Supplementary Material: Colorization of Depth Map via Disentanglement

Chung-Sheng Lai[1], Zunzhi You[2], Ching-Chun Huang[1],
Yi-Hsuan Tsai[3], Wei-Chen Chiu[1]

[1]National Chiao Tung University, Taiwan
[2]Sun Yat-sen University, China [3]NEC Labs America

1 Analysis on Model Learning

In order to know whether the structure sub-network S and appearance sub-network E can disentangle the structure and appearance information, we feed our network with pure red, green and blue image with depth value equal to 0.2 (input depth map is normalized) as the appearance reference RGB-Depth image-pairs for performing colorization. The results are shown in Fig. 1.

In the results, when the appearance reference is the pure color, since there is no appearance information, the image quality is less realistic. However, our model is still able to reconstruct the output that reflects the structure information of the depth input. When the appearance reference is the normal image, our model is able to reconstruct the corresponding appearance and also the structure according to the inputs. This demonstrates that our sub-networks can disentangle the factors for appearance and structure well.

2 Network Architecture

We adopt the same structure for both the sub-network S and the appearance sub-network E. There are four residual blocks with different spatial sizes. Except for the first residual block, the first convolutional layer of each blocks has the spatial size with stride equals to 2. The mixing sub-network M consists of 4 2-stride transpose convolutional layers, each followed by a normalization layer, a ReLU activation layer and 4 residual blocks. There are skip connections between structure sub-network S and mixing sub-network M at the same resolution. The details are shown in Fig. 2.

3 Implementation Details

When utilizing our time invariant property in the training process, we do not pick the frame that is only one frame away from the other. The reason is that the structures of consecutive frames are too similar to each other or even identical, which would not satisfy the time invariant property to perform disentanglement.

