DEN: Disentangling and Exchanging Network for Depth Completion

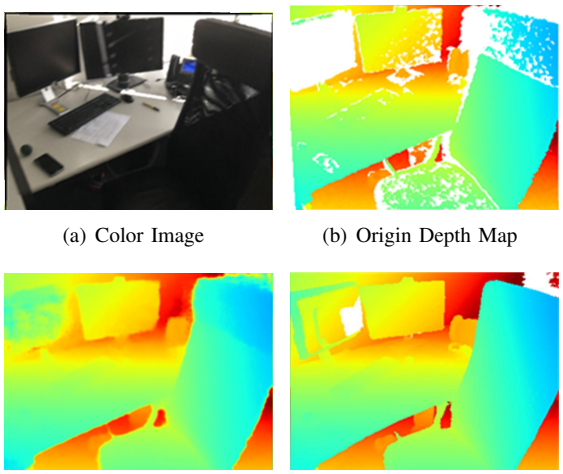
You-Feng Wu*, Vu-Hoang Tran†, Ting-Wei Chang‡, Wei-Chen Chiu‡, and Ching-Chun Huang‡
*Department of Electrical Engineering, National Chung Cheng University, Chiayi, Taiwan
Email: ubeewu@gmail.com
†HCMC University of Technology and Education, Ho Chi Minh City, Vietnam
Email: hoangtv@hcmute.edu.vn
‡Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan
Email: da20jeja1996@gmail.com
Email: walon@cs.nctu.edu.tw
Email: chingchun@cs.nctu.edu.tw

***Abstract*—In this paper, we tackle the depth completion problem. Conventional depth sensors usually produce incomplete depth maps due to the property of surface reflection, especially for the window areas, metal surfaces, and object boundaries. However, we observe that the corresponding RGB images are still dense and preserve all of the useful structural information. The observation brings us to the question of whether we can borrow this structural information from RGB images to inpaint the corresponding incomplete depth maps. In this paper, we answer that question by proposing a Disentangling and Exchanging Network (DEN) for depth completion. The network is designed based on the assumption that after suitable feature disentanglement, RGB images and depth maps share a common domain for representing structural information. So we firstly disentangle both RGB and depth images into domain-invariant content parts, which contain structural information, and domain-specific style parts. Then, by exchanging the complete structural information extracted from the RGB image with incomplete information extracted from the depth map, we can generate the complete version of the depth map. Furthermore, to address the mixed-depth problem, a newly proposed depth representation is applied. By modeling depth estimation as a classification problem coupled with coefficient estimation, blurry edges are enhanced in the depth map. At last, we have implemented ablation experiments to verify the effectiveness of the proposed DEN model. The results also demonstrate the superiority of DEN over some state-of-the-art approaches.**

I. Introduction

Depth estimation is a longstanding task in computer vision. A complete depth map can provide valuable information for many related tasks, including AR/VR applications, object detection, and autonomous driving. Nowadays, different kinds of commercial devices are available in the market for depth map acquisition. Some popular ones are Microsoft Kinect, Azure Kinect, and Intel RealSense. These devices can generate useful raw depth maps but are still not yet perfect. For highly reflective surfaces, shiny and bright regions, transparent objects, they usually produce imperfect or incomplete depth maps [1]. Besides, the estimated depth may be blurred or missing in the edge areas owing to significant depth differences between the foreground and background. Also, the resolution of the produced depth map is usually lower than the RGB image [2]. To obtain a complete and high-resolution depth map, a well-designed algorithm for depth completion thus becomes emergent.

This work was supported by Ministry of Science and Technology of Taiwan under the Grants No.109-2221-E-009-112-MY3, 109-2634-F-009-020, 109-2218-E-009-025, 106-2628-E-009-015-MY3.



(a) Color Image (b) Origin Depth Map (c) Ours (d) Ground Truth

Fig. 1. Example results from DEN network's output. The original depth map (b) in ScanNet data set is partially missing and DEN network filled in the missing part (c).

Given a raw depth map captured by a commercial depth camera, we aim to generate a complete depth map by leveraging the high-resolution RGB image as an auxiliary. Although some related works have been proposed for depth estimation, some challenges still remain. In this work, we mainly focus on the three major problems that modern depth estimation methods may encounter.

1) Mixed depth pixels, which leads to distorted and blurry edges between foreground and background, usually emerge when considering depth estimation as a regression problem [3]. It also causes severe interference on point cloud projection

and depth estimation. We introduced a newly proposed depth representation that takes advantage of both depth coefficients (DC) based representation [4] and space-increasing discretization (SID) [5]. The modification significantly relieves the mixed depth effects.

2) We treat RGB images as the auxiliary to assist the depth completion task. However, the excessively rich image texture may be added to the estimated depth map unexpectedly. Hence, disentangling only useful structural information from color images becomes a critical must to complete the missing depth.

3) The spatial scale offset problem inherently exists as a challenge for monocular depth estimation. Although the relative depth between objects can be roughly estimated, the absolute estimation retains an offset. We deal with this problem by mutually referring to the monocular image and the incomplete depth map.

