde{O}$. By construction a 3D point X visible in O (written as $X \in O$) is not visible in \tilde{O} iff $X \notin O^{[P]}$. Compared to $O^{[P]}$ the true depth map P may contain additional surfaces occluding X at the pose of P , hence $X \notin O^{[P]}$ (or $X \notin \tilde{O}$) implies $X \notin P$ (equivalence holds if P has no additional occluders not present in $O^{[P]}$ blocking X , see Fig. 4).¹ Together with $X \notin P$ implying $X \notin P^{[O]}$ we have that $X \notin \tilde{O}$ is a sufficient condition for $X \notin P^{[O]}$, or $O \setminus \tilde{O} \subseteq O \setminus P^{[O]}$. \square

4. Architecture and loss

We follow the U-Net architecture very similar to [15]. Its a feed-forward convolutional network consists of an encoder and a symmetric decoder, where a number of skip connections is introduced by concatenating the features from the encoder layer to the corresponding decoder layer. The network takes an input depth image of size 512×384 and passes the input through multiple convolutional and deconvolutional layers (with stride 2) to predict a 2D displacement field of the same size as the input. Here the displacements (similar to pixel-shifts [21]) indicate shifts between the source pixel and target pixel location. Note that target pixel locations are the pixels with missing depth. The task of the network is to predict the corresponding source pixel locations (from which the known depth value is subsequently copied) instead of directly hallucinating the depth value at the target pixel location. The details of the architecture can also be found in Fig. 5. Note that experimentally we observe (validated in the result section) that dis-

placement estimation network performs better than the direct depth prediction network. In contrast to the single image depth prediction networks [3] (RGB to depth), in our case the depth for unknown regions is indirectly estimated by copying from known depth map portions.

We utilize a masked ℓ_1 loss \mathcal{L}_{ℓ_1} in this work. The mask Ω is considered as the pixels with the unknown depths $\Omega := O \setminus \tilde{O}$. We also incorporate a total variation loss \mathcal{L}_{tv} to ensure smooth depth predictions, and we further leverage a content loss [8], \mathcal{L}_c , to preserve structure of the depth image as described below:

$$\mathcal{L}_{tv} = \sum_{y \in \Omega} \left(\|O_y - \hat{O}_y\|_1 + \lambda \|\nabla \hat{O}_y\|_1 \right) \quad (2)$$

$$\mathcal{L}_c = \gamma \sum_{y \in \Omega} \|\phi_l(O)_y - \phi_l(\tilde{O})_y\|_1 \quad (3)$$

where \hat{O}_y is the predicted depth at the pixel y and ϕ_l are the feature descriptors at the layer l . Note that the depth is predicted only at the unknown pixels and the feature descriptors in the loss (3) is only considered for the last two layers. The network is trained to minimize the sum of the above loss $\mathcal{L}_{\ell_1} = \mathcal{L}_{tv} + \mathcal{L}_c$. λ and γ are chosen as 10^{-3} and 10^{-5} respectively.

5. Experiments

The proposed depth completion network (named as **Depth-Flow-Net**) is evaluated on the widely used SUNCG [18] and NYU Depth v2 [12] datasets. The loss \mathcal{L}_{ℓ_1} is minimized using ADAM with a mini-batch of size 10. The weight decay is set to 10^{-5} . The network is trained for 100 epochs with an initial learning rate 0.001 which is gradually decreased by a factor of 10 after every 10 epochs. The network is trained with Tensorflow on a desktop equipped with a NVIDIA Titan X GPU, and evaluated on an Intel CPU of 3.10GHz.

Baseline Methods We compare the proposed network against the following baselines:

¹The vertical line segments are not visible from O .