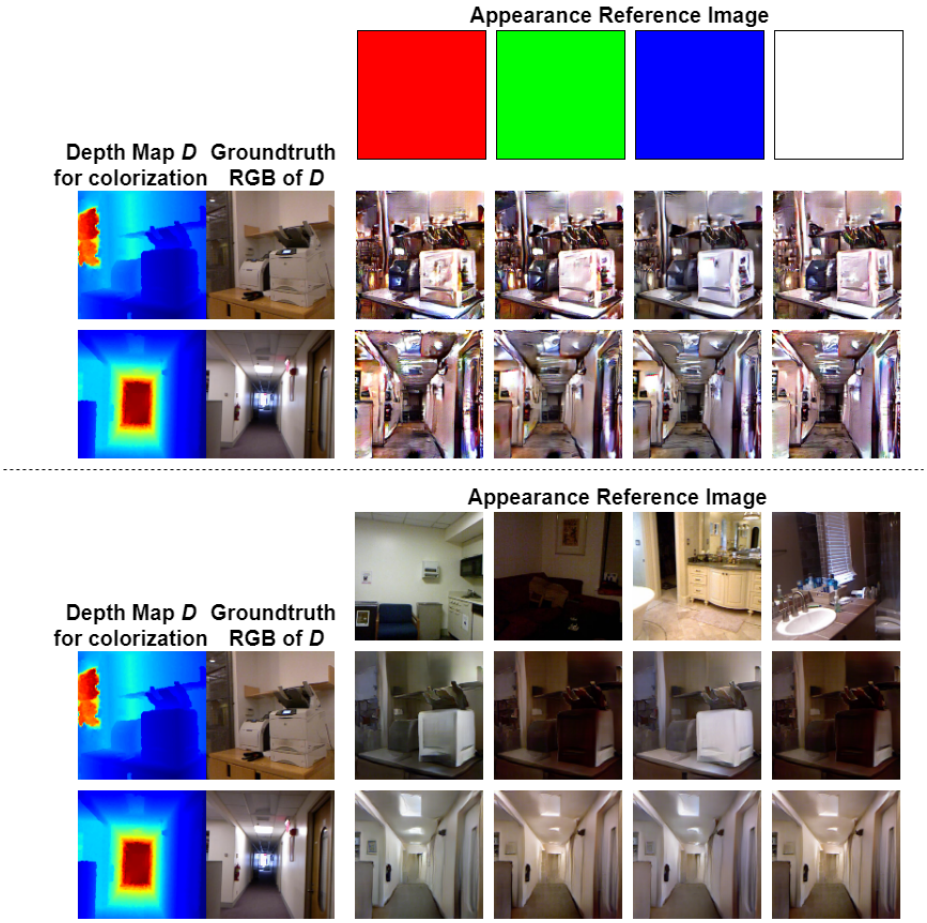

Fig. 1. The qualitative results of our proposed model using "solid color, single value depth map" or the "real RGB-Depth image pair" as the appearance reference.

In practice, we pick RGB-Depth pairs at timestamp T and $T+10$ for NYU-Depth V2, and timestamp T and $T+50$ for SceneNet RGB-D dataset.

We adopt two discriminators to train our full model when utilizing adversarial learning like Pix2Pix GAN [1]. Our model is trained to fool both discriminators with different output patch sizes. The discriminators match our output images from size 128×128 to patch 14×14 and 30×30 respectively. We only apply adversarial learning at the first output of DRIT-arch, since it is the only place that we do not have ground-truths of output images.

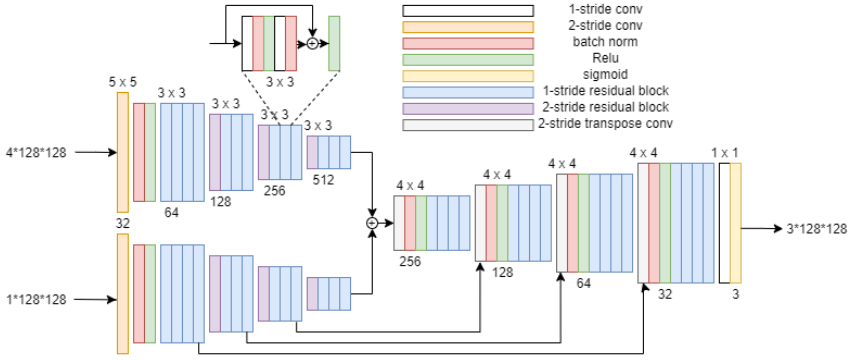

Fig. 2. The detailed structure of our model. There are four residual blocks with different spatial size (denoted as blue blocks). The first convolution layer of each block has the spatial size with stride equal to 2. In addition, each residual block has its own skip connection with the element-wise addition before the last activation function. The decoder outputs a colorized depth map with sigmoid as output activation function. The numbers above each blocks indicate the kernel size of each convolutional layers. The numbers underneath each blocks indicate the numbers of kernel filters.

4 More Qualitative Results

We show more example results for our colorized outputs in comparison with other image-to-image translation approaches in Fig. 3 to 5 and a supplementary video. In addition, we present qualitative results for the consistency of object detection in Fig. 6.

5 Modified Version of CycleGAN

As the architectures of pix2pix and CycleGAN do not support the appearance input, there could raise the concern of potentially having the unfair comparison with respect to them. Hence, we further modify the CycleGAN for making it learn the translation between RGB image and depth map concatenated with appearance reference. However, such modified CycleGAN turns out to totally ignore the depth map and just output the RGB image in appearance reference (in the direction of translating from depth map concatenated with appearance reference to RGB image), in which such results again clearly point out the problematic applications of pix2pix/CycleGAN on the task of depth map colorization.