To fulfill our ideas and address the issues mentioned above, we proposed a Disentangling and Exchanging Network (DEN). In the network, we introduce the concept of feature disentanglement, domain adaption, and content transfer to deal with the depth completion task. In detail, we applied the idea of domain adaptation and utilized adversarial networks to disentangle domain-invariant structural content and domain-specific style content from both color and depth images. Next, by transferring the complete structural content from the image domain to the depth domain, the depth map is then guided to complete. To train DEN, we constrain our network by the domain/content adversarial loss, the reconstruction loss, the cycle-consistency loss, the depth classification loss, and the surface normal estimation loss provided by [6]. The performance evaluation on ScanNet [7] dataset demonstrates that DEN can produce superior results over the state-of-the-art methods.

II. Related Work

This section will briefly introduce previous works over the three key research fields that are related to our model: (a) depth estimation, (b) depth representation, and (c) image-to-image translation.

A. Depth Estimation

1) Monocular Depth Estimation: Many deep CNN structures show the high-potential ability to understand geometric information of a single image. Thus, some researchers have used deep CNN to address the depth estimation task. Alhashim *et al.* [8] proposed an auto-encoder framework and utilized pretrained DenseNet-169 as its encoder network. Chen *et al.* [9] further introduced the concept of using adversarial networks. On the other hand, Xu *et al.* [10] applied a multi-scale CRF model, which enables using a color image to guide depth estimation. Later on, the same team *et al.* [11] then extended the idea to tackle the sequential depth estimation task by deep neural networks. Recently, Liana *et al.* [12] proposed an FCRN framework to improve the quality of depth estimation. It intended to solve the chessboard artifact caused by deconvolution while up-sampling the feature maps. By replacing the original large kernel with several small ones within the convolution layer, FCRN has efficiently reduced the chessboard artifact and preserves the local features even better.

2) Sparse Depth Map Completion: In the monocular depth estimation task, the spatial-scale offset problem still remains unsolved. Therefore, some recent works [13], [3] focus on another related task, depth completion, which considers integrating the features from both images and sparse depth maps to avoid the spatial-scale offset efficiently.

Zhang *et al.* [6] proposed learning networks to predict the surface normal and object boundaries from a given color image. An optimization setting for depth estimation, which is physically subject to the surface normals, object boundaries, and the spare depth maps, is then formulated. The method has achieved state-of-the-art performance on ScanNet dataset. However, the result is significantly affected by the accuracy of the surface normal prediction and boundary detection. Also, time-consuming optimization is another concern of the method. To maintain the efficiency of the overall system, researchers have recently worked on designing network structures for depth completion and training them in an end-to-end manner. For instance, Shivakumar *et al.* [2] proposed a spatial pyramid pooling module in their network which enables the model to capture features in different scales while downsampling. Jeon *et al.* [14] also worked on the model structure design and proposed the LapDEN framework based on the Laplacian pyramid model.

B. Depth Representation

Many related methods model depth completion/estimation as a regression problem, but the regression loss would encourage generating mixed depth pixels. Especially near the depth boundaries, the produced depth maps contain blurry depth edges and present inaccurate results. To reduce the mixed depth pixels, Fu *et al.* [5] first considered the depth estimation task as an ordinal classification problem instead of a regression problem. By leveraging cross-entropy loss, it somehow avoids depth mixing. However, dividing the depth dynamic range into multiple depth levels would cause quantization errors and inevitably damage the depth precision. Recently, to deal with the quantization error problem, Imran *et al.* [4] proposed a novel depth representation which attaches a coefficient for each depth level to interpolate the continuous depth.

C. Image-to-Image Translation

Image-to-image translation has become a popular topic lately and has been widely discussed in many different research fields. It refers to the problem of generating a target image map upon the extracted feature from a reference image map. Many tasks can be formulated as Image-to-image translation problems, including but not limited to image segmentation, image inpainting, and depth estimation. Some learning networks have been proposed to realize different Image-to-image translation tasks. They go from a pair-wise

