Learning Scene Structure Guidance via Cross-Task Knowledge Transfer for Single Depth Super-Resolution

Baoli Sun[1], Xinchen Ye[1,2]*, Baopu Li[3], Haojie Li[1,2], Zhihui Wang[1,2], Rui Xu[1,2]

[1]International School of Information Science & Engineering, Dalian University of Technology, China
[2]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, China
[3] Baidu Research, USA

Abstract

Existing color-guided depth super-resolution (DSR) approaches require paired RGB-D data as training samples where the RGB image is used as structural guidance to recover the degraded depth map due to their geometrical similarity. However, the paired data may be limited or expensive to be collected in actual testing environment. Therefore, we explore for the first time to learn the cross-modality knowledge at training stage, where both RGB and depth modalities are available, but test on the target dataset, where only single depth modality exists. Our key idea is to distill the knowledge of scene structural guidance from RGB modality to the single DSR task without changing its network architecture. Specifically, we construct an auxiliary depth estimation (DE) task that takes an RGB image as input to estimate a depth map, and train both DSR task and DE task collaboratively to boost the performance of DSR. Upon this, a cross-task interaction module is proposed to realize bilateral cross-task knowledge transfer. First, we design a cross-task distillation scheme that encourages DSR and DE networks to learn from each other in a teacher-student role-exchanging fashion. Then, we advance a structure prediction (SP) task that provides extra structure regularization to help both DSR and DE networks learn more informative structure representations for depth recovery. Extensive experiments demonstrate that our scheme achieves superior performance in comparison with other DSR methods.

1. Introduction

To better understand a scene image, depth information is supplemented to the RGB images, providing the key clue about the scene and enabling wide applications in 3D reconstruction [9], autonomous navigation [20], monitoring [2], and so on. However, acquiring depth information for indoor and outdoor scenes needs expensive cost and great efforts, especially for high-quality and high-resolution (HR) depth maps. As such, one of the effective post processing techniques, Depth Super-Resolution (DSR), is greatly desired to yield HR depth maps to alleviate this problem. Many efforts have been taken along the direction of DSR. Usually, fine scene structures are easily lost or severely destroyed in low-resolution (LR) depth map because of the limited spatial resolution. An RGB image and its associated depth map are the photometric and geometrical representations of the same scene, and have a strong structural similarity. Most existing DSR methods learn structural complementarity from RGB images to recover the degraded depth maps.

Previous color-guided DSR methods take advantage of RGB-D image pairs via a two-way fusion architecture, in which an extra branch is required to extract structural guidance from RGB image. As illustrated in Figure 1 (a), RGB image and LR depth map are often processed by separate branches and filtered together through a joint branch to output the HR result [26, 31, 4, 21]. However, due to the simple feature aggregation at a specific layer in the middle of the network, high-frequency structure information from RGB image is more likely to be lost in the process of feature extraction. Therefore, as shown in Figure 1 (b), some novel methods [18, 12, 49] incorporate a new paradigm of feature aggregation, i.e, multi-scale fusion, to allow the network to learn rich hierarchical features at different levels. This in turn makes the network to retain more spatial details for recovering both fine-scale and large-scale structures.

Although existing color-guided DSR methods have demonstrated remarkable progress, several limitations still remain. First, these methods require paired RGB-D data as training examples to jointly recover the degraded depth

*Corresponding author: yexch@dlut.edu.cn. This work was supported by National Natural Science Foundation of China (NSFC) under Grant 61702078, 61772108, 61976038, 61772106.

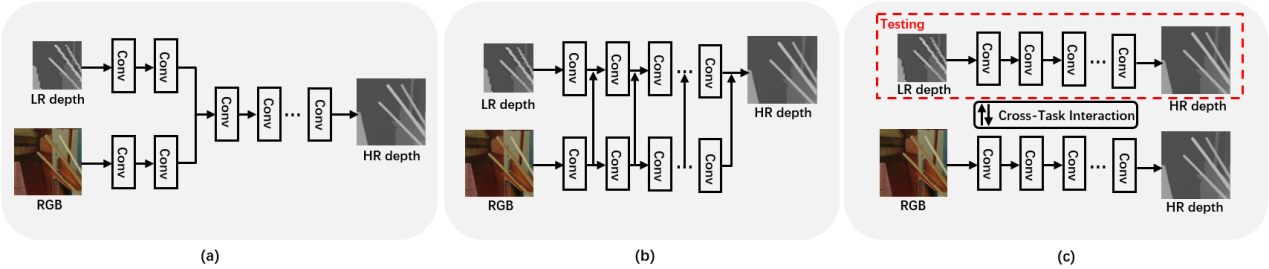


Figure 1. Color-guided DSR paradigms. (a) Joint filtering, (b) Multi-scale feature aggregation, (c) Our cross-task interaction mechanism to distill knowledge from RGB image to DSR task without changing its network architecture.

map. However, the paired data may be limited or expensive to be collected in actual testing environment. For example, RGB image and depth map are captured by separate depth and RGB sensors with different resolutions and views, thus needing accurate calibration and rectification between them to obtain the registered pairs. Actually, most of real-world applications still come with only a single LR depth map, which raises the above question. Second, considering the memory consumption and computing burden, the processing on the HR RGB data also hinders the practical application. Moreover, although RGB features can be used as structural guidance to resolve the degradation in DSR, RGB discontinuities do not always coincide with those of depth map (structure inconsistency), which results in noticeable artifacts such as texture copying and depth bleeding. Therefore, how to leverage RGB information to help recover the depth map and simultaneously satisfy the actual testing environment, still needs to be studied.

Motivated by the above analysis, this paper breaks away from the shackles of general paradigms and introduces a novel scene structure guidance learning method for the task of DSR, as shown in Figure 1 (c). We explore for the first time to learn the cross-modality knowledge at training stage, where both RGB and depth modalities are available, but test on the target dataset, where only single depth modality exists. Our key idea is to distill the knowledge of scene structural guidance from RGB modality to the single DSR task without changing its network architecture.