- The straight forward network for predicting depth directly (named as **Depth-Net**) instead of predicting depth-flow (Depth-Flow-Net) at the unknown pixels. Depth-Net is employed as baseline in this work.
- Low rank completion (**LR**) [20]: The missing depth values are computed by low rank matrix completion with low gradient regularization.²
- Highly sparse inpainting (**Semantic**) [14]: Region-based depth recovery for highly sparse depth maps.³ This method requires semantic labels of different objects present in the depth image. Fortunately, the datasets used in this work contain semantic labels which have been employed during the evaluation of this baseline. Note that none of the other methods including ours do not require semantic labels.
- PDE-based inpainting (**PDE**) [17]: Partial differential equation based anisotropic diffusion model for image inpainting is executed as baseline for RGB inpainting.
- Mumford-Shah inpainting (**MS**) [4]: The traditional image inpainting based on Mumford-Shah-Euler model is also evaluated as baseline.⁴

5.1. Depth image completion

SUNCG [18] is a large-scale synthetic dataset contains 45,622 depth images of different scenes with realistic rooms and furniture layouts. The NYU depth dataset [12] consists of 1,449 real depth images of indoor environment of commercial and residential buildings. The datasets also consists of semantic object labels which are not utilized in this work. In each epoch we select a batch of 2,000 depth image pairs (original and with occlusions) generated by our augmentation technique. More augmented images are generated by random cropping and flipping the images in the left/right direction.

Quantitative Evaluation The proposed network is evaluated for the task of generating new depth views. For this task a set of 100 testing image pairs of the same scene at different view-points and orientations is generated from SUNCG [18] datasets.⁵ One is considered as the depth at the source pose and the other considered as depth at the target pose. The depth image at the source pose is first warped at the target pose and then fed into the depth completion network Depth-Flow-Net. In Table 1, we display the mean and median depth prediction error evaluated only at the unknown pixels.

For NYU Depth v2 [12] datasets, no complete 3D model is available. Thus, we rely on the “dual warping” technique

²code is available at <https://github.com/xuehy/depthInpainting>

³code is available at <https://uk.mathworks.com/fileexchange/64546>

⁴code is available at <https://uk.mathworks.com/fileexchange/55326>

⁵code is available at <https://github.com/shurans/sscnet>

Table 1. Depth completion Comparison : SUNCG datasets [18]

	LR [20]	Semantic [14]	Ours
Mean error	0.34m	0.33m	0.28m
Median error	0.25m	0.23m	0.05m

Table 2. Depth completion Comparison : NYU datasets [12]

	LR [20]	Semantic [14]	Ours
Mean error	0.42m	0.37m	0.38m
Median error	0.26m	0.20m	0.06m

(strategy 1) to generate test data. In Table 2 we observe that existing depth inpainting algorithms [20] and [14] perform comparably, whereas the proposed depth prediction method improves the median error significantly. A detailed description of the runtime can also be found in Table 3. We observe Depth-Flow-Net is fastest among the depth completion benchmark methods. Note that all the methods including ours are evaluated on a CPU. The estimation of 2D displacements, depth completion, and generation of new view are included in the runtime.

Qualitative Evaluation The network is again used to evaluate for the task of generating new depth views. Separate sets of depth images are chosen as test sets. Each depth image of the test set is warped w.r.t. a randomly sampled target pose (with position and orientation variations from the range of $[-1m, 1m]$ and $[-15^\circ, 15^\circ]$). The proposed depth completion network is then employed for depth completion at the occluded regions. The novel view generation results are plotted in Fig. 8. Although, direct depth estimation network Depth-Net produces reasonable solutions in some cases, we observe that proposed displacement-field based Depth-Flow-Net produces more consistent solutions. The estimated displacements are displayed by green lines segments while target pixels are marked with blue dots.

In order to enhance the quality of completed depth maps, we utilize an ensemble-inspired framework: after warping the current depth maps to multiple nearby poses, the induced occlusions are completed using Depth-Flow-Net. The resulting depth maps are warped back to the original pose and are subsequently merged using a pixel-wise median filter (which we use as a efficient surrogate for a more refined approach for depth maps fusion such as [11]). Examples for such 3D scene completion can be found in Fig. 8(c) and Fig. 8(e). Note that signed distance functions (i.e. volumetric fusion) could be applied to obtain smoother surfaces. However, we leave this for future work. More results can be found in the supplementary material.