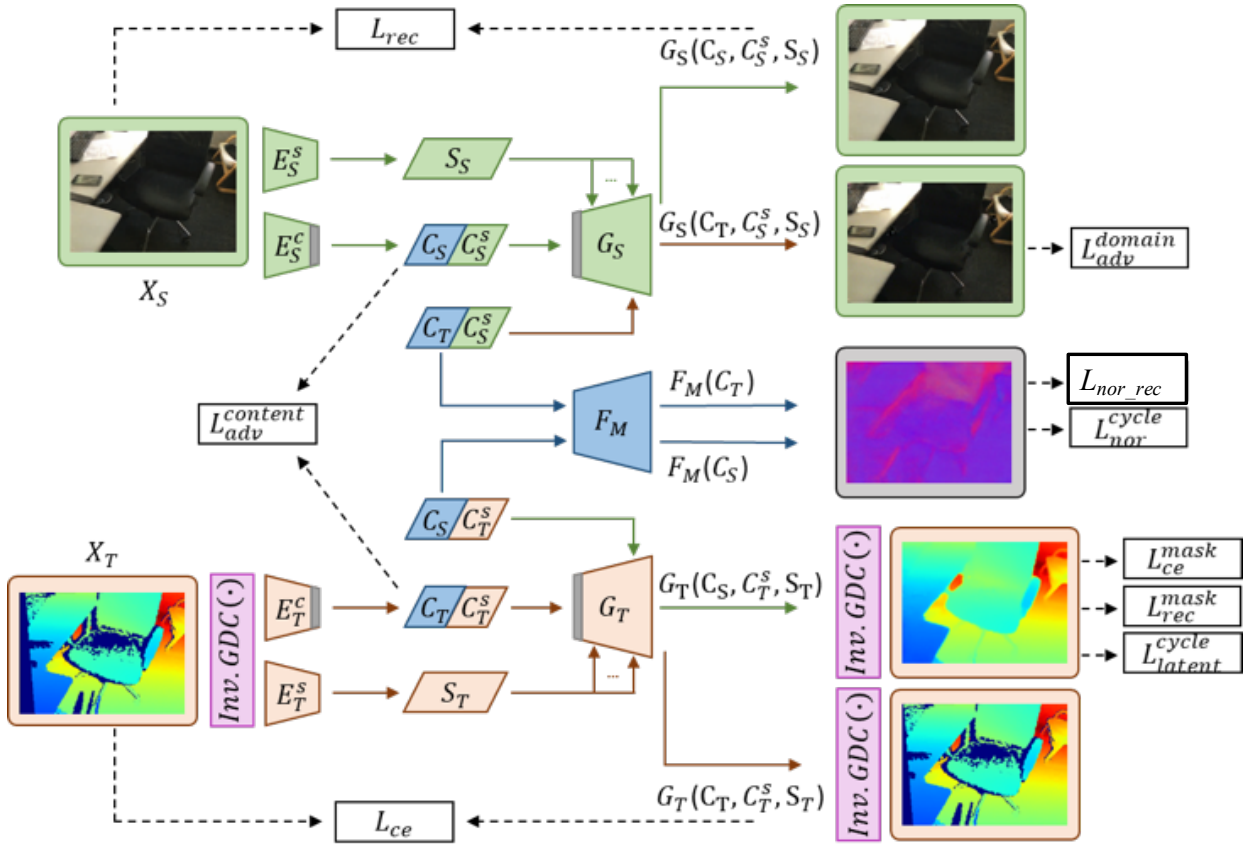


Fig. 2. The proposed DEN framework. The upper(green) and lower(red) part are auto-encoder structure for RGB image and depth image, respectively. C_S, C_T are common content and C_S^S, C_T^S are specific content from RGB and depth image, respectively. We designed an adversarial network to guide the extraction of common content then further exchange C_S and C_T for style-transferring. We borrow the idea of [6] and leverage on surface normal reconstruction F_M to constrain C_S and C_T we get from E_S^C and E_T^C.

network setting [15] to an unpaired input setting [16]. They can also be trained in a supervised or unsupervised [17] way. Some novel ideas, which are widely applied to image-to-image translation tasks, are introduced in the following.

Owing to the difficulty of obtaining pair-wise data, Zhu *et al.* [16] proposed CycleGAN to deal with unpaired image-to-image translation. Many works have reported success in leveraging the cycle consistency loss to constrain network learning. Some researchers treat unsupervised image translation as a domain adaptation problem. For instance, Liu *et al.* [17] applied the idea of distribution alignment between two domains. It assumes that both the source and target domains can be mapped to shared latent space. Image translation is then achieved in this space. However, it may not be true that hypothesizing the feature distributions extracted from both domains can always be well aligned. Continuing to the trend, Lee *et al.* [18] modified CycleGAN's structure and proposed the DRIT framework. It separates the encoder into a common content encoder and a specific attribute encoder for feature disentangling. Furthermore, Tran *et al.* [19] introduced a unified network to combine domain adaptation, feature disentanglement, and style transfer. By adapting the common content of two domains, image translation can be accomplished by exchanging the common content between the two domains.

III. Method

In this section, we will introduce our method, DEN, for depth completion. We will first describe our network architecture; Then, we would detail the proposed depth representation; at last, we explain the training constraints we designed for network learning.

A. Network Architecture

Our architecture, shown in Figure 2, mainly consists of two auto-encoder networks. One deals with RGB images, and the other one processes depth maps. We inherit the idea from DRIT [18] to design two encoders, E^S and E^C, named as the style encoder and the content encoder. However, funded on a different precondition from that of DRIT [18], we hypothesize that a complicated RGB image contains some specific texture content, which a depth map does not have. These specific texture content, containing hidden structural information, should be further separated from the output of the content encoder so that the shared common content can be extracted. Thus, unlike DRIT, which rigidly maps the content parts extracted from

both domains to the same latent space (domain adaptation), we left a free zone in each content latent space, named as specific content C^S. Thus, each content encoder can freely extract both the specific content information and the common structural content, C, rigidly bounded by the domain discriminator, D_M, and the surface normal decoder, F_M. Here, F_M is designed to determine the surface normal given C because we believe that surface normal[6] information embeds the scene structure and is highly related to depth content. Besides, to ensure the completeness of encoded information, the decoders, G_S and G_T, are designed to generate the reconstructed images conditioning on the style S, common content C, and specific content C^S. Furthermore, to help align the latent space of content features C_T and C_S, we share weights at the last layers of encoders E_T^C, E_S^C and the first layers of decoders G_S, G_T respectively.