Specifically, as illustrated in Figure 2, inspired by the success of multi-task learning [52, 22, 46], we construct an auxiliary depth estimation (DE) task that takes RGB image as input to estimate a depth map. Upon this, we propose a cross-task interaction module to realize bilateral knowledge transfer between DSR task and DE task. Different from the commonly used distillation techniques [37, 16, 22], we first design a cross-task distillation that encourages DSR network (DSRNet) and DE network (DENet) to learn from each other, i.e., the roles of teacher and student will dynamically switch between both tasks based on their current performances on depth recovery in the iterative collaborative training. A multi-space distillation scheme is introduced to distill knowledge from the perspective of output and affinity spaces, which can better describe the essential structural characteristics of depth map. Moreover, to address the problem of RGB-D structure inconsistency, we construct a structure prediction (SP) task that provides extra structure regularization to help both DSRNet and DENet learn more informative structure representations for depth recovery. We come up with an uncertainty-induced attention fusion module to provide a reasonable input for the SP network (SPNet), in which the uncertainty maps acquired from both DSRNet and DENet are used to re-weight their features for strengthening effective structural knowledge. Extensive experiments demonstrate that our single DSR method even outperforms the color-guided DSR methods on benchmark datasets in terms of both accuracy and runtime. The main contributions are summarized as follows,

- So far as we know, our proposed paradigm of DSR is the first work that learns with multiple modalities as inputs at training stage, but tests on only single LR depth modality.
- A cross-task distillation scheme is proposed to encourage DSRNet and DENet to learn from each other in a collaborative training mode.
- A structure prediction network is advanced to provide structure regularization for helping DSRNet resolve the problem of structural inconsistency.

2. Related Work

Depth Super-Resolution. Compared to single DSR methods [5, 32, 43, 35], color-guided DSR methods [18, 12, 40, 49] have been widely proposed to improve the quality of depth map by the guidance of color image. Li *et al.* [26] proposed a joint filtering approach that leverages color image as guidance to enhance the spatial resolution of depth map. Hui *et al.* [18] employed a multi-scale fusion strategy that fuses the rich hierarchical color features at different levels to resolve ambiguity in DSR. Wen *et al.* [41] presented a data-driven filter to infer an initial HR depth map with the guidance of color image, then proposed a coarse-to-fine network to progressively recover the depth map. Guo *et al.* [12] proposed a hierarchical feature driven method that constructs an input pyramid and a guidance pyramid for multi-level residual learning. Wang *et al.* [40] proposed to upsample the depth map with the help of edge map learned from

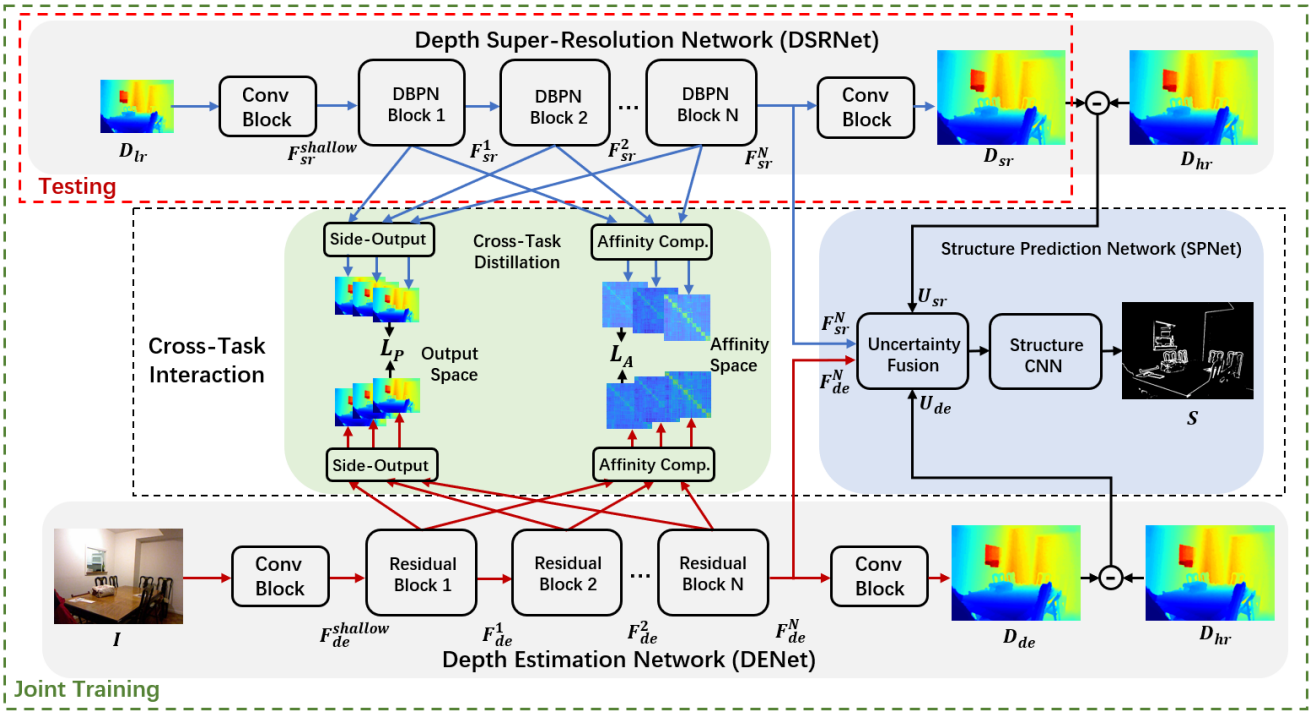


Figure 2. Illustration of our proposed framework, which consists of DSRNet, DENet, and the middle cross-task interaction module. We supervise the outputs of DSRNet and DENet with the same groundtruth depth map D_{hr}. In testing phase, DSRNet is the final choice to predict HR depth map from only LR depth map without the help of color image.

the color image.

Monocular Depth Estimation. Due to the strong ability of CNN in feature extraction, many supervised monocular depth estimation methods [45, 10, 42, 23] continue to improve the performance of depth estimation. Laina *et al.* [24] proposed a fully convolutional architecture to model the mapping between color image and depth map. In [25], a two-stream CNN is proposed to simultaneously predict depth and depth gradients for accurate depth estimation. Wang *et al.* [38] presented a depth estimation network via a semantic divide-and-conquer strategy, in which a scene is decomposed into semantic segments and then predicts depth for each segment. In contrast, unsupervised methods [8, 10, 42, 51] use video or stereo data during training without the need of groundtruth depth maps. Wong *et al.* [42] learned a robust representation with a two-branch decoder to estimate the depth map. Different from the above works, we introduce an auxiliary monocular depth estimation task for achieving the cross-model knowledge learning at training stage.

