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

Abstract. Vision perception is one of the most important components for a computer or robot to understand the surrounding scene and achieve autonomous applications. However, most of the vision models are based on the RGB sensors, which in general are vulnerable to the insufficient lighting condition. In contrast, the depth camera, another widely-used visual sensor, is capable of perceiving 3D information and being more robust to the lack of illumination, but unable to obtain appearance details of the surrounding environment compared to RGB cameras. To make RGB-based vision models workable for the low-lighting scenario, prior methods focus on learning the colorization on depth maps captured by depth cameras, such that the vision models can still achieve reasonable performance on colorized depth maps. However, the colorization produced in this manner is usually unrealistic and constrained to the specific vision model, thus being hard to generalize for other tasks to use. In this paper, we propose a depth map colorization method via disentangling appearance and structure factors, so that our model could 1) learn depth-invariant appearance features from an appearance reference and 2) generate colorized images by combining a given depth map and the appearance feature obtained from any reference. We conduct extensive experiments to show that our colorization results are more realistic and diverse in comparison to several image translation baselines.

Keywords: Depth Colorization, Disentanglement, Image Translation

1 Introduction

Recognizing the surrounding objects or the environment based on the visual sensory is one of the fundamental topics in computer vision. Most of the existing computer vision algorithms, including object recognition, simultaneous localization and mapping (SLAM) for robot navigation and position, or the ones used in autonomous driving, are applied on the images or videos taken by RGB cameras. Under the condition of having sufficient lighting environment, appearance details of objects can be well captured in RGB images, in which the vision models are trained to perform recognition on such images. As the RGB image is formed by recording the scattered light coming from the illuminated surface of the

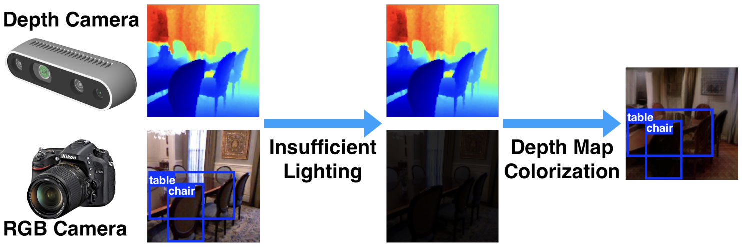

Fig. 1. RGB and Depth cameras are two most popular visual sensors nowadays, while most of the typical computer vision models (e.g., object detection) are trained upon the data acquired by RGB cameras. However, if there is no sufficient illumination in the environment, the RGB camera usually cannot well capture image details, thus leading to inaccurate recognition. In comparison, the active sensing of depth cameras is able to function well under such low-lighting situation to produce depth maps. In this paper we aim to utilize this advantage of depth cameras and colorize the depth map, such that the vision models would still be able to perform their tasks.

surrounding environment into the camera, appearance details shown in the image would gradually diminish when the environmental illumination becomes lower, which leads to undesirable performance of recognition for vision models. The case of having insufficient lighting is actually quite common in our daily life, e.g., indoor navigation or surveillance in a dark room, autonomous driving in the evening, or cave exploration by a robot. How to maintain the ability of visual perception in such cases is an important topic for the research community.

Depth camera is another popular visual sensor for perceiving the depth information and it is nowadays equipped to various robots and autonomous vehicles, where the rough structure/shape of the surrounding objects is well preserved in the resultant depth maps. Depth camera can still function smoothly in the low-lighting environment via its active sensing, e.g., based on the (infrared) laser design. However, although depth camera is able to provide more robust sensory ability against different illumination conditions, it is incapable of capturing appearance details (e.g., color and texture) of the objects as the RGB cameras do. Thus, the complementary property of RGB and depth cameras has attracted wide research interests [21,23,14] to have them integrated together for achieving better performance in visual perception and recognition.

Nevertheless, the combination between RGB and depth cameras does not guarantee to fully resolve the challenge for recognizing objects in the low-lighting environment. Therefore, several works [3,2,1,5,19] propose to tackle this problem from another perspective via performing the colorization on depth maps, where the colorized images are used as input for computer vision models to perform their recognition tasks. We observe that these works are either based on the hand-crafted colorization approaches or aiming to find the specific colorization manner in order to boost the performance of a certain computer vision model, and thus the colorized results are usually unrealistic and hard to be used for different tasks. In this paper, we instead propose to focus on the problem of learning depth map colorization without being constrained on any specific applications and target for generating realistic colorization results. In particular, the resultant colorization produced by our model is expected to paint the given depth map with (photo-)realistic textures and still maintain the overall structure/shape of the objects, such that the RGB-based vision models can be easily adapted to the colorized depth maps with less efforts.

We tackle the depth map colorization problem based on a hypothesis: an RGB image is composed of the structure factor and the appearance factor, where the former can be well captured by the corresponding depth map. This hypothesis of the disentanglement for RGB images is realized by our proposed model, which has three main components: structure sub-network, appearance sub-network, and a mixing sub-network (see Fig. 2). Given a depth map that we would like to colorize and a reference RGB-Depth image-pair as the source of appearance, the mixing sub-network takes 1) the structure factor extracted by the structure sub-network from the given depth map and 2) the appearance factor extracted by the appearance sub-network from the reference image-pair, as the input and then outputs the colorized depth map.