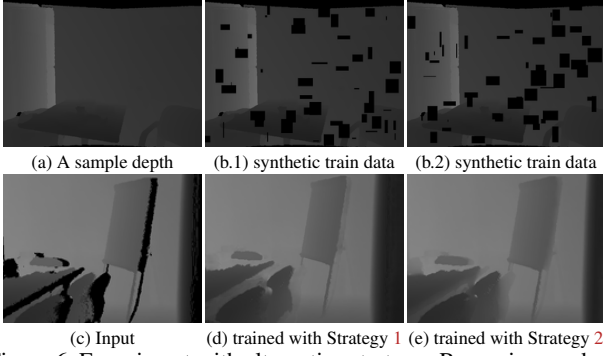


Figure 6. Experiment with alternative strategy: Removing random blocks. We observe more ghosting artifacts if the same network (Depth-Net) is trained with the alternative strategy compared to the “dual warping” strategy.

5.2. Validation of dual warping for dataset creation

To validate our synthetic dataset generation method (strategy 1), we conduct an experiment with a dataset generated by strategy 2. For each depth image the missing regions are generated by removing random regions of pixels. Up to 20% of all pixels are removed. The size of each removed region is chosen uniformly within the range $[1, 50]$ along both the directions. We train Depth-Net for both the datasets generated by strategy 1 and 2. Note that in the current experiment Depth-Net is chosen over Depth-Flow-Net to demonstrate that strategy 1 does not just favor a displacement-based network, but enhances the problem itself. The results are displayed in Fig. 6. We observe more accurate depth prediction with strategy 1. Thus the current experiment validates our argument of minimal domain shift of novel view generation with strategy 1.

5.3. Limitation / Failure cases

Despite the generally good performance of Depth-Flow-Net we have encountered failure cases, usually caused by the following:

- In the presence of large occlusions on foreground objects (e.g. Fig. 7(a)) the background depth may incorrectly spill over into the foreground object (Fig. 7(b,c)).
- Very large occlusions (or otherwise regions with missing depth, such as in Fig. 7(d)) can lead to displacement vectors that point themselves to missing data (Fig. 7(e,f)). In our current approach there is no guarantee that the displacement field always refers to valid depth.

Despite the above limitations, proposed Depth-Flow-Net produces satisfactory results in a wide variety of depth images. A number of examples are included in the supplementary material.

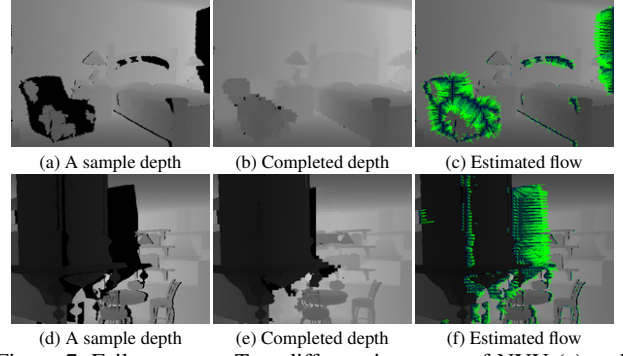


Figure 7. Failure cases: Two different instances of NYU (a) and SUNCG (d) datasets where proposed Depth-Flow-Net fails due to the limitations in the current approach. See text for more details.

Table 3. Runtime Comparison: evaluated on CPU

	MS [4]	PDE [17]	LR [20]	Semantic [14]	Ours
Runtime	114.8s	4.07s	73s	5.81s	0.7s

Table 4. Quantitative comparison of RGB image completion

	PDE [17]	MS [4]	Ours
PSNR	24.9dB	24.7dB	26.08dB

5.4. Novel RGBD image synthesis

We also exploit our image augmentation strategy for new view RGBD image synthesis given a single RGBD image. We utilize a similar augmentation (strategy 1) to generate the ground truth for RGBD image completion. A network similar to Depth-Flow-Net is trained on the augmented RGB-D datasets of (original-occluded) pairs. In contrast to depth estimation it takes 4D channels as input and estimate 2D displacements from source to target regions. Once the network is trained we warp the original view to the target views and complete the missing pixels using the predicted displacement field.

We conduct similar procedure as before for quantitative and qualitative evaluation. A quantitative comparison can be found in Table 4 and a qualitative comparison is displayed in Fig. 9. We observe an improvement of PSNR compared to the traditional in-painting algorithms. More results can be found in the supplementary document.