Next, we exchanged the common contents, C_S and C_T, which are extracted by E_S^C and E_T^C, respectively. By feeding the exchanged latent features into the depth decoder, we are able to generate a completed depth map, $G_T(C_S, C_T^S, S_T)$, which is guided by RGB image's structural information, C_S. Finally, inspired by the idea CycleGAN [16], but applied on latent space instead of image space, we sent the completed depth map ($G_T(C_S, C_T^S, S_T)$) into E_T^C again and check for the consistency between the latent features extracted from the original image and the completed depth map, respectively. By introducing the cycle consistency loss, we make the network training more stable.

B. Depth Representation

As shown in Fig. 3, to deal with the mixed depth pixel problem, inherited from the main idea of depth coefficient (DC) [4], our method, named as general depth coefficient (GDC), also divided the depth interval [α, β] into K bins $\{D_1, D_2, ..., D_K\}$. The depth value d_i is then represented by the depth coefficient vector $c_i = \{W_{i,1}, W_{i,2}, ..., W_{i,K}\}$, which actually is a sparse representation with only three non-zero coefficients attached to the quantized bin and its two neighboring bins. Suppose that k is the index of the depth bin closest to the depth value d_i, then depth coefficient vector will be $c_i = \{0, ..., 0, W_{i,k-1}, W_{i,k}, W_{i,k+1}, 0, ..., 0\}$, where the coefficient $W_{i,k}$ for the quantized bin D_k is 0.5. The coefficients for two neighbors D_{k-1} and D_{k+1} are $W_{i,k-1}$ and $W_{i,k+1}$, where $W_{i,k+1} = 0.5 - W_{i,k-1}$. The other coefficients are zero. Given a depth coefficient vector c_i for the i^{th} depth sample, we then can estimate the continuous depth $\hat{d}_i$ in a linear combination manner as Eq. 1.

$$\hat{d}_i = \frac{W_{i,k-1}D_{k-1} + W_{i,k}D_k + W_{i,k+1}D_{k+1}}{W_{i,k-1} + W_{i,k} + W_{i,k+1}} \tag{1}$$

In DEN network training, we force DEN network to predict $\hat{W}_{i,j}$, estimated depth coefficients, so that the following cross entropy loss can be minimized.

$$L_i^{ce}(c_i) = -\sum_{j=1}^{K} W_{i,j} log(\hat{W}_{i,j}), \tag{2}$$

where, the ground truths for depth coefficients can be inversely infer from a depth ground truth d_i as follows:

$$W_{i,k-1} = \frac{d_i - 0.5(D_k + D_{k+1})}{D_{k-1} - D_{k+1}}, \tag{3}$$

$$W_{i,k+1} = 0.5 - W_{i,k-1}. \tag{4}$$

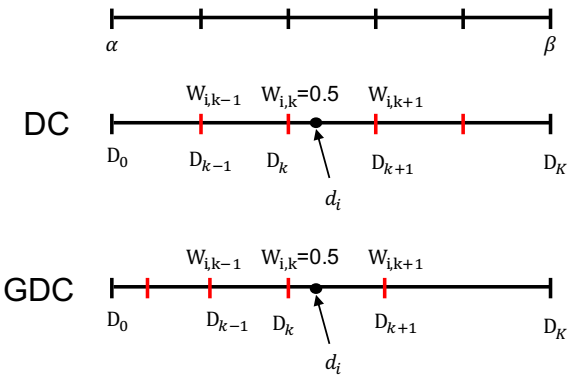


Fig. 3. General Depth Coefficient (GDC) and Depth Coefficient (DC).

In the DC method [4], a conventional way, uniform discretization (UD), is applied for quantization. As shown in Fig. 3, the depth interval is divided into several bins with identical size. However, if considering the geometry property of image projection, the precision of depth estimation would decay as the distance grows. Thus, it may not be the best way to treat the quantization step in the nearer depth field the same as the further ones. To solve this problem, different from [4], we introduce spacing increasing discretization (SID) [5] defined as Eq. 5 for quantization. It divides the depth interval [α, β] into K non-uniform bins by emphasizing the closer depth and allowing more error tolerance in the further areas.

$$D_i = e^{log(\alpha) + log(\beta - \alpha) * i / K} \tag{5}$$

C. Loss Functions

1) Reconstruction Loss: Just like most of the other auto-encoders, we have to make sure images reconstructed by the extracted content and style parts still retain most of the important components of the original image. Given C_S and S_S, we formulate the reconstruction loss between the reconstructed and original RGB images by mean square error (L2 loss). On the other hand, we use cross entropy in Eq. 2 to calculate the reconstruction loss of depth maps conditioned on C_T and S_T.