Based on our proposed model, a depth map can be colorized into different appearances by utilizing various reference RGB-Depth image-pairs. Each of our designed model is learned to obtain its function, i.e, extraction of structure/appearance factors for structure/appearance sub-networks and image generation for mixing sub-network. In addition, we apply several designs to improve our training procedure, such as the random flipping of reference image-pair and the time-invariant property of a video sequence. Experiments are conducted on the NYU-Depth v2 [20] and the SceneNet RGB-D [17] datasets. We provide the quantitative and qualitative evaluation on both the quality and diversity of the colorized depth maps, and demonstrate the efficacy of our method on maintaining performance for RGB-based computer vision models, in comparison to several baselines.

2 Related Work

Depth Colorization Previous works [3,1,2,5,19] on depth colorization mainly focus on how to transfer the depth maps into the format compatible with RGB images, such that the computer vision models which are primitively learned or designed for other data domains (e.g., RGB images) can still be adopted. For instance, Eitel *et al.* [3] propose a hand-crafted way to map the normalized depth values into RGB color channels (i.e., from highest depth values to lowest ones, they are gradually mapped into red, green, and blue colors). In [1], since a depth map contains rich 3D information, they instead propose to convert a depth image into a map of 3D surface normal, where the magnitude along each axis of a normal vector is encoded into RGB color channels respectively. Although the colorization obtained by these two methods do have the RGB colors, they are dissimilar to the typical images taken by regular RGB cameras, which may not be utilizable by typical RGB-based computer vision models. More recently, the work of [2] tackles the depth map colorization problem from the perspective of transfer learning, where a deep network is learned to find the optimal transformation from depth maps into RGB images, with respect to a given pre-trained Conv-Net. Since the given Conv-Net is pre-trained for a specific task, the learned transform and the resultant colorized depth maps are actually not realistic and thus less generalizable for direct usage by other models. Our goal in this paper is distinct from aforementioned approaches since we aim to generate the (photo-)realistic colorization on depth maps and our model is not designed specifically for any particular pre-trained models. There are also other approaches in performing colorization on the gray-scale images [7,11]. However, their input gray-scale images are the monochrome photos which already have quite some appearance details. Hence these gray-scale photos are fundamentally different from our target depth maps in this paper, which only represent the rough structure/shape of objects in the surrounding environment.

Image-to-Image Translation Another way to tackle the depth map colorization problem is to treat it as a special case of image-to-image translation task, where we take RGB images and depth maps as two data domains, and learn the translation between them. Image-to-image translation methods have been developed widely. For instance, Isola *et al.* [8] leverage the conditional generative adversarial network (GAN) [4] for learning the translation from one domain to another (with taking one domain as condition), where their method needs the paired data across domains for training. Zhu *et al.* [24] utilize the cycle consistency for learning the translation networks between two image domains, without requiring paired data. However, those methods are only able to produce one-to-one translation, i.e., the mapping between domains is deterministic, and thus the translated outputs lack diversity. Instead, the work of [25] extends the image-to-image translation from one-to-one mapping into one-to-many. With taking the data from one domain as condition, their method learns to model a distribution of plausible translated outputs for another domain based on the conditional generative modeling. However, as which will be shown later in our experiments, applying such models in the task of depth map colorization could suffer from the issue of mode collapse. Moreover, the resultant images may not be sufficiently realistic nor maintain the structure as in the input depth map. In comparison, our method is based on learning the disentanglement of RGB images and can generate realistic colorization with high image quality and diversity.

3 Proposed Method for Depth Map Colorization

As motivated in the introduction, the objective of our proposed method is to colorize a given depth map D by taking the appearance information from a reference of RGB-Depth image-pair $\{I^R, D^R\}$. The architecture of the proposed model is illustrated in Fig. 2, consisting of three sub-networks: structure sub-network S, appearance sub-network E, and mixing sub-network M. In the following, we will describe how we achieve the depth map colorization in details.

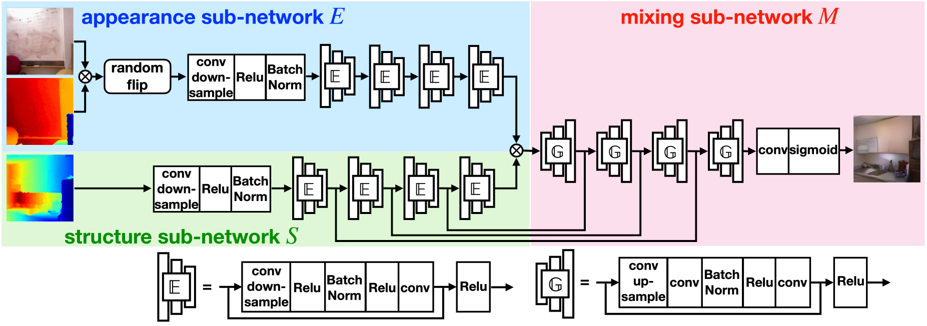

Fig. 2. Overview of the proposed model, which is composed of three main components: structure sub-network, appearance sub-network, and a mixing sub-network (shaded in green, blue, and red background respectively).