6. Evaluation on SUNCG [18] dataset

To evaluate the proposed method on SUNCG [18] datasets, we utilize similar evaluation dataset generation technique as NYU [12] datasets. We warp the current depth maps to multiple nearby poses and the induced occlusions are completed by proposed Depth-Flow-Net. The results are displayed in Fig. 10 and Fig. 11 along with the estimated depth flow. Note that the depth-flow is estimated at every pixels but in the figure we only show the flow at the

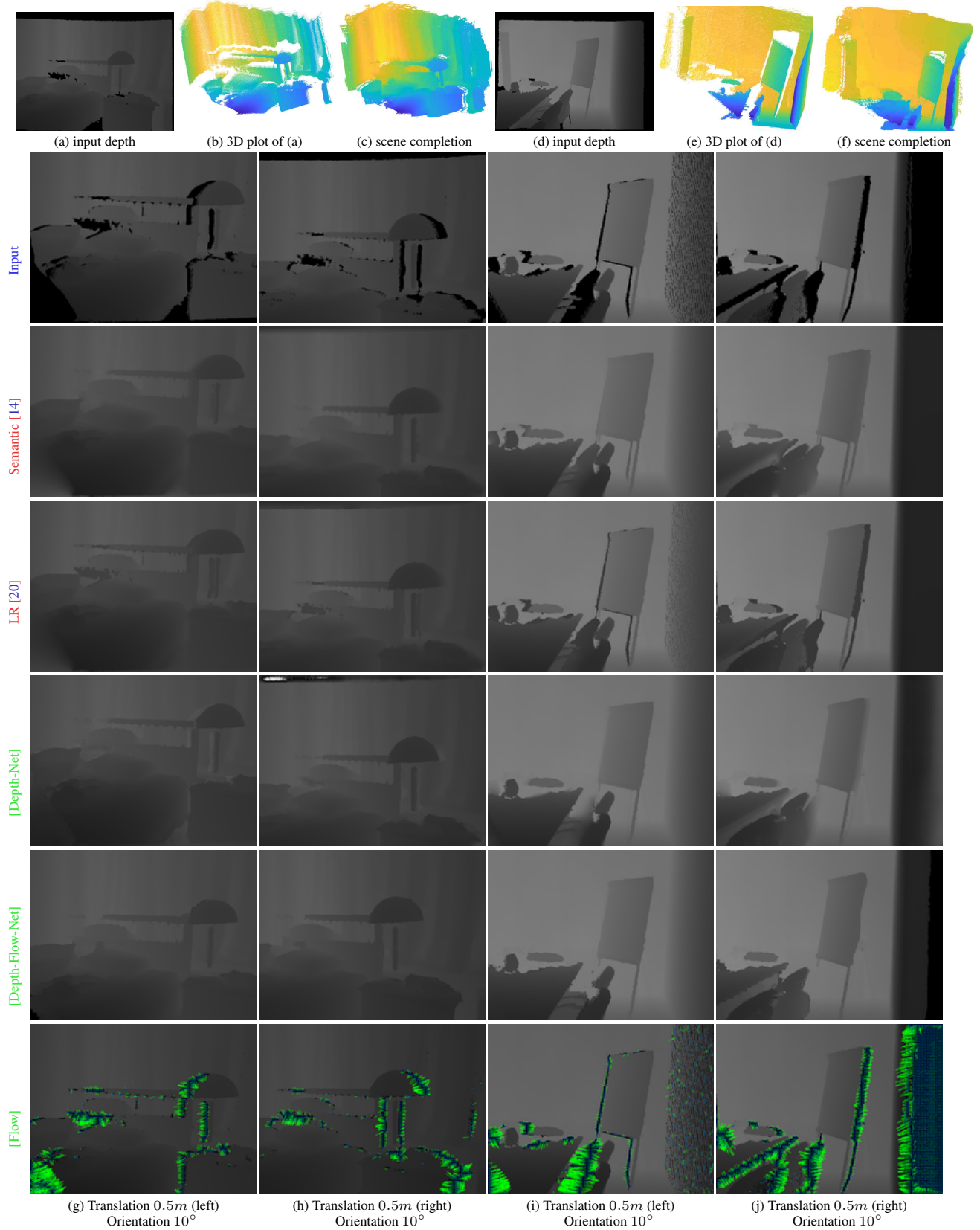


Figure 8. Qualitative results of depth completion methods at different viewing location and orientation on NYU Depth v2 [12] dataset. Images are first warped to the target pose and then use depth completion methods to predict depth at the occluded regions. **Depth-Flow-Net** produces less artifacts and can even hallucinate handles of the chairs. A complete video is shown in the supplementary material.

unknown pixels with a regular 4 pixel interval.