Similar to [19], to constrain the generated RGB image after feature exchanging, $G_S(C_T, C_S^S, S_S)$, we introduce an image discriminator D_S and apply a domain adversarial loss defined as in Eq. 6. By training in an adversarial way, it will ensure that the generated RGB images have a similar style as the real RGB images, X_S.

$$L_{adv}^{domain} = log(D_S(X_S)) + log(1 - D_S(G_S(C_T, C_S^S, S_S))). \tag{6}$$

Also, DEN inpaints the depth map by referring to the common content part from the RGB image C_S. In order to obtain a fine-grained depth prediction, recommended by Ruiz *et al.* [20], we minimized both the masked cross-entropy loss L_{ce}^{mask} and the masked mean square loss L_{rec}^{mask}. They are defined as:

$$L_{rec}^{mask} = \frac{1}{N'}\sum_{i=1}^{N} -m_i(d_i - \hat{d}_i)^2. \tag{7}$$

$$L_{ce}^{mask} = \frac{1}{N'}\sum_{i=1}^{N} l_i^{ce}(m_i, c_i, \hat{c}_i), where \tag{8}$$

$$l_i^{ce} = -m_i \sum_{j=1}^{K} c_{ij} log(\hat{c}_{ij}). \tag{9}$$

Here, m_i is a mask pixel whose value is equal to 1 when ground truth depth is valid and is equal to 0 otherwise. N' and N are the counts of valid pixels and all pixels. $\hat{d}_i$ and d_i are the estimated continuous depth and its ground truth. $\hat{c}_i$ and c_i are the estimated GDC vector and its ground truth, respectively. The final depth reconstruction loss is defined as $L = L_{ce}^{mask} + \alpha L_{rec}^{mask}$.

Moreover, we introduce the surface normal as supervision to guide the network. To make sure the C_S and C_T are able to reflect the structural information, the surface normal estimated by the decoder F_M is forced to be close to the real surface normal N_{nor}^{gth} provided by [6]. The surface normal supervision leads to another mean square loss as in Eq. 10.

$$L_{nor\_rec} = ||N_{nor}^{gth} - F_M(C_S)||_2^2 + ||N_{nor}^{gth} - F_M(C_T)||_2^2. \tag{10}$$

2) Content Adversarial Loss: As mentioned before, we intend to align the target latent space and the source latent space C_T, C_S extracted by the encoders E_S^C, E_T^C. To make sure the distributions of the two content types are well mixed, we introduce a discriminator D_M to determine the domain class of a given content feature. Next, we train the two encoders to maximumly confuse D_M by minimizing the adversarial loss function, which is defined as:

$$\begin{aligned} L_{adv}^{content} = &\frac{1}{2}log(D_M(C_T)) + \frac{1}{2}log(1 - D_M(C_S)) + \\ &\frac{1}{2}log(D_M(C_S)) + \frac{1}{2}log(1 - D_M(C_T)). \end{aligned} \tag{11}$$

While training E_S^C, we encourage the discriminator to label C_S as 1 and C_T a 0; on the other hand, while training E_T^C, we hope the discriminator labels C_S as 0 and C_T as 1. When C_T, C_S are well aligned, the discriminator D_M may not discriminate the difference between the two domains. Ideally, it would output a value near 0.5 which minimizes Eq. 11.

3) Cycle Latent feature and Surface Normal Consistency Losses: We inherit the idea of CycleGAN [16] and introduce the idea of cycle consistency loss but for latent space. We feedback the final output, the completed depth map, $G_T(C_S, C_T^S, S_T)$, into our content encoder E_T^C and constrain its latent features to be consistent with the original latent features C_S. The loss function is defined as:

$$L_{latent}^{cycle} = ||C_S - E_T^C(G_T(C_S, C_T^S, S_T))||_2^2 \tag{12}$$

Following the same concept, we also introduce the cycle surface normal consistency loss defined as

$$L_{nor}^{cycle} = ||N_{nor}^{gth} - F_M(C_S')||_2^2, \tag{13}$$

where N_{nor}^{gth} is the ground truth surface normal of the origin image; $C_S' = E_T^C(G_T(C_S, C_T^S, S_T))$ is the extracted content feature when feeding the generated completed depth map, $G_T(C_S, C_T^S, S_T)$, into the content encoder E_T^C.

IV. Experiment

In this section, we show a series of experiments using the ScanNet dataset [7] to evaluate the performance of our disentangling and exchanging depth completion network. We performed detailed ablation studies to show the impact of the different constraints and architecture choices. We also discuss the estimation at different time steps and visualize our final output with both 2D depth maps and 3D projected point clouds. It helps to demonstrate the improvement of reducing mixed depth pixels. Besides, to prove our original hypotheses about feature disentanglement, we also visualize the disentangled features in both image and feature domains.