Knowledge Distillation. Knowledge distillation [16] is to transfer knowledge from high-capacity model to a compact model to improve the performance of the latter one. It has been wildly applied to many applications, including action recognition [7], style transfer [28], depth estimation [13] and scene parsing [44]. For example, the task of image classification takes class probabilities from teacher network as soft targets to train the student network [37, 16] or transfers the knowledge through intermediate layers [50, 33]. Recently, deep mutual learning [52] proposed a two-way distillation which transfers knowledge between the teacher and student and benefits to both networks. Kundu *et al.* [22] tried to extend cross-model distillation to multiple spatially-structured prediction tasks by using two regularization strategies to minimize the domain discrepancy. Yao *et al.* [46] presented a dense cross-layer mutual distillation mechanism to train the teacher and student collaboratively from scratch. Inspired by the above knowledge distillation techniques, we propose to learn the scene structure guidance for the single DSR task via our designed cross-task distillation. Moreover, another difference to the previous works is that the role of teacher and student is dynamically changed.

3. Method

3.1. Network Architecture

Figure 2 shows the overall architecture of our framework, which mainly consists of three components: depth super-resolution network (DSRNet), depth estimation network (DENet) and the middle cross-task interaction module. Given a collection of paired LR-HR depth maps $\{D_{lr}^{(k)}, D_{hr}^{(k)}\}_{k=1}^{M}$ with the corresponding HR color images $\{I^{(k)}\}_{k=1}^{M}$ as training data, where M is the number of training data, our goal is to learn a model, i.e., DSRNet, that

can predict the super-resolved depth maps $\{D_{sr}^{(k)}\}_{k=1}^{M}$ from their corresponding downsampled versions $\{D_{lr}^{(k)}\}_{k=1}^{M}$.

Specifically, the structure of DSRNet is designed based on a network unit from deep back-projection network (DBPN) [14], which can effectively improve the feature representations through iterative projecting HR representations to LR spatial domain and then mapping back into HR domain[1]. The shallow features $F_{sr}^{shallow}$ are first extracted from D_{lr} through a simple convolutional block (including three convolutional layers), and then sent into N stacked DBPN blocks to obtain HR features $\{F_{sr}^{n}\}_{n=1}^{N}$. D_{sr} is finally reconstructed from F_{sr}^{N} through another convolutional block. DENet takes I as input and estimates the depth map D_{de}. The architecture of DENet is similar to DSRNet, but replaces the DBPN blocks with deeper residual blocks [15] to extract informative features $\{F_{de}^{n}\}_{n=1}^{N}$ from color image.

The cross-task interaction module acts as a bridge to connect DSRNet and DENet, and realizes bilateral knowledge transfer between them. It consists of two components, i.e., a cross-task distillation scheme and a structure prediction network (SPNet), where the former focuses on the interaction between multi-scale features extracted from both networks while the latter uses structure maps as supervision to further guide the learning of both networks.

Note that, at the training stage, DSRNet, DENet and the cross-task interaction module are jointly learned by using both color image and depth map as input. In testing phase, DSRNet is the final choice to predict HR depth map from only LR depth map without the help of color image.

3.2. Cross-Task Distillation

Knowledge distillation is generally viewed as a technique of transferring beneficial information from a top-performing model to the other naive one. Different from the commonly used distillation techniques, in which the teacher network is trained beforehand and fixed under the assumption that it always learns a better representation than the student network, our goal is to train SRNet and DENet collaboratively and encourage them to benefit from each other.

Inspired by mutual learning methods [52, 22], we propose a cross-task distillation scheme, in which the roles of teacher and student will exchange between both tasks based on their current performances on depth recovery in the iterative collaborative training. Specially, at the current round of training, we need to determine the teacher in advance according to their performance at the previous round. We compute the average pixel error between each recovered depth map and its groundtruth for both networks:

$$e_{dsr} = \frac{1}{HW}\sum_{h}^{H}\sum_{w}^{W}|D_{sr}(h,w) - D_{hr}(h,w)|, \quad (1)$$

[1]We direct readers to refer to [14] for more details about the design of DBPN block.

$$e_{de} = \frac{1}{HW}\sum_{h}^{H}\sum_{w}^{W}|D_{de}(h,w) - D_{hr}(h,w)|, \quad (2)$$

where $\{H, W\}$ are the size of the output depth map. If e_{dsr} is smaller than e_{de}, DSRNet has a relatively better performance, and becomes the dominant one to guide the learning of DENet, and vice versa.

Next, in order to distill more meaningful knowledge that can accurately describe the essential structural characteristics of depth map, we introduce a multi-space distillation scheme to condense the knowledge from the perspectives of output and affinity spaces, as shown in Figure 2.

Output Space Distillation. To ensure the transfer of local information from pixel-wise depth values in a depth map, we apply the side-output layer (containing two successive convolutions) on $\{F_{sr}^{n}, F_{de}^{n}\}_{n=1}^{N}$ from both DSRNet and DENet to generate the corresponding multi-scale depth outputs $\{D_{sr}^{n}, D_{de}^{n}\}_{n=1}^{N}$ respectively. Thus, the distillation loss of output space is designed to indirectly align the features between DSRNet and DENet:

$$\mathcal{L}_O = \frac{1}{N}\sum_{i=1}^{N}||D_{sr}^{i} - D_{de}^{i}||_1, \quad (3)$$

Affinity Space Distillation. Color image and its associated depth map are different representations of the same scene and have strong structural similarity. Pixels with similar appearances in a color image have more chances of belonging to the same object, and should have close depth values. Inspired by [39, 6, 47] that consider the nonlocal correlations to strengthen correlated features between pixels and benefit the depth map recovery, we also transfer non-local structure knowledge on affinity space, which is implemented by computing pair-wise similarities between pixels.

Assuming the dimension of feature F is $w \times h \times c$, the reshape function $\mathbb{R}$ recasts F as $\mathbb{R}(F)$ with the dimension of $wh \times c$. The affinity matrix A is defined as:

$$A(F) = \sigma(\mathbb{R}(F) \otimes \mathbb{R}^T(F)), \quad (4)$$

where $\sigma(\cdot)$ is the softmax operation, $\otimes$ is the matrix multiplication and T is the transpose operator. The distillation loss of affinity space is defined as the following,

$$\mathcal{L}_A = \frac{1}{N}\sum_{i=1}^{N}||A(F_{sr}^{i}) - A(F_{de}^{i})||_1, \quad (5)$$

The final distillation loss $\mathcal{L}_{distill}$ is expressed as:

$$\mathcal{L}_{distill} = \mathcal{L}_O + \gamma\mathcal{L}_A. \quad (6)$$

where γ is an adjustment parameter. Note that, $\mathcal{L}_{distill}$ should be imposed on the training of the student, but not the teacher, which are determined by the errors comparison between e_{dsr} and e_{de}.