3.1 Disentanglement via Self-supervised Learning

The main assumption behind our model design is that: an RGB image I can be disentangled into the structure and appearance components, where they should be independent from each other. In particular, we hypothesize that the structure information has been fully maintained in the corresponding depth map D. Thus, the structure sub-network S is designed to extract the structure factor $v_S = S(D)$ from the input depth map D. In addition, while following the assumption above, the appearance information of an RGB image should be obtainable by subtracting the structure information from it. Hence, the appearance sub-network E takes an RGB-Depth image-pair $\{I^R, D^R\}$ as input, and then learns to extract the structure-invariant appearance factor $v_E = E(I^R, D^R)$ from I^R. Upon having both v_S and v_E, the mixing sub-network M combines them and produces the colorization result, which ideally should be a (photo-)realistic RGB image with its structure and appearance similar/related to D and I^R respectively.

Self-Supervised Learning. Our task is to colorize a given depth map D by using the appearance reference from any arbitrary RGB-Depth image-pair $\{I^R, D^R\}$. Since D and $\{I^R, D^R\}$ are unnecessary a pair that belongs to the same scene, there is no dataset under such setting that we can directly use to supervise our models. Moreover, it is impossible to collect a dataset with proper ground truths, i.e., finding multiple real-world images related to the same depth map having different appearance and the corresponding appearance references.

To address the problem of having no proper dataset, we propose a *self-supervised learning* scheme. Basically, we use the RGB-Depth image-pair $\{I^R, D^R\}$ and its depth map D^R as the input for the appearance sub-network E and structure sub-network S respectively. Then the resultant colorization produced by the mixing sub-network M should be able to well reconstruct an RGB image $\hat{I}^R = M(E(I^R, D^R), S(D^R))$. Nevertheless, directly using the objective defined on such reconstruction for training our model could be problematic.

The main reason is as follows. The primary motivation behind our model design is to make the appearance sub-network extract only the structure-invariant appearance information from the reference RGB-Depth image-pair, i.e., to achieve the disentanglement between the structure factor v_S and the appearance factor v_E, such that we are able to flexibly colorize a depth map D with any arbitrary appearance reference for producing diverse colorization results with the well-maintained structure of D. However, regarding our self-supervised learning scheme, since the input D^R for the structure sub-network S is used again in the input pair $\{I^R, D^R\}$ for the appearance sub-network E, the mixing sub-network M could have a trivial solution via learning to ignore $S(D^R)$, as M already receives all the information from E to reconstruct $\hat{I}^R$ to be similar to I^R. In other words, there is no guarantee to achieve the disentanglement between $v_S = S(D^R)$ and $v_E = E(I^R, D^R)$ solely via using the aforementioned reconstruction.

Random Flipping. In order to resolve such issue and still keep the benefit of self-supervised learning, we introduce a *random flipping* step to randomly flip the reference RGB-Depth pair $\{I^R, D^R\}$ before passing it to the appearance sub-network. Such random flipping operation F helps to alleviate the dependency between the inputs for both appearance and structure sub-networks, i.e., $\{F(I^R), F(D^R)\}$ and D^R respectively (note that I^R and D^R are under the same flipping). Therefore, the mixing sub-network is encouraged to jointly consider $v_S = S(D^R)$ and $v'_E = E(F(I^R), F(D^R))$ for achieving the reconstruction of I^R. In particular, the appearance factor v'_E extracted by E is enhanced to be structure-invariant, and our colorization model is encouraged to acquire the structural information mainly from the structure sub-network S, as the input D^R for S and the colorization output $\hat{I}^R = M(v'_E, v_S)$ should be consistent in structure even when the reference RGB-Depth pair $\{I^R, D^R\}$ is flipped. We define an objective to calculate the L1, L2, and the perceptual errors [9]:

$$\mathcal{L}_r = \left\| I, \hat{I} \right\|_1 + \left\| I, \hat{I} \right\|_2 + \sum_l \left\| \phi_l(I), \phi_l(\hat{I}) \right\|_2, \tag{1}$$

where I and $\hat{I}$ are input and reconstructed images respectively, and ϕ_l denotes the feature representation obtained from the l-th layer of an ImageNet-pretrained VGG network using `relu1_1`, `relu2_1`, `relu3_1`, and `relu4_1` layers. The objective based on our self-supervised scheme and the random flipping step considers (1) in reconstruction:

$$\mathcal{L}_{rec} = \mathcal{L}_r(I^R, M(v'_E, v_S)). \tag{2}$$

As shown in Fig. 2, both appearance sub-network E and structure sub-network S share the similar network architecture. The only difference is that S has the skip-connections to the mixing sub-network M over multiple convolution layers, serving a purpose to help the mixing sub-network preserve the structure details of the input depth map D^R.