A. Training Dataset

Since there are more and more commercial RGB-D sensors published and applied nowadays, 3D-geometric related researches have drawn significant attention. It also opened many new applications and research fields. Modern deep learning models show their high potential for 3D semantic understanding and facilitate development in these research fields. However, the biggest constraint of supervised deep learning is the demand for large labeled datasets. To deal with this problem, Dai *et al.* [7] introduced the ScanNet dataset, containing 1513 scenes (over 2.5 million images). Every image is annotated with its camera pose, surface reconstruction, which includes ground truth surface normal and boundaries, and instance-level semantic segmentation. This dataset has helped to achieve state-of-the-art performance on several 3D scene understanding tasks, including 3D object classification and semantic voxel labeling. In this work, we used the ScanNet reconstructed 3D model to generate the depth ground truth. In addition, we borrowed the image surface normal provided by Zhang *et al.* [6] as additional supervision. We followed the setting of Zhang *et al.* [6] which applied 59743 pairs of data from about 1000 scenes for training then tested with another 500 pairs from other scenes for evaluation. Our model is trained and tested using single NVIDIA RTX 2080ti GPU, and we measure an average speed 4.8 seconds per frame.

B. Evaluation Metrics and Comparison Circumstances

We compare the performance of our method under three circumstances: the whole depth map (B), which means comparing all the valid parts in the ground truth; the valid parts of

the raw depth map (Y), which means only comparing the parts that both ground truth and the raw depth map are valid; and the missing parts of the raw depth map (N), which means only comparing the parts that is valid in ground truth but invalid in the raw depth map. We evaluate the performance with mean related error (Rel), root mean square error (RMSE), and root mean square threshold error (tRMSE) to show the performance of reducing mixed depth pixels. We also introduce threshold accuracy (δ) for evaluation. The metrics are defined as follows:

$$Rel = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i}, \tag{14}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{15}$$

$$tRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} min((y_i - \hat{y}_i)^2, t^2), and} \tag{16}$$

$$\delta_i = max(\frac{y_i}{\hat{y}_i}, \frac{\hat{y}_i}{y_i}). \tag{17}$$

Here, y_i is the i^{th} pixel in the depth ground truth y, $\hat{y}_i$ is the i^{th} pixel in the generated depth map $\hat{y}$, and N is the total number of pixels considered in each depth map.

TABLE I
COMPARISON OF DEN AND SOME STATE-OF-THE-ART METHODS ON SCANNET

Obs	Method	Rel	RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
B	Bilateral [21]	0.084	0.411	0.9073	0.9412	0.9584
	DDC [6]	0.087	0.320	0.9213	0.9588	0.9764
	DEN(GDC)	**0.074**	**0.304**	**0.9247**	**0.9621**	**0.9794**
Y	Bilateral [21]	0.049	0.248	0.9588	0.9757	0.9857
	DDC [6]	0.049	0.248	0.9588	0.9757	0.9856
	DEN(GDC)	**0.047**	**0.230**	**0.9617**	**0.9786**	**0.9877**
N	Bilateral [21]	0.226	0.697	0.7560	0.8398	0.8781
	DDC [6]	0.201	0.471	0.8113	0.9092	0.9492
	DEN(GDC)	**0.156**	**0.457**	**0.9160**	**0.9134**	**0.9551**

C. Depth Completion

To evaluate the depth completion performance of our disentangling and exchanging network, we compare to the state-of-the-art CNN-based depth completion methods on the ScanNet dataset. DEN network takes advantage of RGB-D images through inputting pairwise color images and depth images. To generate the completed depth maps in the test phase, the content feature C_S extracted from the RGB image is first concatenated with the target-specific content C_T^S and the target-style feature S_T extracted from the incomplete depth map. Next, the concatenated feature is input to the generator G_T. Table I show the superiority of DEN network. By making good use of the technique of domain adaptation, the network achieves better performance when running depth completion tasks on partial missing depth images. Our DEN network, shown in Figure 2, adopts novel designs that combine a part of the feature extracting ideas from DRIT [18] and Hoang's method *et al.* [19]. Hence, it is able to efficiently guide depth completion tasks according to RGB images and exclude the rich texture information that may mislead the result. This advantage is especially magnified when comparing on the circumstance N, where we achieve significant improvement compared with other previous works. Since the missing parts are our main target in the depth completion tasks, the improvement shows the stability and feasibility of DEN network.

TABLE II
COMPARISON OF DIFFERENT DEPTH REPRESENTATIONS. SP STANDS FOR SPARSE DEPTH MAP (USED BY REGRESSION MODELS), DC STANDS FOR DEPTH COEFFICIENT (WITH UD DISCRETIZATION STRATEGY) AND GDC STANDS FOR GENERAL DEPTH COEFFICIENT (WITH SID DISCRETIZATION STRATEGY).

Obs	Method	Rel	RMSE	tRMSE
B	DEN(SP)	0.0892	0.3194	0.2383
	DEN(DC)	0.0778	0.3072	0.2220
	DEN(GDC)	**0.0748**	**0.3043**	**0.2195**
Y	DEN(SP)	0.0552	0.2312	0.1709
	DEN(DC)	0.0479	**0.2290**	**0.1634**
	DEN(GDC)	**0.0470**	0.2300	0.1636
N	DEN(SP)	0.1890	0.4948	0.3713
	DEN(DC)	0.1658	0.4665	0.3400
	DEN(GDC)	**0.1567**	**0.4574**	**0.3332**
Obs	Method	1.25	1.25^2	1.25^3
B	DEN(SP)	0.9102	0.9591	0.9793
	DEN(DC)	0.9235	0.9616	0.9793
	DEN(GDC)	**0.9247**	**0.9621**	**0.9794**
Y	DEN(SP)	0.9594	0.9788	0.9883
	DEN(DC)	0.9616	**0.9787**	**0.9879**
	DEN(GDC)	**0.9617**	0.9786	0.9877
N	DEN(SP)	0.7653	0.9014	0.9527
	DEN(DC)	0.9115	0.9111	0.9544
	DEN(GDC)	**0.9160**	**0.9134**	**0.9551**