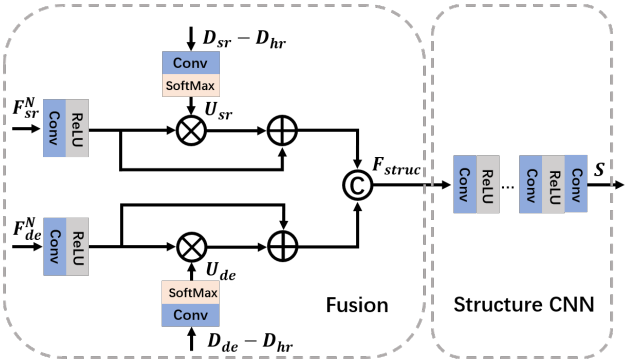


Figure 3. The proposed SPNet. We fuse F_{sr}^N and F_{de}^N by our proposed uncertainty-induced attention fusion module.

3.3. Structure Prediction

The goal of SPNet is to predict a structure map S from the last feature maps F_{sr}^N and F_{de}^N generated by DSRNet and DENet respectively. Through the supervision with the groundtruth structure map S_{gt}[2], SPNet can provide extra structure regularization to help both DSRNet and DENet learn more informative structure representations to alleviate the problem of RGB-D structure inconsistency. As shown in Figure 3, SPNet consists of a fusion module and a structure CNN, where the latter is a lightweight network with five stacked 'Conv+ReLU' layers and a last 'Conv' layer.

Usually, the erroneous recovery of DSR and DE tasks occurs in the regions around depth boundaries and fine structures in a depth map, which are subject to higher recovery uncertainty. Therefore, instead of simply concatenating F_{sr}^N and F_{de}^N and sending into the structure CNN, we design an uncertainty-induced attention fusion module to strengthen these informative structure features by attending the recovery uncertainty to the feature map. Thus, we first compute the uncertainty maps U_{sr} and U_{de} of both networks by activating the recovery errors:

$$U_{sr} = \sigma(Conv_{1\times1}(D_{sr} - D_{hr})) \tag{7}$$

$$U_{de} = \sigma(Conv_{1\times1}(D_{de} - D_{hr})) \tag{8}$$

where $Conv_{1\times1}$ is the 1×1 convolution to adjust the channels. Then, we use the uncertainty maps to re-weight F_{sr}^N, F_{de}^N and fuse them through an attention module:

$$F_{struc} = [F_{sr}^N * (1 + U_{sr}), F_{de}^N * (1 + U_{de})] \tag{9}$$

where F_{struc} is the fused features for structure prediction. $[\cdot]$ denotes concatenation operation and $*$ is element-wise multiplication. Note that, through back-propagating the network gradients from SPNet in the backward information flow, the parameters of DSRNet and DENet can be updated.

[2] S_{gt} is obtained by computing the difference between adjacent pixels (gradients) in D_{HR}. Following [30], we calculate it by a convolution layer with a fixed kernel to extract the high-frequency parts from D_{HR}.

Algorithm 1 Training Details

Input: Training data D_{lr}, D_{hr}, I, S_{gt}

———————— Step 1 ————————
Randomly initialize DSRNet and DENet
for i = 1; i $\leq$ 100 **do**
    Train DSRNet and DENet with $\mathcal{L}_{DSR}$ and $\mathcal{L}_{DE}$, respectively
———————— Step 2 ————————
Randomly initialize SPNet
for i = 101; i $\leq$ max epoch **do**
    Compute the average error value e_{dsr} and e_{de} according to Eq.(1) and Eq.(2)
    if $e_{dsr} \leq e_{de}$ **then**
        Fix DSRNet and update DENet with
        $\mathcal{L} = \mathcal{L}_{DE} + \rho_1\mathcal{L}_{struc} + \rho_2\mathcal{L}_{distill}$
    else
        Fix DENet and update DSRNet with
        $\mathcal{L} = \mathcal{L}_{DSR} + \rho_1\mathcal{L}_{struc} + \rho_2\mathcal{L}_{distill}$

Output: D_{sr}

3.4. Training Algorithm

The training process of our framework can be divided into two steps, as presented in Algorithm 1. First, we separately train DSRNet and DENet with the groundtruth D_{hr}. The losses are defined as follows:

$$\mathcal{L}_{DSR} = ||D_{sr} - D_{hr}||_1, \tag{10}$$

$$\mathcal{L}_{DE} = \lambda\frac{1 - SSIM(D_{de}, D_{hr})}{2} + (1-\lambda)||D_{de} - D_{hr}||_1, \tag{11}$$

where $\mathcal{L}_{DSR}$ is a common pixel-wise L1 loss for the task of DSR. Following [10], $\mathcal{L}_{DE}$ is set as a combination of the reconstruction loss (L1 loss) and structural similarity (SSIM). λ is an adjustment parameter.

Then, we introduce the cross-task distillation between both networks with the loss $\mathcal{L}_{distill}$ in Eq. (6). At the same time, we randomly initialize SPNet, and train it together with DSRNet and DENet. The loss for SPNet is defined as:

$$\mathcal{L}_{struc} = ||\mathbb{G}(F_{struc}) - S_{gt}||_1, \tag{12}$$

where $\mathbb{G}(\cdot)$ denotes SPNet, F_{struc} is the fused structure feature in Eq.(9) and S_{gt} is the ground truth of the structure. If DSRNet is chosen as the student, the parameters of DENet are fixed at the current epoch, and DSRNet is updated with the following loss:

$$\mathcal{L} = \mathcal{L}_{DSR} + \rho_1\mathcal{L}_{struc} + \rho_2\mathcal{L}_{distill}, \tag{13}$$

where ρ_1, ρ_2 are the trade-off parameters. Otherwise, DSRNet is fixed, and DENet is updated with the loss $\mathcal{L}$ by replacing $\mathcal{L}_{DSR}$ with $\mathcal{L}_{DE}$.