Time Invariant Property (TIP). To further improve the self-supervised signal to deal with the lack of ground truths to train our colorization network, we

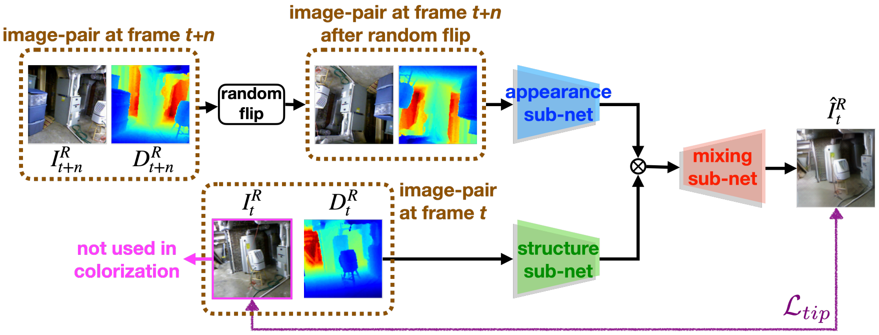

Fig. 3. Illustration of using Time Invariant Property (TIP) to provide additional supervised signals for our model training.

propose another design to facilitate the model training, named as *Time Invariant Property (TIP)*. The basic idea is to leverage the characteristics of the RGB-D video: We assume that the consecutive frames in an RGB-D video sequence share similar or even the identical appearances, but only have difference in the structure which is related to their depth maps. For instance, the RGB-D video sequences used for our experiments are taken in indoor scene by a moving camera, where the textures/appearances between consecutive frames are similar to each other.

This assumption then provides additional supervision to train the colorization model. One training scheme is illustrated in Fig. 3. Given an RGB-D video sequence, we use the depth map at time stamp t as the input to the structure sub-network S, and its neighboring RGB-Depth image-pair at time stamp $t+n$ as the appearance reference, to perform the colorization. Since the appearance features among neighboring frames are similar, we treat the RGB image at time stamp t as the ground truth of colorization for network training. However, there could exist a potential concern where the depth map structure of the appearance reference could be similar to the target depth map, which may break the disentanglement assumption. Fortunately, the proposed random flipping operation can well decorrelate these two depth maps and ensure our time invariant property. The reconstruction loss with TIP is similar to (2):

$$\mathcal{L}_{tip} = \mathcal{L}_r(I^R_t, M(v'_{E,t+n}, v_{S,t})), \tag{3}$$

where $v'_{E,t+n}$ denotes the extracted appearance factor from the reference pair I^R_{t+n}, D^R_{t+n} at time stamp t+n, $v_{S,t}$ denotes the extracted structure factor for the input depth map D^R_t at time stamp t, and I^R_t denotes the image at time stamp t.

3.2 Adversarial Learning and Cycle Consistency

To further improve the robustness of our proposed model (denoted as "Full Model" in the following sections), we introduce two additional training techniques, as shown in Fig. 4. First, we utilize the adversarial learning approach [4] via

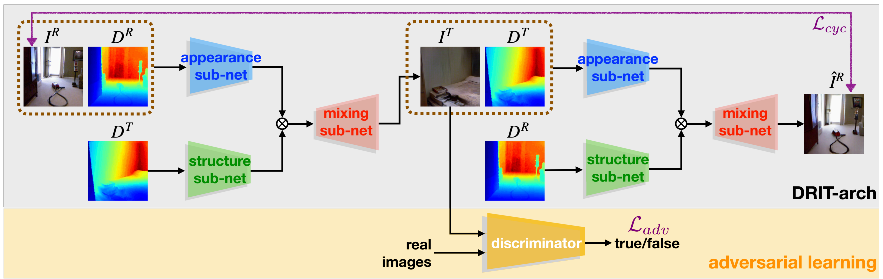

Fig. 4. Illustration of the DRIT-arch and adversarial learning used for improving our model training (cf. Section 3.2).

utilizing a discriminator on the colorization output to make the results more realistic. Second, as inspired by a recent work (i.e., DRIT [12,13]) on learning disentanglement for improving image-to-image translation, we apply a similar idea (denoted as *DRIT-arch*) to enforce cycle consistency and stabilize our network training. To better explain the overall procedure of DRIT-arch as illustrated in Fig. 4, we denote the reference RGB-Depth image-pair as $\{I^R, D^R\}$, and name the input depth map for colorization as D^T. Note that D^T is unrelated to the reference $\{I^R, D^R\}$. Then our network first produces the colorized output $I^T = M(E(I^R, D^R), S(D^T))$, which should have the appearance factor from $\{I^R, D^R\}$ and a similar structure to D^T. Then, we pair I^T and D^T as a new source of appearance reference and use it to colorize D^R, where the resultant colorization should well reconstruct I^R:

$$\mathcal{L}_{cyc} = \mathcal{L}_r(I^R, M(E(I^T, D^T), S(D^R))). \tag{4}$$

This cycle consistency provides us another supervision to train the network. It is worth mentioning that the cycle consistency can be applied without relying on the TIP assumption. Since there is no ground truth for I^T, we adopt an adversarial loss $\mathcal{L}_{adv}$ [16] to make the colorized output similar to real images.

Overall Objective. The overall objective for our model training involves the above-mentioned self-supervised loss via random flipping and time invariant property in (3), the cycle-consistency loss in (4), and the adversarial loss $\mathcal{L}_{adv}$:

$$\mathcal{L}_{all} = \mathcal{L}_{tip} + \mathcal{L}_{cyc} + \lambda_{adv}\mathcal{L}_{adv}, \tag{5}$$

where λ_{adv} serves to balance the loss function, which is set as 0.001 in this work.