6 Replacing Random Flipping with Random Cropping

As described in the main manuscript, the **random flipping** operation is applied on the reference RGB-Depth pair to alleviate the dependencies between the inputs for both appearance and structure sub-networks, thus encouraging

the disentanglement between the appearance and structure factors in our self-supervised learning scheme for colorization. Here we further study another possible choice for such operation, i.e. **random cropping**, to replace the random flipping. Basically, we randomly crop the reference RGB-Depth pair as the input for the appearance sub-network, in which it would have the different structure but ideally still the same appearance information with respect to the input depth map for the structure sub-network (i.e. the complete depth map). We experiment such random clipping operation on the model variant without using DRIT-arch and adversarial learning for simplicity, and get 11.9929 and 93.8137 on PSNR and FID respectively. In comparison with our original model design of adopting random flipping (12.7394 and 55.9795 in terms of PSNR and FID, cf. Table.2 in the main manuscript), the utilization of random clipping operation results in comparable reconstruction but slightly worse image fidelity and diversity. We believe that such degradation in performance is likely to be stemmed from the potentially overlapping area between the randomly-cropped reference RGB-Depth pair and the complete depth map, which would cause the leak of structure information from the appearance sub-network during the model training thus leading to inferior disentanglement and colorization. Meanwhile, this study clearly verifies the benefit brought by our current design choice of using random flipping operation in the proposed model.

7 Different Forms of Input Depth Map

In the current setting of our proposed method, all the input depth maps are min-max normalized by adopting the minimum and maximum depth values of the whole dataset as reference (denoted as *dataset-wise normalization* here). Here we additionally study two other possible forms of input depth maps: (1) the surface-normal maps directly derived from the corresponding depth maps; and (2) the depth maps min-max normalized by their own minimum and maximum depth values (denoted as *image-wise normalization*). Please note that these two forms are both scale-invariant with respect to the numerical range of the depth maps. We experiment these forms of input depth map based on the model variant without using DRIT-arch and adversarial learning for simplicity, where the quantitative and qualitative comparisons are provided in Table 1 and Fig. 7 respectively. We observe from the qualitative examples that both using surface-normal maps or the depth maps normalized image-wisely could sometimes lead to unsmooth/inconsistent colorization on the planar surfaces. Moreover, these two forms of input depth maps give slightly worse performance in terms of PSNR and FID, hence verifying the benefit for our current model design of adopting the dataset-wise normalization on the input depth maps.

8 Input for Appearance Sub-network E

The input of appearance sub-network E in our current model design is the appearance reference RGB-D pair $\{I^R, D^R\}$, as we assume that the appearance

Table 1. Quantitative evaluation on different forms of the input depth map (cf. Sec. 7).

Metrics	PSNR	FID
Surface Normal	12.6979	63.4613
Image-wise Normalization	12.3315	71.2023
Our Model (Dataset-wise Normalization)	**12.7394**	**55.7394**

information of an RGB image I^R should be obtainable by subtracting the structure information (provided by D^R) from it. Here we try to play with another design choice where the appearance sub-network E learns to directly extract the appearance information from the RGB image I^R only. Again, we perform the experiments based on the model variant without using DRIT-arch and adversarial learning for simplicity. In results, We get 12.7665 and 59.9241 on PSNR and FID respectively for using RGB image only as the input for the appearance sub-network E, which are actually comparable to our current model setting of taking RGB-D pair as input for E (cf. Table.2 in the main manuscript, 12.7394 and 55.9795 in terms of PSNR and FID). However, as shown by the qualitative examples in Fig. 8, we observe that using RGB image only as input for E could sometimes lead to fuzzier boundaries (in the upper example) or some artifacts (in the lower example). Therefore we still adopt the RGB-D pair $\{I^R, D^R\}$ as the appearance reference in our full model.