Table 1. Quantitative DSR results (in MAD) on Middlebury 2005 dataset. 'DSRNet w/o CT' and 'DSRNet' denote the results of the proposed method without and with cross-task interaction scheme, respectively.

	Art				*Books*				*Dolls*				*Laundry*				*Moebius*				*Reindeer*			
	×2	×4	×8	×16	×2	×4	×8	×16	×2	×4	×8	×16	×2	×4	×8	×16	×2	×4	×8	×16	×2	×4	×8	×16
Bicubic	0.48	0.97	1.85	3.59	0.13	0.29	0.59	1.15	0.20	0.36	0.66	1.18	0.28	0.54	1.04	1.95	0.13	0.30	0.59	1.13	0.30	0.55	0.99	1.88
DJF [26]	0.12	0.40	1.07	2.78	0.05	0.16	0.45	1.00	**0.06**	0.20	0.49	0.99	0.07	0.28	0.71	1.67	0.06	0.18	0.46	1.02	0.07	0.23	0.60	1.36
MSG [18]	-	0.46	0.76	1.53	-	0.15	0.41	0.76	-	0.25	0.51	0.87	-	0.30	0.46	1.12	-	0.21	0.43	0.76	-	0.31	0.52	0.99
DGDIE [11]	0.20	0.48	1.20	2.44	0.14	0.30	0.58	1.02	0.16	0.34	0.63	0.93	0.15	0.35	0.86	1.56	0.14	0.28	0.58	0.98	0.16	0.35	0.73	1.29
DEIN [48]	0.23	0.40	0.64	**1.34**	0.12	0.22	0.37	0.78	0.12	0.22	0.38	0.73	0.13	0.23	0.36	0.81	0.11	0.20	0.35	0.73	0.15	0.26	0.40	0.80
CCFN [41]	-	0.43	0.72	1.50	-	0.17	0.36	0.69	-	0.25	0.46	0.75	-	0.24	0.41	**0.71**	-	0.23	0.39	0.73	-	0.29	0.46	0.95
GSRPT [4]	0.22	0.48	0.74	1.48	0.11	0.21	0.38	0.76	0.13	0.28	0.48	0.79	0.12	0.33	0.56	1.24	0.12	0.24	0.49	0.80	0.14	0.31	0.61	1.07
DSRN [40]	0.12	0.25	0.61	1.80	**0.04**	**0.11**	0.28	0.69	**0.06**	**0.14**	**0.33**	0.73	**0.06**	**0.15**	0.43	1.24	**0.05**	**0.13**	0.29	**0.67**	**0.07**	**0.15**	**0.35**	0.92
DSRNet w/o CT	0.16	0.31	0.59	1.55	0.10	0.15	0.31	0.73	0.12	0.21	0.39	0.69	0.12	0.21	0.40	0.82	0.11	0.16	0.32	0.74	0.13	0.22	0.38	0.87
DSRNet	**0.11**	**0.25**	**0.53**	1.44	0.05	**0.11**	**0.26**	**0.67**	0.07	0.16	0.36	**0.65**	**0.06**	0.16	**0.36**	0.76	0.07	**0.13**	**0.27**	0.69	0.08	0.17	**0.35**	**0.77**

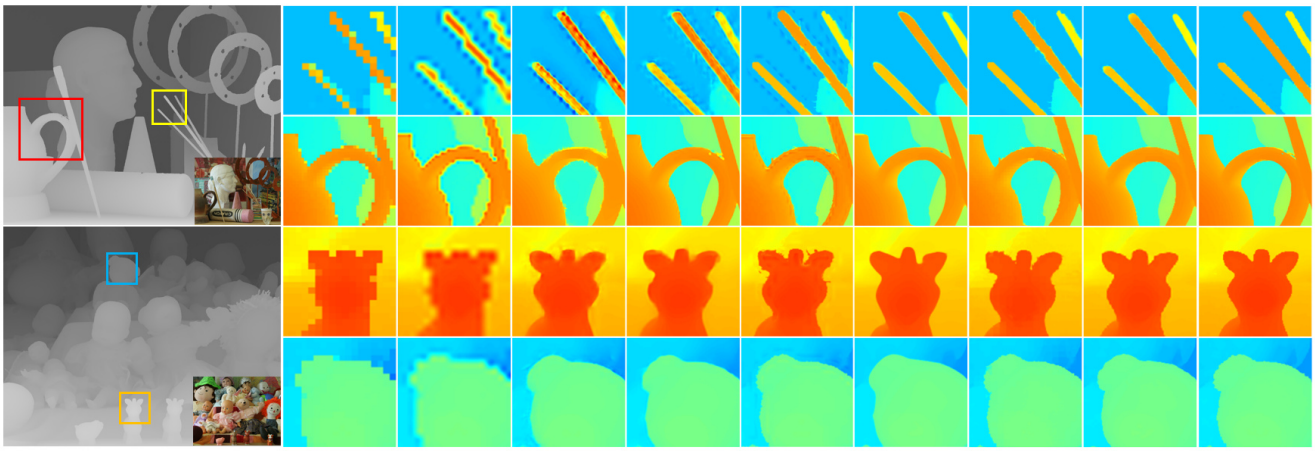


Figure 4. Visual comparison of ×8 DSR results on *Art* and *Dolls* in Middlebury: (a) Groundtruth depth maps, (b) LR patches, results from (c) Bicubic, (d) DJF [26], (e) DGDIE [11], (f) DEIN [48], (g) GSRPT [4], (h) DSRN [40], (i) Ours, and (j) Groundtruth.