Implementation Details. We follow the standard training scheme of GAN [4] to optimize the objective in (5), using the Adam optimizer [10] with a fixed learning rate of 0.001. First, we train our model from scratch only using time invariant property with random flipping via (3) for 100 epochs as a warm-up stage. After the model is more stable and able to produce reasonable results, we adopt the full loss function via (5) to make outputs more realistic and encourage our model to keep the appearance factor along with output images. Furthermore, we use PatchGANs [8] and least-squares objective [16] for stable training. More details are provided in the supplementary material about the network architecture and training procedure. Our project page is at https://github.com/alanlai199/ColorizeDepthNet

4 Experimental Results

Dataset. We adopt two datasets in our experiments: NYU Depth v2 [20] and SceneNet RGB-D [17]. The NYU Depth v2 dataset is composed of a collection of RGB-D video sequences and is originally proposed for learning indoor scene segmentation task. Here we use all the 284 raw video clips in the training set as we would like to leverage the time invariant property within the video sequences, while all the 654 RGB-D images in its test set (originally for evaluating segmentation) are used for testing. The SceneNet RGB-D dataset has a large scale collection of synthetic RGB-D videos, which are with photo-realistic quality in rendering. Here we randomly select 50,728 short video clips and 493 RGB-D images of different room-layouts from its training and test sets for our model learning and testing respectively.

4.1 Evaluation Metrics and Baselines for Comparison

There are two different quantitative evaluation schemes for the colorization results in our experiments. For the first evaluation scheme, we aim to quantify the performance of reconstruction, i.e., colorization on a depth map by using its corresponding ground truth RGB image. Here we adopt the well-know PSNR (peak signal-to-noise ratio, higher the better) metric for the assessment on the reconstructed image with respect to the ground truth RGB image.

For the second scheme, we target to evaluate the image quality and the diversity of the colorized depth maps. Here we adopt Fréchet Inception distance (FID [6], lower the better), which is commonly used in GAN-related works, as our metric. FID basically compares the similarity between two sets of data based on the distance between their distributions in the space of Inception feature representation [22]. Regarding both NYU Depth v2 and SceneNet RGB-D datasets, we choose 5 distinct RGB-Depth image-pairs from each of their test sets as our appearance references, and perform colorization on all the testing depth maps of each dataset. The colorization outputs are then compared with the real RGB images in the testing set, by using FID scores.

Three models of image-to-image translation are used as our baselines for comparison, including CycleGAN [24], Pix2Pix [8], and BicycleGAN [25]. We take RGB images and depth maps as two different data domains for training the baselines. Both CycleGAN and Pix2Pix can only produce one-to-one mapping, which means they can only generate one RGB output image for each input depth map. BicycleGAN acts more similar to our model; ideally it is able to colorize a given depth map into various appearances, where the appearance feature could

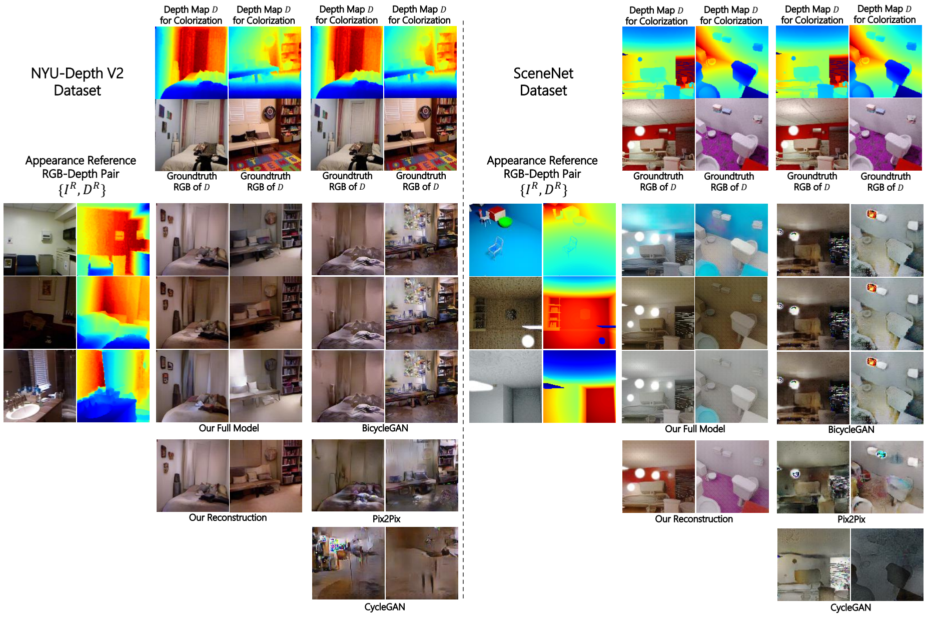
NYU-Depth V2 Dataset
Depth Map D for Colorization
Depth Map D for Colorization
Depth Map D for Colorization
Depth Map D for Colorization
Appearance Reference RGB-Depth Pair {I^R, D^R}
Groundtruth RGB of D
Groundtruth RGB of D
Groundtruth RGB of D
Groundtruth RGB of D
Our Full Model
BicycleGAN
Our Reconstruction
Pix2Pix
CycleGAN
SceneNet Dataset
Depth Map D for Colorization
Depth Map D for Colorization
Depth Map D for Colorization
Depth Map D for Colorization
Appearance Reference RGB-Depth Pair {I^R, D^R}
Groundtruth RGB of D
Groundtruth RGB of D
Groundtruth RGB of D
Groundtruth RGB of D
Our Full Model
BicycleGAN
Our Reconstruction
Pix2Pix
CycleGAN