References

1. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 2

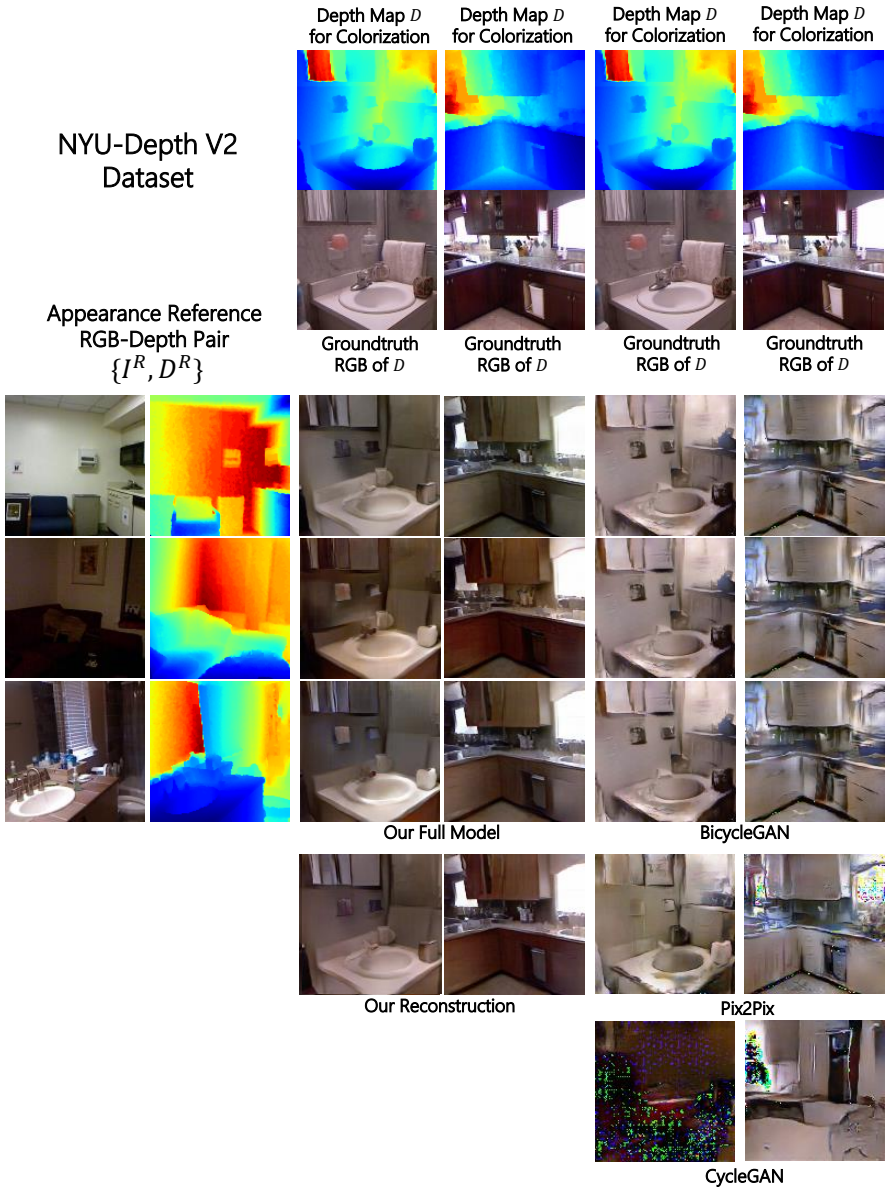

Fig. 3. The qualitative results of our proposed model and other image-to-image translation models on NYU-Depth V2.