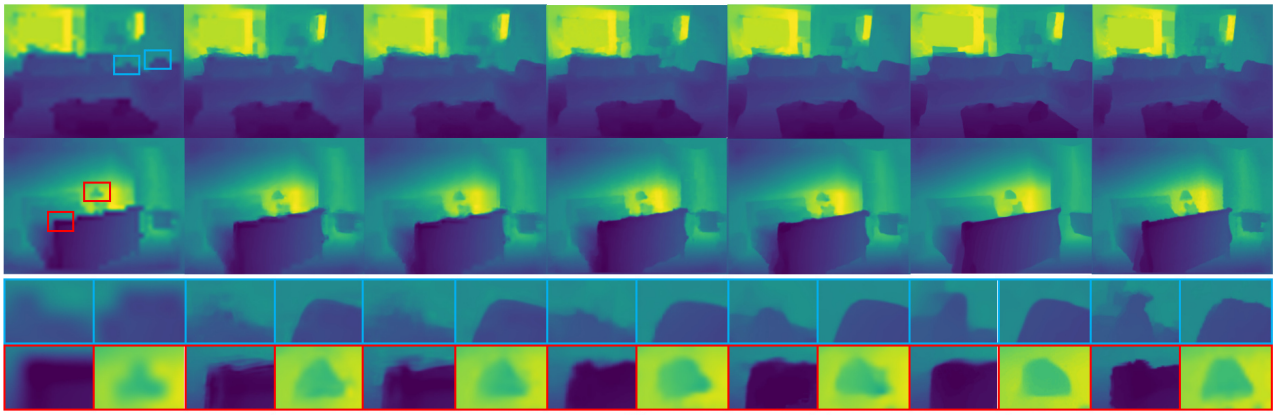
4. Experiments

We conduct experiments on four datasets, i.e., Middlebury [17], MPI Sintel [3], NYUv2 [34] and ToFMark [5]. First, following MSG [18], we use 34 RGB-D images from Middlebury dataset [17] and 58 RGB-D images from MPI Sintel dataset [3] as training data. To evaluate our performance, we test on Middlebury 2005 (6 standard RGB-D images: *Art, Books, Moebius, Dolls, Laundry and Reindeer*). We also evaluate the generalization performance on ToFMark dataset (3 real-world depth maps captured by Time of Flight (ToF) sensor). Another training and testing dataset is NYU v2 dataset captured by Kinect sensor. Following the widely used data splitting manner, we use the first 1000 RGB-D images for training and the rest 449 RGB-D images for evaluation. For both experiment settings, we randomly extract 10000+ patches of fixed size of 256×256 from HR depth maps and downsample HR depth maps by bicubic interpolation to get LR inputs. We augment the training data by 180° rotation. We choose Mean Absolute Difference (MAD) and Root Mean Square Error (RMSE) as the evaluation metrics. Lower MAD and RMSE values indicate higher recovery quality.

During training, we set the number of DBPN and Residual blocks as $N = 5$. We set the trade-off parameters as $\gamma = 0.5$, $\lambda = 0.2$, $\rho_1 = 0.1$ and $\rho_2 = 0.1$. For optimization, we use the Adam optimization algorithm with momentum $= 0.9$, $\beta_1 = 0.9$, $\beta_2 = 0.99$ and $\epsilon = 10^{-8}$ to train our models. The initial learning rate is set to 1e-3 and decreased by multipling by 0.1 for every 50 epochs. We implemented our method using PyTorch with one RTX 2080Ti GPU.

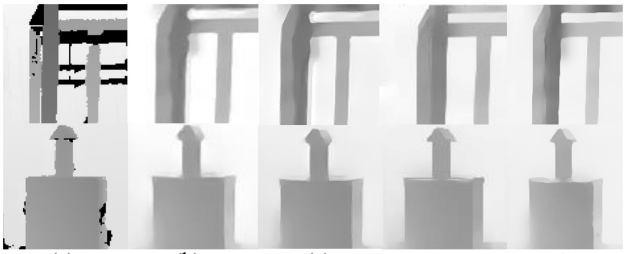
4.1. Comparison with State-of-The-Art

Middlebury Dataset. We compare our method with state-of-the-art (SOTA) DSR methods on Middlebury under four different up-scaling factors (×2, ×4, ×8 and ×16). All the compared methods and ours are deep learning based methods, which are trained and tested under the same conditions for fair comparison. Table 1 summarizes the quantitative results on the Middlebury dataset. 'DSRNet w/o CT' and 'DSRNet' denote the results of the proposed method without and with cross-task interaction scheme, respectively. Benefiting from the backbone of DBPN network, the performance of 'DSRNet w/o CT' goes beyond most of the previous methods, but slightly inferior to the recent SOTA



(a) (b) (c) (d) (e) (f) (g)

Figure 5. Visual comparison of $\times 16$ DSR results on NYU v2 dataset. (a) LR depth maps; Results from (b) DJF [26], (c) DJFR [27], (d) P2P [19], (e) GbFT [1], (f) Ours, and (g) Groundtruth.



(a) (b) (c) (d) (e)

Figure 6. Visualization on *Shark* in ToFMark dataset. (a) Groundtruth patches, Results from (b) DGDIE[11], (c)GSRPT[4], (d) DSRN[40], and (e) Ours.

Table 2. Quantitative DSR results (in RMSE) on NYUv2 dataset.

	DJF [26]	DJFR [27]	DGDIE [11]	GbFT [1]	P2P [19]	PAC [36]	DKN [21]	Ours
$\times 4$	3.54	3.38	1.56	3.35	4.12	2.39	1.62	**1.49**
$\times 8$	6.20	5.86	2.99	5.73	6.48	4.59	3.26	**2.73**
$\times 16$	10.21	10.11	5.24	9.01	10.17	8.09	6.51	**5.11**

Table 3. Quantitative DSR results (in MAD) on ToFMark dataset.

	MSG [18]	DEIN[48]	DGDIE [11]	GSRPT [4]	DSRN [40]	Ours
Books	12.26	12.78	12.31	13.21	11.15	**11.03**
Shark	14.11	15.11	14.06	15.03	13.26	**13.08**
Devil	12.45	14.25	9.66	12.27	9.54	**9.33**

methods, e.g., DSRN. When training together with DENet assisted by cross-task interaction, the performance of our DSRNet steps further compared to 'DSRNet w/o CT', and is comparable to DSRN, even better than it on at least half of the cases. Note that, different from these color-guided methods, we stand on single DSR without the help of color image in testing phase, but achieves satisfactory results on both accuracy and runtime (evaluated later in this section). The $\times 8$ DSR results are visually shown in Figure 4. For the structural details, e.g., the stick and the teapot handle in *Art*, and the toy's contour in *Dolls*, we clearly recover these regions without introducing the texture copying artifacts thanks to our structure prediction network (SPNet).

NYU Dataset. We further evaluate our method on NYU-v2 dataset compared with other SOTA methods in Table 2. Our method performs the best on all the cases with different up-scaling factors. Figure 5 presents visual comparisons under the $\times 16$ DSR case. Observing from the recovered depth maps and enlarged patches, our method achieves the clearest results and preserves the sharpest and correct object contours, which demonstrates our effectiveness to learn the information of scene structure guidance from color image.