7. Conclusion

In this work we develop a technique to complete 3D scenes indirectly by filling occluded regions in warped depth (and optionally RGB) images. Hence, we are able to avoid a costly volumetric representation and consequently work in higher-resolution image-space. Our main contribution is the generation of training data via dual warping, which adds realistic occlusion patterns to given depth images. Therefore large amounts of training data are easy to acquire. We also perform a thorough evaluation to demonstrate the effectiveness of our weakly supervised approach and to show the efficiency of a proposed depth completion network. Further, the flexibility of the proposed method is emphasized by an evaluation on RGB-D data. Currently, the proposed method is limited to generating depth images of relatively nearby poses, which is a restriction addressed in future research.

References

- [1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proc. of CVPR*, pages 5828–5839, 2017.
- [2] Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans. *CVPR*, 2018.
- [3] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.
- [4] Selim Esedoglu and Jianhong Shen. Digital inpainting based on the mumford–shah–euler image model. *European Journal of Applied Mathematics*, 13(4):353–370, 2002.
- [5] John H Gennari, Pat Langley, and Doug Fisher. Models of incremental concept formation. *Artificial intelligence*, 40(1-3):11–61, 1989.
- [6] Christine Guillemot and Olivier Le Meur. Image inpainting: Overview and recent advances. *IEEE signal processing magazine*, 31(1):127–144, 2014.
- [7] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017.
- [8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.
- [9] Kevin Lai, Liefeng Bo, and Dieter Fox. Unsupervised feature learning for 3d scene labeling. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 3050–3057. IEEE, 2014.
- [10] DC Marr. A computational investigation into the human representation and processing of visual information. *Freeman, San Francisco, CA*, 1982.
- [11] Paul Merrell, Amir Akbarzadeh, Liang Wang, Philippos Mordohai, Jan-Michael Frahm, Ruigang Yang, David Nistér, and Marc Pollefeys. Real-time visibility-based fusion of depth maps. In *IEEE 11th International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007.
- [12] Derek Hoiem Pushmeet Kohli Nathan Silberman and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *Proc. of ECCV*, pages 746–760, 2012.
- [13] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 702–711. IEEE, 2017.
- [14] Said Pertuz and Joni Kamarainen. Region-based depth recovery for highly sparse depth maps. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 2074–2078. IEEE, 2017.
- [15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [16] Ling Shao, Jungong Han, Dong Xu, and Jamie Shotton. Computer vision for RGB-D sensors: Kinect and its applications [special issue intro.]. *IEEE transactions on cybernetics*, 43(5):1314–1317, 2013.
- [17] Jianhong Shen and Tony F Chan. Mathematical models for local nontexture inpaintings. *SIAM Journal on Applied Mathematics*, 62(3):1019–1043, 2002.
- [18] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proc. of CVPR*, pages 190–198, 2017.