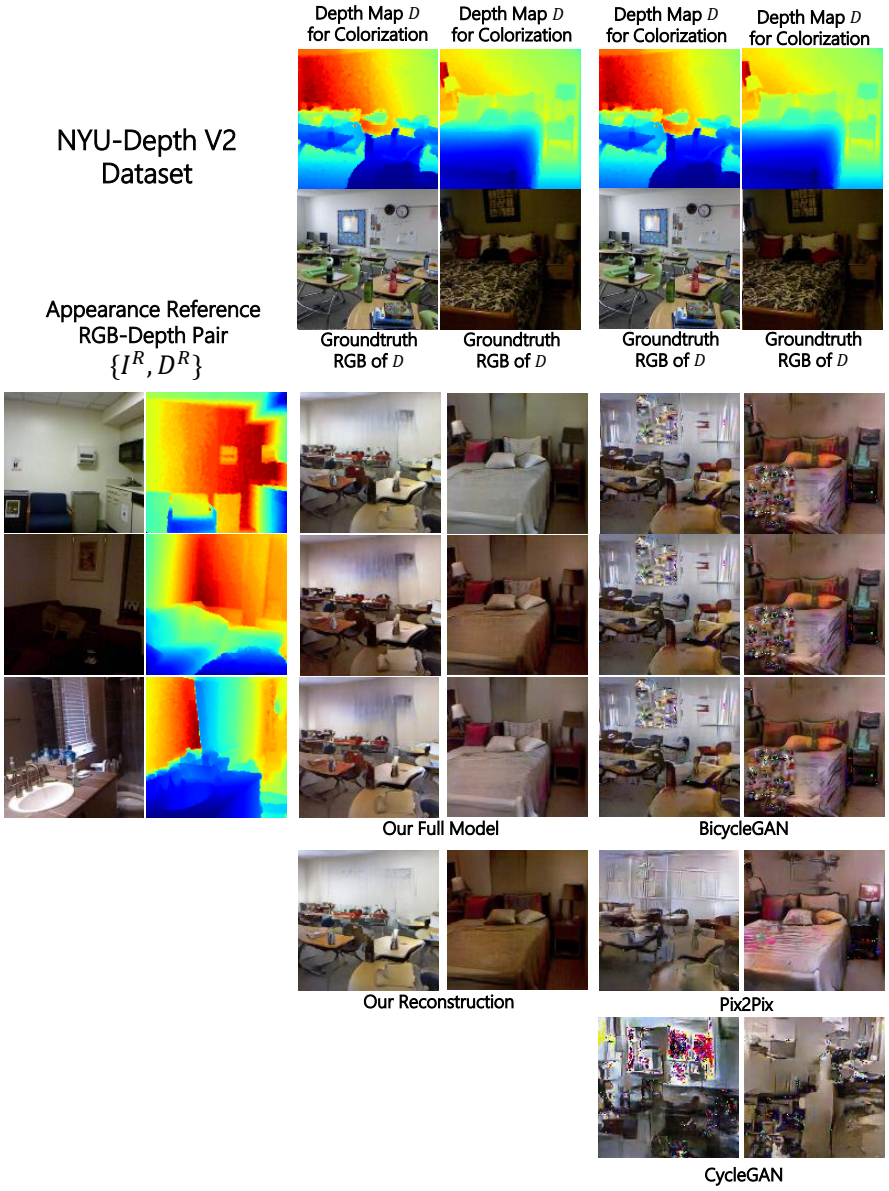

Fig. 4. The qualitative results of our proposed model and other image-to-image translation models on NYU-Depth V2.