Fig. 5. Qualitative results of our framework and image-to-image translation baselines.

be extracted from a reference RGB image, which is then combined with a given depth map as the input to achieve colorization. Note that, the generators in all baseline models have similar architecture as the one for BicycleGAN, in which the capacity of the generator is larger than our proposed model, i.e., 4×10^7 v.s. 1.5×10^7 (ours) in terms of the number of parameters.

4.2 Quantitative and Qualitative Results

In Fig. 5 and Table 1, we show the quantitative and qualitative evaluations respectively for our proposed model and the image-to-image translation baselines. For both evaluation schemes described previously, we observe that our full model performs favorably on colorization in comparison to baselines, where the resultant images have clearer edges, realistic appearance, and larger variety/diversity. In particular, our colorization results demonstrate the flexibility of our proposed method for adding different appearances into the same depth map via learning disentanglement, while the BicycleGAN baseline suffers from the mode collapse problem (i.e., produces the same colorization result no matter which appearance reference is given). As both the Pix2Pix and CycleGAN can only produce one-to-one mapping, it is not surprising to see lower diversity. In addition, we often observe noisy patterns in the baseline results, in which it verifies the difficulty of such depth colorization task for image-to-image translation models. Based on our reconstruction results, it is also worth noting that the appearance features extracted from the appearance reference are actually high-level and invariant to

Table 1. Quantitative evaluation on the NYU Depth v2 and SceneNet RGB-D datasets, in terms of PSNR for reconstruction and FID [6] for both quality/diversity of the colorization results, with comparison to the image-to-image translation baselines.

Dataset	NYU Depth v2		SceneNet RGB-D	
Metrics	PSNR	FID	PSNR	FID
Cycle GAN	8.9948	245.2066	9.7859	187.3561
Pix2Pix	10.9974	142.2589	11.8593	152.3537
BicycleGAN	10.7301	145.3382	15.5830	192.1874
Our Full Model	**12.1115**	**45.1402**	**22.3333**	**92.7267**

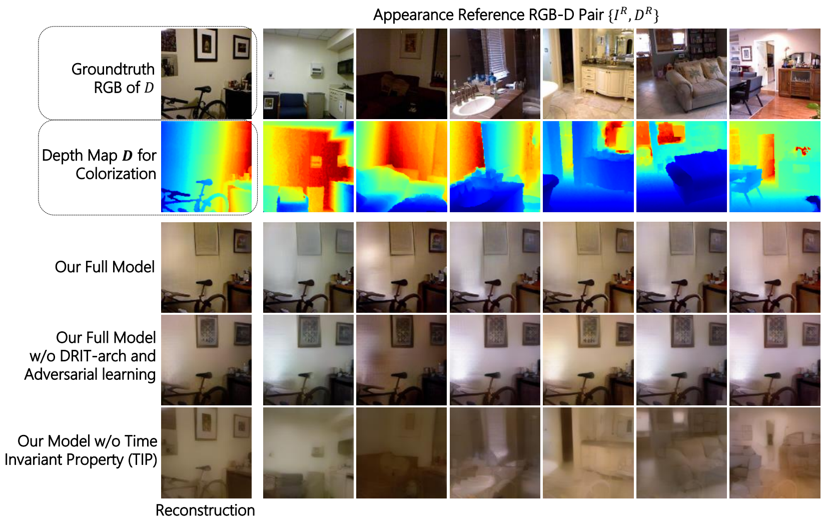
Appearance Reference RGB-D Pair $\{I^R, D^R\}$
Groundtruth RGB of D
Depth Map $\boldsymbol{D}$ for Colorization
Our Full Model
Our Full Model w/o DRIT-arch and Adversarial learning
Our Model w/o Time Invariant Property (TIP)
Reconstruction

Fig. 6. Qualitative examples of our design choices for model variants.

the structure, e.g., the objects on the same position in both ground truth RGB image and our reconstruction may not have the identical appearance.

4.3 Ablation Study

We perform an ablation study to investigate the contributions of our design choices in the proposed model on NYU Depth v2. The quantitative and qualitative evaluations of different variants are provided in the Fig. 6 and Table 2. Having all the designs in the full model achieves the best performance in terms of FID, showing that the results are the most realistic and diverse ones compared to other model variants. Also, both the random flipping operation and time invariant property (TIP) contribute to model learning, thus helping to produce more realistic colorization. Especially, the TIP plays an important role to largely improve the diversity as indicated by the FID scores. Regarding the DRIT-arch

Table 2. Ablation study of our design choices on NYU Depth v2.