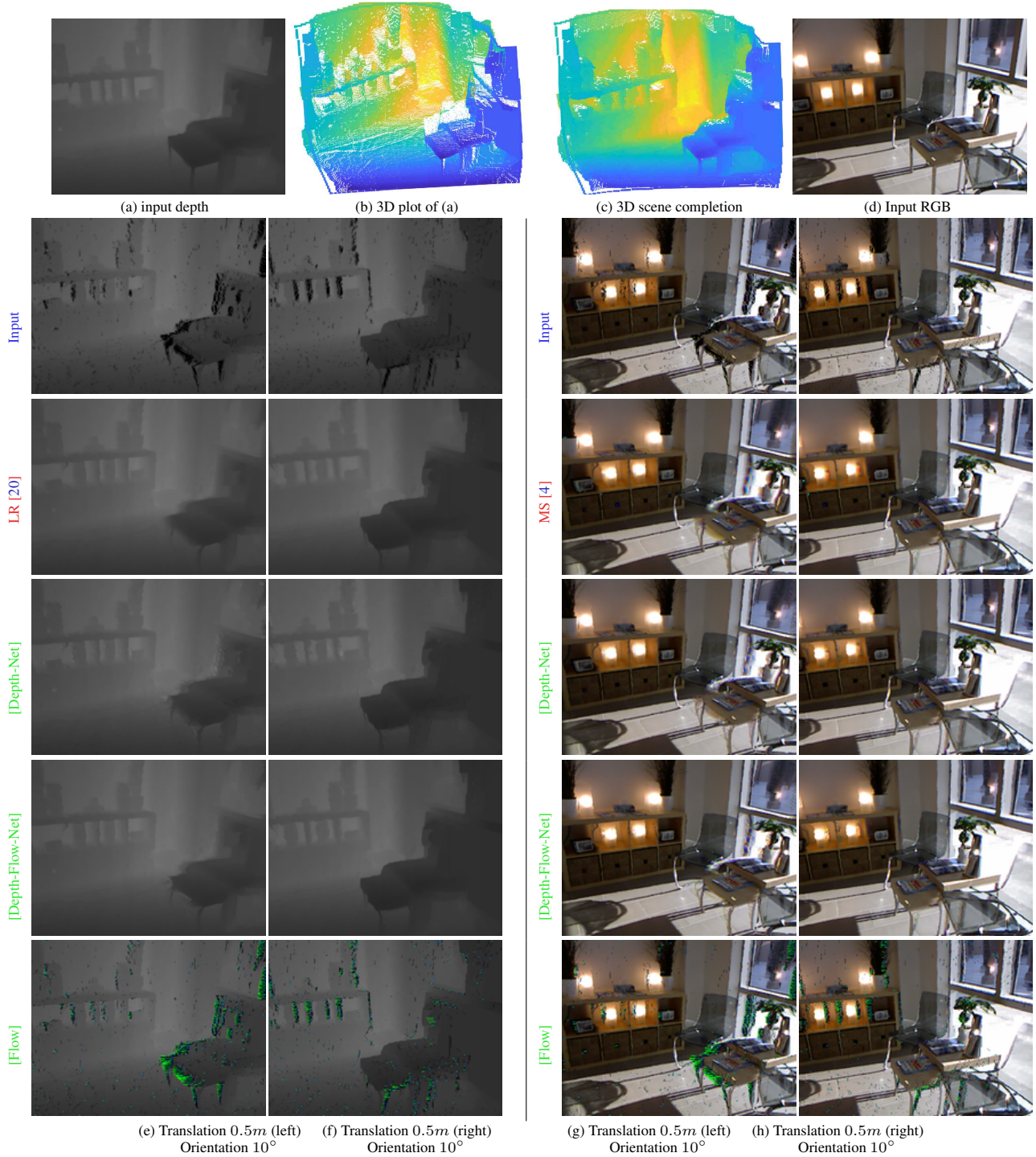


Figure 9. (a) Qualitative results of depth completion methods at different viewing location and orientation. The same network used for RGB image completion to obtain a new RGB image at the target pose. A complete video is shown in the supplementary material.

[19] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. In *Advances In Neural Information Processing Systems*, pages 540–550, 2017.

ances In Neural Information Processing Systems, pages 540–550, 2017.