D. Ablation Study

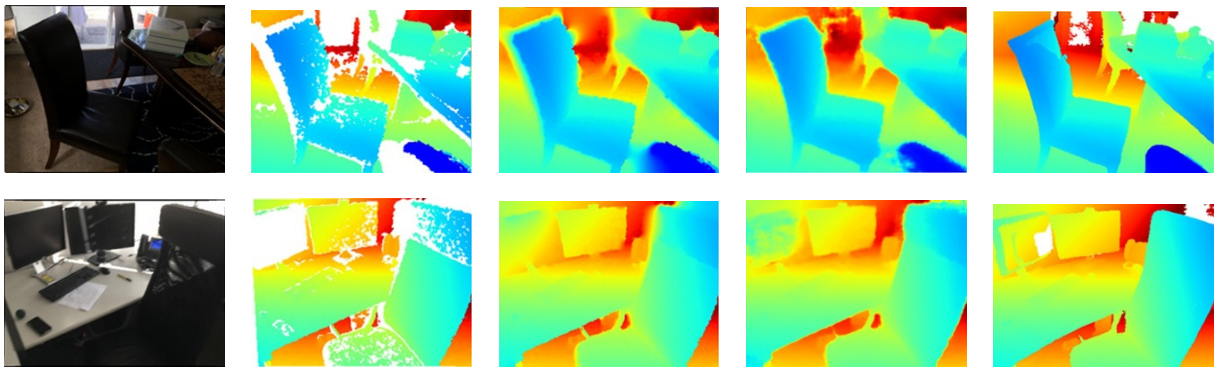
In this section, we investigate the importance of each technique applied in our depth completion network, including the new depth representation, the cycle consistency constraints, and the regression constraints.

Depth Representation. In this paper, we proposed a new depth representation: named as GDC. We performed a series of experiments to test three different kinds of depth representations on our DEN framework with all constraints applied. As shown in Table II, although the performance of GDC on the valid parts of the raw depth map (Y) does not show significant improvement than the original DC, its performance on the missing part (N), which we aim to complete, is significantly better than other depth representations. Overall, GDC shows the best performance on the whole depth map (B).

TABLE III
ABLATION STUDY OF ADDITIONAL CONSTRAINTS.

L_{ce}^{mask}	L_{rec}^{mask}	L_{nor}^{cycle}	L_{latent}^{cycle}	tRMSE	$\delta < 1.25$
✓				0.2575	0.8902
✓	✓			0.2292	0.9168
✓	✓	✓		0.2235	0.9210
✓	✓	✓	✓	**0.2195**	**0.9247**

Cycle Consistency and Reconstruction Regression Constraints. Next, we investigate the influence of the additional constraints we applied to improve the performance. Note that



(a) Color Image (b) Sparse Depth Map (c) [6] (d) Ours (e) Ground Truth

Fig. 4. Visual 2D depth map comparison between state-of-the-art.

we have added the masked reconstruction mean square loss L_{rec}^{mask} to increase the precision of the completed depth map. We also inherited the idea of CycleGAN [16] and introduced the cycle latent feature consistency loss and cycle surface normal consistency loss to our network. To know the effectiveness, we did the ablation study on the four constraints, L_{ce}^{mask}, L_{rec}^{mask}, L_{nor}^{cycle}, L_{latent}^{cycle}. tRMSE and threshold accuracy, two evaluation metrics that are related to the accuracy of bin classification, are used to evaluate the performance on the whole depth maps. As shown in Table III, we can see the improvement by adding L_{rec}^{mask} is significant. Furthermore, the cycle consistency losses also slightly enhance the stability of DEN network and help to refine our result.