Metrics	PSNR	FID
Our Full Model	12.1115	**45.1402**
w/o Random Flipping	13.8649	58.9894
w/o Time Invariant Property	**16.1137**	138.9675
w/o DRIT-arch & Discriminator	12.7394	55.9795

and adversarial learning with discriminator, they together benefit the training of disentanglement and improve the image quality of colorization. It is also worth noting that, although the model variant without TIP can achieve the best performance in PSNR, it can only perform well in the case of reconstruction (i.e., colorization on a depth map by having its corresponding ground truth RGB image as the appearance reference) but fail to nicely paint the depth map with other appearance sources, which leads to much worse FID scores and colorization results with artifacts as clearly shown in Fig. 6 (some structure information from the appearance reference stains the colorization in the bottom row).

Table 3. Comparisons between different methods in terms of consistency in average precision for object detection on NYU Depth v2 (top) and SceneNet (bottom) test sets. Here, we randomly select 5 appearance reference image-pairs and exclude them from the testing set for the "Our Full Model" setting, while using depth maps and corresponding reference image-pairs for the "Reconstruction" setting.

	chair	sofa	bed	tv	table	person	sink	fridge	toilet	oven
Ill-Lighted	3.5	9.5	19.4	15.3	12.6	26.8	28.4	31.3	33.2	12.6
CycleGAN	8.4	1.7	0.2	0.1	4.9	0.6	0.0	0.0	0.0	0.0
Pix2Pix	38.0	46.6	52.0	10.9	37.8	35.9	39.7	55.7	76.0	5.5
BicycleGAN	46.2	31.4	45.4	7.3	35.9	31.7	20.3	38.1	20.9	2.5
Reconstruction	**71.0**	**76.3**	82.5	**39.7**	**77.7**	**69.0**	59.7	**77.9**	88.1	**46.2**
Our Full Model	69.9	71.4	**85.0**	36.1	75.1	59.6	**61.7**	66.4	**90.0**	35.7

	chair	toilet	bench	bowl	pottedplant
Ill-Lighted	18.7	25.0	25.0	0.0	0.0
CycleGAN	33.0	15.9	8.3	3.6	0.0
Pix2Pix	0.0	0.0	4.9	12.8	6.1
BicycleGAN	23.4	21.5	25.0	6.3	0.0
Reconstruction	**70.3**	**75.4**	32.1	**68.8**	**66.7**
Our Full Model	65.1	67.0	**68.8**	62.5	44.4

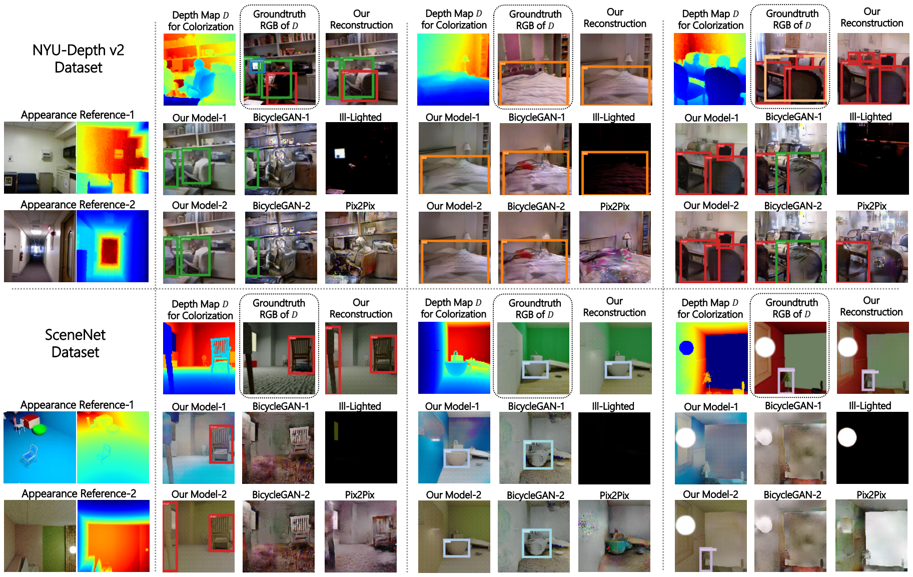

Fig. 7. Examples for recognition consistency. Our reconstruction and colorization show higher detection consistency with respect to the original image than the other methods.

4.4 Recognition and Temporal Consistency

As motivated in this paper, we would like to colorize the depth maps such that the vision models originally trained on RGB images can still function reasonably on the colorization results even when the illumination is insufficient. Therefore, given a depth map D and its corresponding RGB image I taken under sufficient lighting, we expect that the vision model would have similar/consistent recognition outputs across I and the colorization of D produced by our proposed method. In order to verify if such consistency exists, we adopt an off-the-shelf object detector, YOLOv3 [18] pre-trained on the COCO dataset [15], to perform the object detection on both ground truth RGB images and the colorization results, and then evaluate the consistency between their detection results.