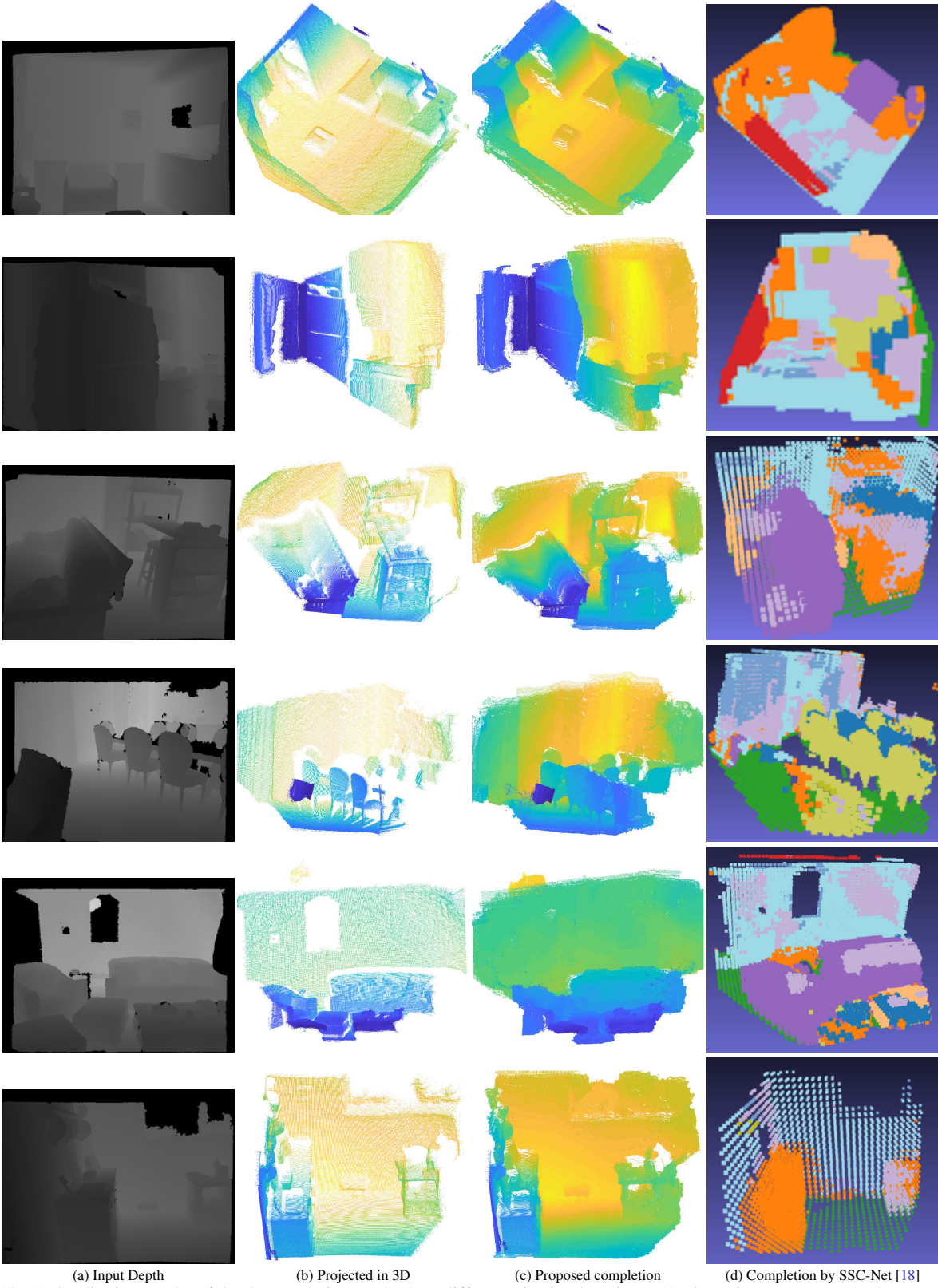


Figure 10. (a) Qualitative results of depth completion methods at different viewing location and orientation on NYU [12] datasets. Images are first warped to the target pose and then use proposed depth completion method to predict depth at the occluded regions and subsequently merged.