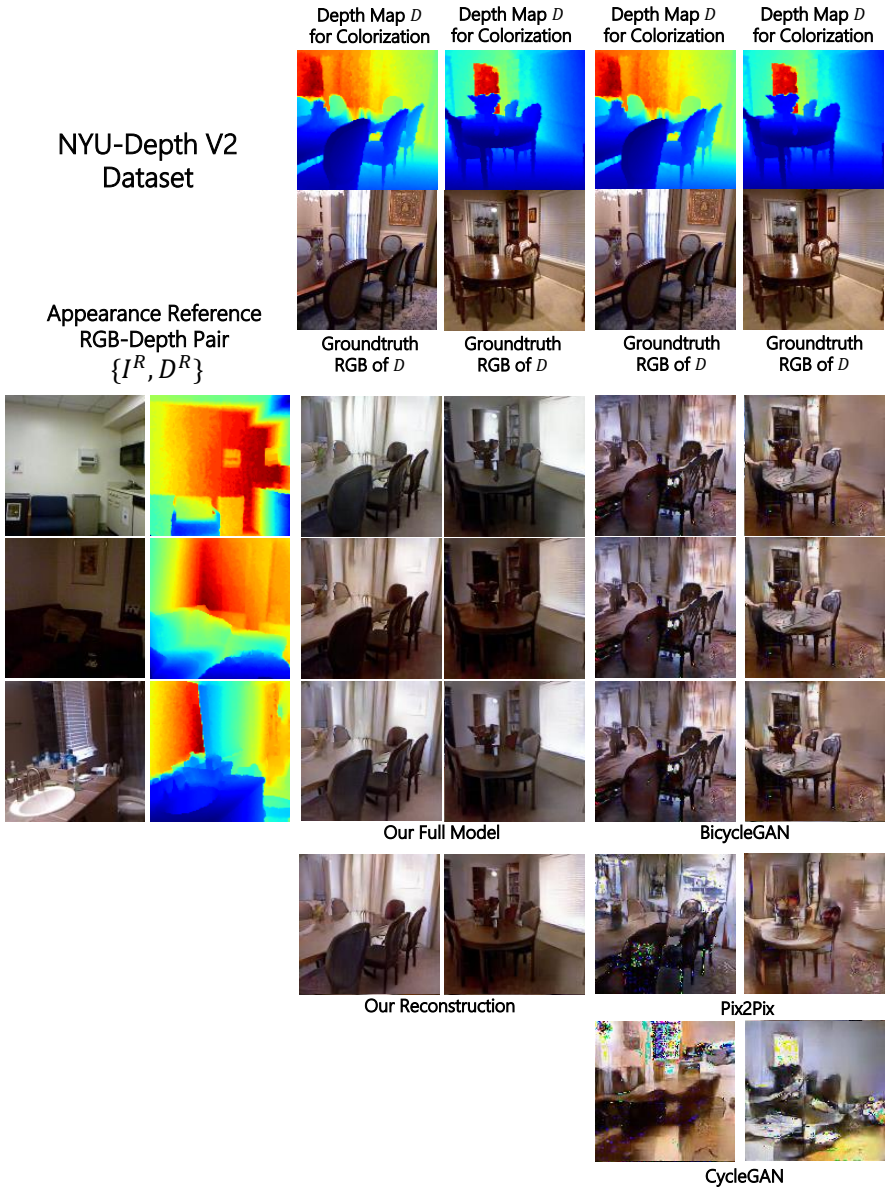

Fig. 5. The qualitative results of our proposed model and other image-to-image translation models on NYU-Depth V2.