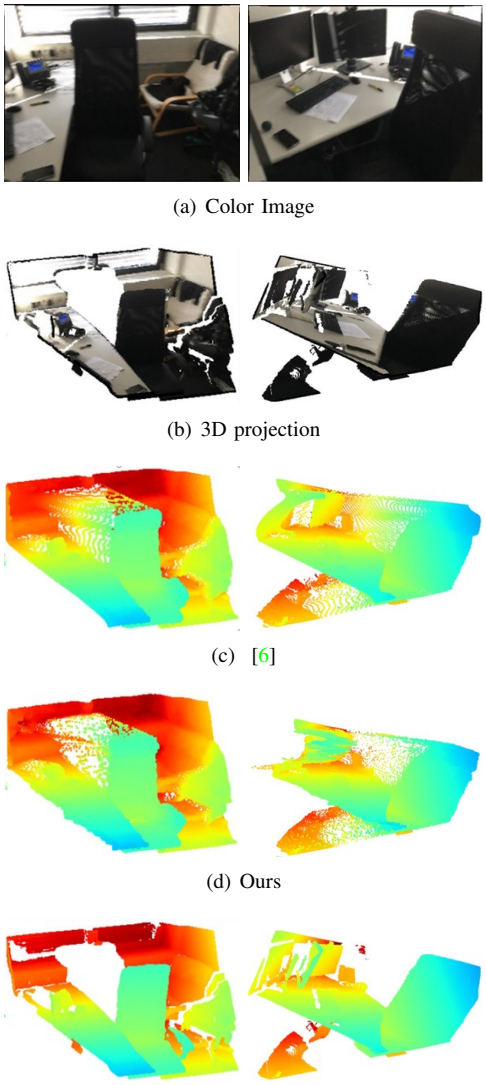
E. Visualization

In this section, we demonstrate our depth completion results by visualizing the 2D depth maps and 3D point clouds. For comparison, we also visualize other results produced from the state-of-the-art method [6].

2D Depth Map. As shown in Figure 4, we can see that DEN does a better job in terms of separating foreground objects and background objects. Our process also generates fewer mixed depth pixels (the light blue pixels behind the chair back on the top row).

3D Point Cloud. By utilizing the ground truth depth map, we can back-project the 2D image and depth map into the 3D point cloud format, which provides us a better visualization to understand the depth estimation performance. As shown in Figure 5, we may see more evidence that DEN model can reduce the mixed depth pixels.

Besides, in order to understand what kind of information are embedded in disentangled features, we visualize them in image domain as shown in Fig. 6. Suppose as in Fig. 6 (b), we want to visualize the specific content (C_S^S), so we will suppress the other information (i.e. S_S and C_S) in feature domain before reconstructing. So in this Fig., all the specific information about RGB image are preserved. If we visualize the common content and style (RGB image style), we can see some structural information are extracted and all of the color



(a) Color Image

(b) 3D projection

(c) [6]

(d) Ours

(e) Ground Truth

Fig. 5. 3D point cloud visualization and comparison. We can see clearly that our method has fewer mixed depth pixels and separates the foreground objects and background objects better, such as the chair back. Ground Truth of the 3D point cloud is generated by back-projecting the ground truth depth map.

(a) Color Image (b) Specific Content

(c) Common Content + Style (d) Common Content

Fig. 6. Disentangled parts in image domain.

information are removed as shown in Fig. 6 (c). And in Fig. 6 (d), because the suppression of style information make it difficult to observe but we still can recognize some structural patterns embedded in the image.

V. Conclusion

In this paper, we proposed the DEN framework and utilized the RGB image to guide the depth completion task. As we discussed before, there are three major problems that modern depth estimation models may encounter. 1) Mixed depth pixels. We deal with this problem by introducing the new depth representation, GDC. We also combine the cross-entropy and mean square losses to improve the precision of depth estimation. 2) Excessive rich texture details on RGB images would cause undesired depth estimation results. We deal with this problem by disentangling only the structural information from the RGB image and utilize it on depth completion. 3) Spatial scale offset, which is no longer a significant problem since we applied the sparse depth image for reference. Though there is a lot of difference between depth maps and RGB images, our DEN network shows the effectiveness of extracting structural information from both domains. We applied DEN on the ScanNet dataset and demonstrated its superiority over the state-of-the-art approaches.

References

[1] A. Maimone and H. Fuchs, "Reducing interference between multiple structured light depth sensors using motion," in *2012 IEEE Virtual Reality Workshops (VRW)*. IEEE, 2012, pp. 51–54.

[2] S. S. Shivakumar, T. Nguyen, S. W. Chen, and C. J. Taylor, "Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion," *arXiv preprint arXiv:1902.00761*, 2019.

[3] M. Carvalho, B. Le Saux, P. Trouvé-Peloux, A. Almansa, and F. Champagnat, "Deep depth from defocus: how can defocus blur improve 3d estimation using dense neural networks?" in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[4] S. Imran, Y. Long, X. Liu, and D. Morris, "Depth coefficients for depth completion," *arXiv preprint arXiv:1903.05421*, 2019.

[5] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.

[6] Y. Zhang and T. Funkhouser, "Deep depth completion of a single rgb-d image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 175–185.

[7] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5828–5839.

[8] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv:1812.11941*, 2018.

[9] R. Chen, F. Mahmood, A. Yuille, and N. J. Durr, "Rethinking monocular depth estimation with adversarial training," *arXiv preprint arXiv:1808.07528*, 2018.

[10] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3917–3925.

[11] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5354–5362.

[12] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 239–248.

[13] Y. Liao, L. Huang, Y. Wang, S. Kodagoda, Y. Yu, and Y. Liu, "Parse geometry from a line: Monocular depth estimation with partial laser observation," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 5059–5066.

[14] J. Jeon and S. Lee, "Reconstruction-based pairwise depth dataset for depth image enhancement using cnn," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 422–438.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[17] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*, 2017, pp. 700–708.

[18] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 35–51.

[19] V.-H. Tran and C.-C. Huang, "Domain adaptation meets disentangled representation learning and style transfer," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, Oct 2019, pp. 2998–3005.

[20] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2074–2083.

[21] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*. Springer, 2012, pp. 746–760.