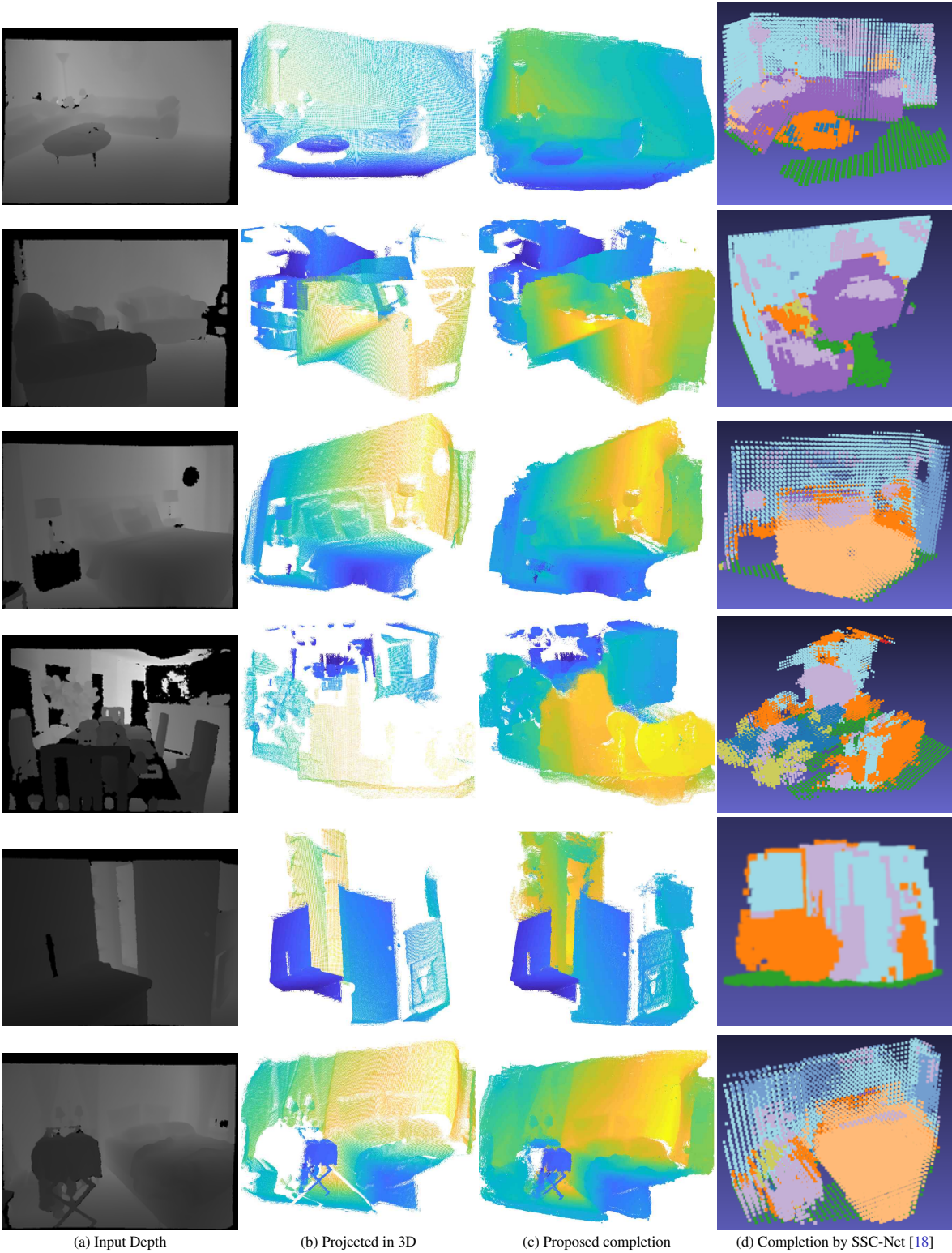


Figure 11. (a) Qualitative results of depth completion methods at different viewing location and orientation on NYU [12] datasets. Images are first warped to the target pose and then use proposed depth completion method to predict depth at the occluded regions and subsequently merged.

- [20] Hongyang Xue, Shengming Zhang, and Deng Cai. Depth image inpainting: Improving low rank matrix completion with low gradient regularization. *IEEE Transactions on Image Processing*, 26(9):4311–4320, 2017.
- [21] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. *arXiv preprint arXiv:1801.09392*, 2018.
- [22] John Zelek and Nolan Lunscher. Point cloud completion of foot shape from a single depth map for fit matching using deep learning view synthesis. In *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*, pages 2300–2305. IEEE, 2017.
- [23] Yinda Zhang and Thomas Funkhouser. Deep Depth Completion of a Single RGB-D Image. *arXiv preprint arXiv:1803.09326*, 2018.