The metric of consistency is defined upon the average precision of object detection on colorization results by considering the detection results on the original RGB images as the ground truths bounding boxes (IoU ≥ 0.5). Note that we do not perform any fine-tuning on the detector towards colorization. As COCO dataset has 80 object categories where most of them do not appear in NYU-Depth-v2, we manually select 10 object classes which frequently appear in NYU-Depth-v2 as our targets for verification. Similarly, for SceneNet, 5 object classes are chosen. In addition to having the comparison with the aforementioned image-translation methods in terms of consistency, we introduce another baseline which applies gamma-correction on the original RGB images for simulating the photos taken under ill-lighting situation (gamma equals to 10 in our experiments).

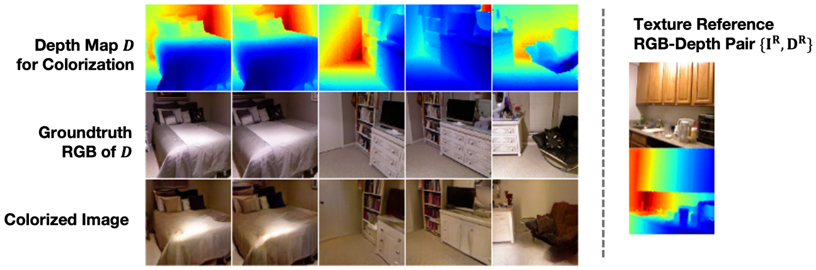

Fig. 8. Example results of colorizing a depth video.

The quantitative and qualitative results are shown in Fig. 7 and Table 3, respectively. The colorization results produced by our full model (using the appearance reference distinct from the original image) not only have higher consistency in comparison to other baselines, but also obtain comparable performance with respect to the reconstruction (i.e., depth map colorized by the appearance reference from its original image). These results validate the benefit of our depth map colorization for maintaining recognition ability of vision models up to a certain degree under the ill-lighting environment.

Moreover, we experiment on colorizing a video sequence of depth maps, based on a fixed RGB-Depth image-pair as the appearance reference, in order to testify the temporal consistency of our colorization results, i.e., the same object should be colorized similarly across video frames. As shown in the qualitative results of Fig. 8, our model is able to produce temporally smooth colorization, and we are able to well recognize the objects despite their different appearances compared to original RGB images.

5 Conclusions

We present a method for colorizing the depth map via disentanglement of image appearance and structure. A practical application is to provide an alternative, clear, and colorful view of the ill-lighting scene. Unlike previous works which usually produce unrealistic images, our model focuses on generating realistic colorization with the flexibility of using any reference image as the source of appearance information. Several self-supervised designs are adopted to realize our model training, such as random flipping, time invariant property (TIP), adversarial learning, and cycle consistency. The ablation study demonstrates the contributions of each design to encourage the image disentanglement that benefits colorization. Results on both the recognition and temporal consistencies further verifies the applicability of our proposed colorization model.

Acknowledgment. This project is supported by MOST109-2636-E-009-018, MOST-109-2634-F-009-020, and MOST-109-2634-F-009-015. Thanks to the National Center for High Performance Computing for computation facilities.

References

1. Bo, L., Ren, X., Fox, D.: Unsupervised feature learning for rgb-d based object recognition. In: Experimental robotics. pp. 387–402. Springer (2013) 2, 3
2. Carlucci, F.M., Russo, P., Caputo, B.: $(DE)^2CO$: Deep depth colorization. IEEE Robotics and Automation Letters **3**(3), 2386–2393 (2018) 2, 3, 4
3. Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M., Burgard, W.: Multimodal deep learning for robust rgb-d object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 681–687. IEEE (2015) 2, 3
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014) 4, 7, 8
5. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: European conference on computer vision. pp. 345–360. Springer (2014) 2, 3
6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems. pp. 6626–6637 (2017) 9, 11
7. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. ACM Transactions on Graphics (ToG) **35**(4), 1–11 (2016) 4
8. 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. pp. 1125–1134 (2017) 4, 9
9. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016) 6
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 8
11. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: European conference on computer vision. pp. 577–593. Springer (2016) 4
12. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European conference on computer vision (ECCV). pp. 35–51 (2018) 8
13. Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M., Yang, M.H.: Drit++: Diverse image-to-image translation via disentangled representations. International Journal of Computer Vision pp. 1–16 (2020) 8
14. Li, Y., Zhang, J., Cheng, Y., Huang, K., Tan, T.: Df^{2}net: Discriminative feature learning and fusion network for rgb-d indoor scene classification. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018) 2
15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 13
16. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2794–2802 (2017) 8, 9
17. McCormac, J., Handa, A., Leutenegger, S., Davison, A.J.: Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation?

In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2678–2687 (2017) 3, 9

18. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018) 13
19. Schwarz, M., Schulz, H., Behnke, S.: Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features. In: 2015 IEEE international conference on robotics and automation (ICRA). pp. 1329–1335. IEEE (2015) 2, 3
20. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012) 3, 9
21. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567–576 (2015) 2
22. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015) 9
23. Yuan, Y., Xiong, Z., Wang, Q.: Acm: Adaptive cross-modal graph convolutional neural networks for rgb-d scene recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9176–9184 (2019) 2
24. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017) 4, 9
25. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Advances in neural information processing systems. pp. 465–476 (2017) 4, 9