ToFMark Dataset. We evaluate the generalization ability of our method on ToFMark dataset. Following DGDIE [11], we fill the missing points in depth maps and downsample them by $\times 2$ downsampling factors, then send them to the $\times 2$ model trained on Middlebury dataset to acquire the generalization results. As shown in Table 3, our method obtains the best objective generalization results for all three test examples. Figure 6 shows the visual comparison on *Shark*. Our method has higher visual quality and less blur than others, especially at the boundary regions.

Runtime. In Figure 8, we summarize the overall performance by the tradeoff between accuracy and runtime. We measure the runtime for $\times 8$ DSR to their full resolution (about 1080×1320) on Middlebury dataset. The color-guided methods i.e., JGF [29] MSG [18], and DSRN [40], run slower than ours due to the usage of HR color images in testing phase. MS is the single DSR version of MSG, but still inferior to ours for both accuracy and speed. Owing to the cross-task interaction, we achieve satisfactory recovered results with minimum inference time.

4.2. Ablation Study

In this section, we further verify the key designs of our cross-task interaction, i.e., cross-task distillation and structure prediction task by ablation study.

We report the $\times 8$ DSR results on Middlebury and NYU-v2 dataset under different experimental settings, as shown

Figure 7. Visual analysis of the fusion process. (a) RGB-D pairs, Output features from (b) DSRNet and (c) DENet, Re-weighted features (d) and (e) corresponding to (b) and (c), respectively, (f) Fused features, (h) Predicted structure map; (i) Groundtruth structure map.

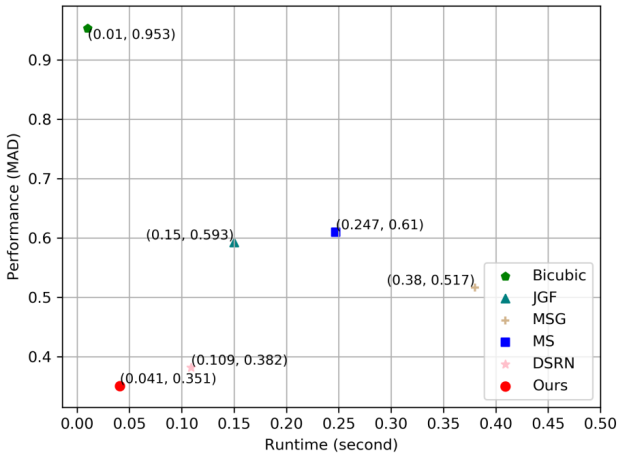


Figure 8. Runtime and performance analysis.

Table 4. The impact of each component in cross-task distillation.

DSRNet	DENet	$\mathcal{L}_O$	$\mathcal{L}_A$	SPNet	Middlebury	NYUv2
✓					0.398	3.13
	✓				0.419	3.26
✓	✓	✓			0.377	2.88
✓	✓		✓		0.384	2.93
✓	✓	✓	✓		0.372	2.85
✓	✓	✓	✓	✓	**0.356**	**2.75**

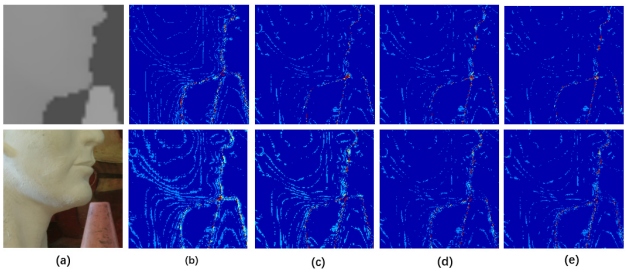


Figure 9. Visualization of recovery errors variation in DSRNet at different distillation stage. (a) LR depth and HR RGB patches; (b) Error maps w/o cross-task distillation; Error maps with cross-task distillation after (c) $140th$ and (d) $180th$ training epoches; (e) Error maps with cross-task distillation and structure prediction.

in Table 4. The 1^{st} and 2^{nd} rows are the objective results of initial DSRNet and DENet trained separately without cross-task interaction scheme. DENet performs relatively inferior to DSRNet, because estimating depth map from only color image is more difficult than from the degraded LR version. Next, $\mathcal{L}_S$ and $\mathcal{L}_A$ distill knowledge from output space and affinity space respectively. When gradually integrating each loss into the baseline, it brings a progressive performance improvement (6.5% and 8.9% increase on Middlebury and NYU respectively), which verifies the effectiveness of cross-task distillation strategy is remarkable. Besides, we also visualize the error maps between recovered depth map and groundtruth to validate the process of our iterative collaborative training in Figure 9. As the epoch goes, the recovery errors for both DSRNet and DENet are decreased at the boundary regions. Finally, after introducing the SPNet (the final case in Table 4), the values of both error metrics are further decreased (4.3% error reduction against the previous case). Figure 9(e) also validates this from visual performance.

Table 5. Ablation study of uncertainty-induced attention fusion module (abbreviated as U) on $\times 8$ DSR cases. The numerical results represent the MADs computed between S and S_{gt}.

	Art	*Books*	*Dolls*	*Laundry*	*Moebius*	*Reindeer*
SPNet w/o U	1.403	2.292	1.730	1.846	1.379	1.565
SPNet w/ U	**1.386**	**2.259**	**1.715**	**1.821**	**1.349**	**1.552**

Besides, we conduct ablation study to evaluate the effectiveness of our uncertainty-induced attention fusion module as presented in Table 5. Compared to simply concatenating features in channel dimension and sending into SPNet (SPNet w/o U), our proposed fusion module can offer significant assistance for SPNet to predict the structure map. We also visualize the feature maps in the fusion process in Figure 7. The object boundaries, e.g., the regions around the toy head and the strip behind, are obviously highlighted by the uncertainty map, which facilitate the final structure perception of SPNet.

5. Conclusion

For the first time, we explored to learn the cross-modal knowledge from both RGB and depth modalities at training stage, but test on only single depth modality. A cross-task interaction module is advanced to realize bilateral knowledge transfer between DSRNet and the auxiliary DENet in a well-designed collaborative training mode. Experiments show our method's superior performance for both accuracy and runtime. In the future work, we may extend our cross-task interaction to more guided image restoration tasks.

References

[1] Badour Albahar and Jia-Bin Huang. Guided image-to-image translation with bi-directional feature transformation. In *IEEE ICCV*, pages 9015–9024, 2019.