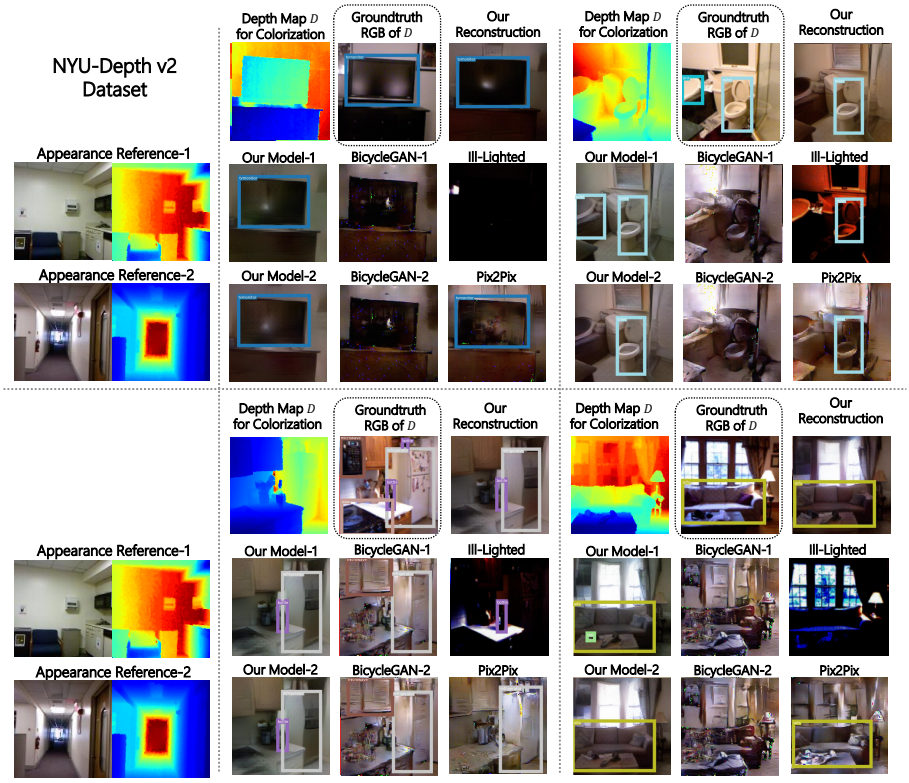

Fig. 6. Examples for verifying recognition consistency. For each example, the YOLOv3 detection outputs on the original RGB image, reconstruction made by our proposed model, our colorization, BicycleGAN results, Pix2Pix result, and ill-lighted image are provided for comparison. Our reconstruction and colorization show higher consistency with respect to the original image than the others in terms of object detection.

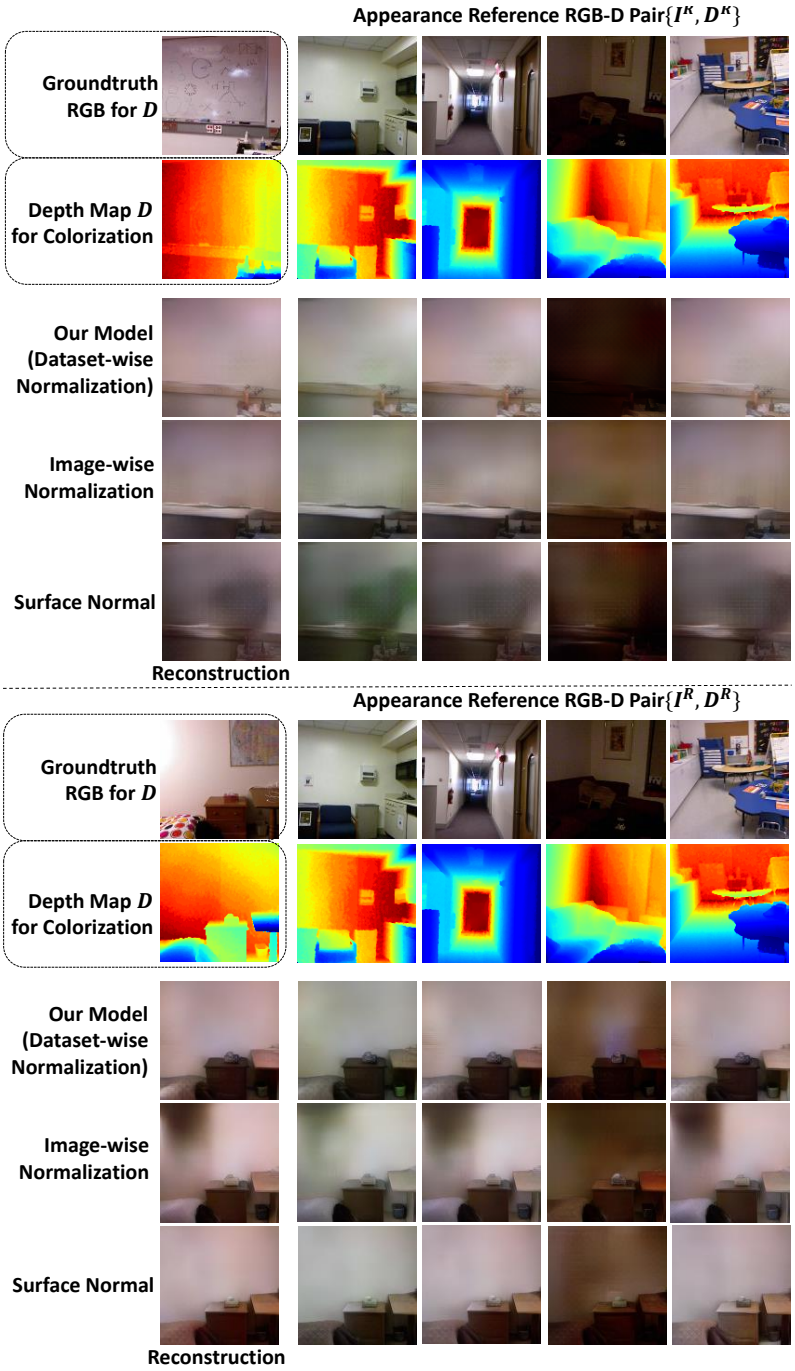

Fig. 7. Qualitative examples on having different forms for the input depth map (cf. Sec. 7).