[2] Guido Borghi. Combining deep and depth: Deep learning and face depth maps for driver attention monitoring. *CoRR*, abs/1812.05831, 2018.

[3] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625, Oct. 2012.

[4] Riccardo de Lutio, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. Guided super-resolution as pixel-to-pixel transformation. In *IEEE International Conference on Computer Vision, ICCV*, pages 8828–8836, 2019.

[5] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proc. ICCV*, 2013.

[6] Yukang Gan, Xiangyu Xu, Wenxiu Sun, and Liang Lin. Monocular depth estimation with affinity, vertical pooling, and label enhancement. In *European Conference on Computer Vision (ECCV)*, 2018.

[7] Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Computer Vision ECCV 15th European Conference*, volume 11212, pages 106–121, 2018.

[8] Ravi Garg, B. G. Vijay Kumar, Gustavo Carneiro, and Ian D. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *Computer Vision ECCV 14th European Conference*, volume 9912, pages 740–756, 2016.

[9] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 963–968, 2011.

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 6602–6611, 2017.

[11] Shuhang Gu, Wangmeng Zuo, Shi Guo, Yunjin Chen, Chongyu Chen, and Lei Zhang. Learning dynamic guidance for depth image enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 712–721, 2017.

[12] Chunle Guo, Chongyi Li, Jichang Guo, Runmin Cong, Huazhu Fu, and Ping Han. Hierarchical features driven residual learning for depth map super-resolution. *IEEE Trans. Image Processing*, 28(5):2545–2557, 2019.

[13] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy S. J. Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *Computer Vision ECCV 15th European Conference*, volume 11215, pages 506–523, 2018.

[14] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1664–1673, 2018.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 770–778, 2016.

[16] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[17] Heiko Hirschmüller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, 2007.

[18] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *Computer Vision ECCV 14th European Conference*, pages 353–369, 2016.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5967–5976, 2017.

[20] Christian Kerl, Jürgen Sturm, and Daniel Cremers. Dense visual SLAM for RGB-D cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2100–2106, 2013.

[21] Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for guided depth map upsampling. *CoRR*, abs/1903.11286, 2019.

[22] Jogendra Nath Kundu, Nishank Lakkakula, and Venkatesh Babu Radhakrishnan. Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In *IEEE International Conference on Computer Vision, ICCV*, pages 1436–1445, 2019.

[23] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R. Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2656–2665, 2018.

[24] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *Fourth International Conference on 3D Vision, 3DV 2016*, pages 239–248, 2016.

[25] Jun Li, Can Yuce, Reinhard Klein, and Angela Yao. A two-streamed network for estimating fine-scaled depth maps from single RGB images. *Comput. Vis. Image Underst.*, 186:25–36, 2019.

[26] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *Computer Vision - ECCV - 14th European Conference*, pages 154–169, 2016.

[27] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Joint image filtering with deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1909–1923, 2019.

[28] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI, Melbourne*, pages 2230–2236, 2017.

[29] Ming-Yu Liu, Oncel Tuzel, and Yuichi Taguchi. Joint geodesic upsampling of depth images. In *IEEE Conference*

on Computer Vision and Pattern Recognition, CVPR, pages 169–176, 2013.

[30] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 7766–7775, 2020.

[31] Jinshan Pan, Jiangxin Dong, Jimmy S. J. Ren, Liang Lin, Jinhui Tang, and Ming-Hsuan Yang. Spatially variant linear representation models for joint filtering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1702–1711, 2019.

[32] Gernot Riegler, Matthias Rüther, and Horst Bischof. Atgv-net: Accurate depth super-resolution. In *Computer Vision ECCV 14th European Conference*, pages 268–284, 2016.

[33] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *3rd International Conference on Learning Representations, ICLR*, 2015.

[34] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Computer Vision ECCV 12th European Conference on Computer Vision*, volume 7576, pages 746–760, 2012.

[35] Xibin Song, Yuchao Dai, and Xueying Qin. Deeply supervised depth map super-resolution as novel view synthesis. *IEEE Trans. Circuits Syst. Video Techn.*, 29(8):2323–2336, 2019.

[36] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik G. Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *IEEE CVPR*, pages 11166–11175, 2019.

[37] Gregor Urban, Krzysztof J. Geras, Samira Ebrahimi Kahou, Özlem Aslan, Shengjie Wang, Abdelrahman Mohamed, Matthai Philipose, Matthew Richardson, and Rich Caruana. Do deep convolutional nets really need to be deep and convolutional? In *ICLR*, 2017.

[38] Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 538–547, 2020.

[39] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 7794–7803, 2018.

[40] Zhihui Wang, Xinchen Ye, Baoli Sun, Jingyu Yang, Rui Xu, and Haojie Li. Depth upsampling based on deep edge-aware learning. *Pattern Recognition*, 103:107274, 2020.

[41] Yang Wen, Bin Sheng, Ping Li, Weiyao Lin, and David Dagan Feng. Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution. *IEEE Trans. Image Processing*, 28(2):994–1006, 2019.

[42] Alex Wong and Stefano Soatto. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5644–5653, 2019.

[43] Jun Xie, Rogério Schmidt Feris, Shiaw-Shian Yu, and Ming-Ting Sun. Joint super resolution and denoising from a single depth image. *IEEE Trans. Multimedia*, 17(9):1525–1537, 2015.

[44] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 675–684, 2018.

[45] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3917–3925, 2018.

[46] Anbang Yao and Dawei Sun. Knowledge transfer via dense cross-layer mutual-distillation. *CoRR*, abs/2008.07816, 2020.

[47] Xinchen Ye, Shude Chen, and Rui Xu. DPNet: Detail-preserving network for high quality monocular depth estimation. *Pattern Recognition*, 109, 2020.

[48] Xinchen Ye, Xiangyue Duan, and Haojie Li. Depth super-resolution with deep edge-inference network and edge-guided depth filling. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1398–1402, 2018.

[49] Xinchen Ye, Baoli Sun, Zhihui Wang, Jingyu Yang, Rui Xu, Haojie Li, and Baopu Li. Pmbanet: Progressive multi-branch aggregation network for scene depth super-resolution. *IEEE Trans. Image Process.*, 29:7427–7442, 2020.

[50] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *5th International Conference on Learning Representations, ICLR*, 2017.

[51] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian D. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 340–349, 2018.

[52] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 4320–4328, 2018.