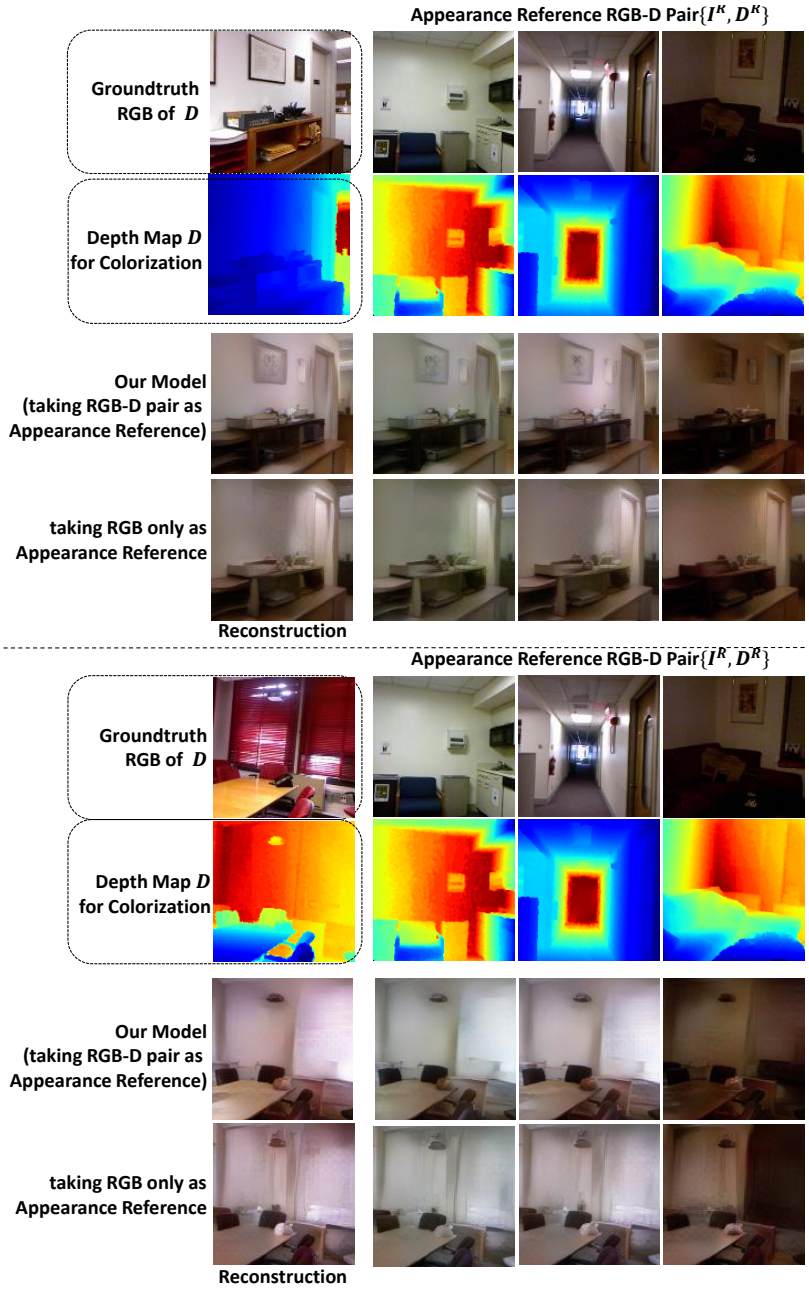

Fig. 8. Qualitative examples of adopting different inputs for appearance sub-network (cf. Sec. 8).