Discrete Cosine Transform Network for Guided Depth Map Super-Resolution
SUPPLEMENTARY MATERIALS

Zixiang Zhao[1,2] Jiangshe Zhang[1*] Shuang Xu[1,3*] Zudi Lin[2] Hanspeter Pfister[2]
[1]Xi'an Jiaotong University, Xi'an, China
[2]Harvard University, Cambridge MA, USA
[3]Northwestern Polytechnical University, Xi'an, China
zixiangzhao@stu.xjtu.edu.cn, jszhang@mail.xjtu.edu.cn, xs@nwpu.edu.cn,
{linzudi,pfister}@g.harvard.edu

Abstract

In this documents, we provide additional supplementary information for the paper "Discrete Cosine Transform Network for Guided Depth Map Super-Resolution". This file contains:
(I) The calculation details for Discrete Cosine Transform (DCT) Module which is mentioned in Sec. 3.1.
(II) The detail architecture for each module in DCTNet framework.
(III) Detailed introduction to the training&testing datasets in Sec. 4.1.
(IV) More visual exhibitions for the highlighting edge attention weights, as shown in Sec. 4.2.2 and Fig. 3 (left) in the original paper.
(V) More qualitative comparison results are displayed, including more error maps corresponding to samples in 4×, 8× and 16×upscaling.
(VI) The network architectures for ablation experiment Exp. 1,2,3 and 5, as referred in Sec. 4.4.
(VII) More ablation experiments in Sec. 4.4.
(VIII) Visulization of the common/modality-specific features.
(IX) Representative failure cases to show our model limitaion, as mentioned in Sec. 4.5.

1. Calculation details for DCT Module

First, we review the notations in the original paper, which are illustrated in Tab. 1. If R and $\tilde{L}$ in the same scene are given, H can be obtained by minimizing the following optimization function:

$$\mathcal{F} = \frac{1}{2}\|H - L\|_2^2 + \frac{\lambda}{2}\|\mathcal{L}(H) - \mathcal{L}(\tilde{R}) \circ \mathcal{W}(\tilde{R})\|_2^2, \tag{1}$$

where $\mathcal{L}(\cdot)$ is the Laplacian filter, $\mathcal{W}(\cdot)$ can be regarded as a given threshold function to select the edges useful for GDSR. Note that $\mathcal{W}(\cdot)$ is given in advance. $\circ$ denotes element-wise multiplication and λ is the tuning parameter. The solution can be acquired by $\frac{\partial \mathcal{F}}{\partial H} = 0$, that is

$$\frac{\partial \mathcal{F}}{\partial H} = H - L + \lambda \mathcal{L}\left\{\mathcal{L}(H) - \mathcal{L}(\tilde{R}) \circ \mathcal{W}(\tilde{R})\right\} = 0. \tag{2}$$

Consider that the kernel of Laplacian filter is symmetric, and we have $\mathcal{L}\left\{\mathcal{L}(H)\right\} = \mathcal{L}^2(H)$ and $\mathcal{L}\left\{\mathcal{L}(\tilde{R})\right\}$ can be obtained in the same way. Then Eq. (2) is rewritten as

$$H - L + \lambda \mathcal{L}^2(H) - \lambda \mathcal{L}\left(\mathcal{L}(\tilde{R}) \circ \mathcal{W}(\tilde{R})\right) = 0, \tag{3}$$

*Corresponding authors.

Notation	Description
$R \in \mathbb{R}^{M\times N\times 3}$	the input HR RGB image
$\tilde{L} \in \mathbb{R}^{m\times n}$	the input LR depth map
$H \in \mathbb{R}^{M\times N}$	the HR depth map to be acquired
$\tilde{R} \in \mathbb{R}^{M\times N}$	the Y channel in YCrCb space of R
$L \in \mathbb{R}^{M\times N}$	the upsampled image of $\tilde{L}$
$\Phi_P^R \in \mathbb{R}^{M\times N\times 64}$ (abbreviated as Φ^R)	the feature map of R
$\Phi_P^L \in \mathbb{R}^{M\times N\times 64}$ (abbreviated as Φ^L)	the feature map of L

Table 1. The notations mentioned in the original paper.

and further,

$$H + \lambda\mathcal{L}^2(H) = \lambda\mathcal{L}\left(\mathcal{L}(\tilde{R}) \circ \mathcal{W}(\tilde{R})\right) + L. \tag{4}$$

The right side of Eq. (4) can be abbreviated as E for simplicity, *i.e.*,

$$\lambda\mathcal{L}(\mathcal{L}(\tilde{R}) \circ \mathcal{W}(\tilde{R})) + L \triangleq E. \tag{5}$$

Inspired by [9] which presents a recipe for solving a discretization of Poisson equation numerically by Fourier transform rapidly, we consider Eq. (4) as a discretization of 2D Poisson equation. Here we assume a "reflection padding" extension at the boundary of the image when performing convolution calculations, which makes zero gradients on the image boundary. Thus, PE Eq. (4) is with the *Neumann boundary condition* (NBC). This boundary value problem, *i.e.* PE with the NBC, can hence be solved via *Discrete Cosine Transform* (DCT), and the boundary conditions are successfully matched. Details of the proof can be found in [12] or [9].

Thus, we implement the DCT on both sides of Eq. (4):

$$\mathcal{F}_c(H) + \lambda K^2 \circ \mathcal{F}_c(H) = \mathcal{F}_c(E), \tag{6}$$

where $\mathcal{F}_c(\cdot)$ is the DCT, $K_{ij} = \cos\left(\frac{i-1}{M}\pi\right) + \cos\left(\frac{j-1}{N}\pi\right)$ and $1 \le i \le M, 1 \le j \le N$. Finally, the HR depth image are acquired by:

$$\mathcal{F}_c(H) = \mathcal{F}_c(E) \oslash \left(I + \lambda K^2\right), \tag{7}$$

$$H = \mathcal{F}_c^{-1}\left\{\mathcal{F}_c(E) \oslash \left(I + \lambda K^2\right)\right\}, \tag{8}$$

where $\mathcal{F}_c^{-1}(\cdot)$ is the inverse DCT operation, $\oslash$ is the element-wise division, and I is the identity matrix.

2. The detail architecture for DCTNet

We illustrate the detail architecture for *semi-coupled feature extraction* (SCFE) module, *guided edge spatial attention* (GESA) module and *depth reconstruction* (DR) module of DCTNet framework in Fig. 1.

3. Detailed introduction to datasets

We adpot three widely-used benchmarks (NYU v2 [11], Middlebury [3, 10], Lu [8] datasets), and a new RGBD benchmark dataset (RGBDD dataset [2]) for the guided depth super-resolution task. The preprocessing and separation of NYU v2, Middlebury, and Lu datasets follows [4–6, 13], and that of RGBDD dataset follows [2].

- NYU v2 dataset[1] [11]: it consists of 1449 RGBD image pairs captured by the Microsoft Kinect [15]. We use the first 1,000 images in this dataset for training, and the rest 449 images for testing.
- Middlebury dataset[2] [3, 10]: we use 30 image pairs from 2001-2006 datasets provided by Lu *et al*. [8] for testing.

[1] https://cs.nyu.edu/ silberman/datasets/nyu_depth_v2.html
[2] https://vision.middlebury.edu/stereo/data/

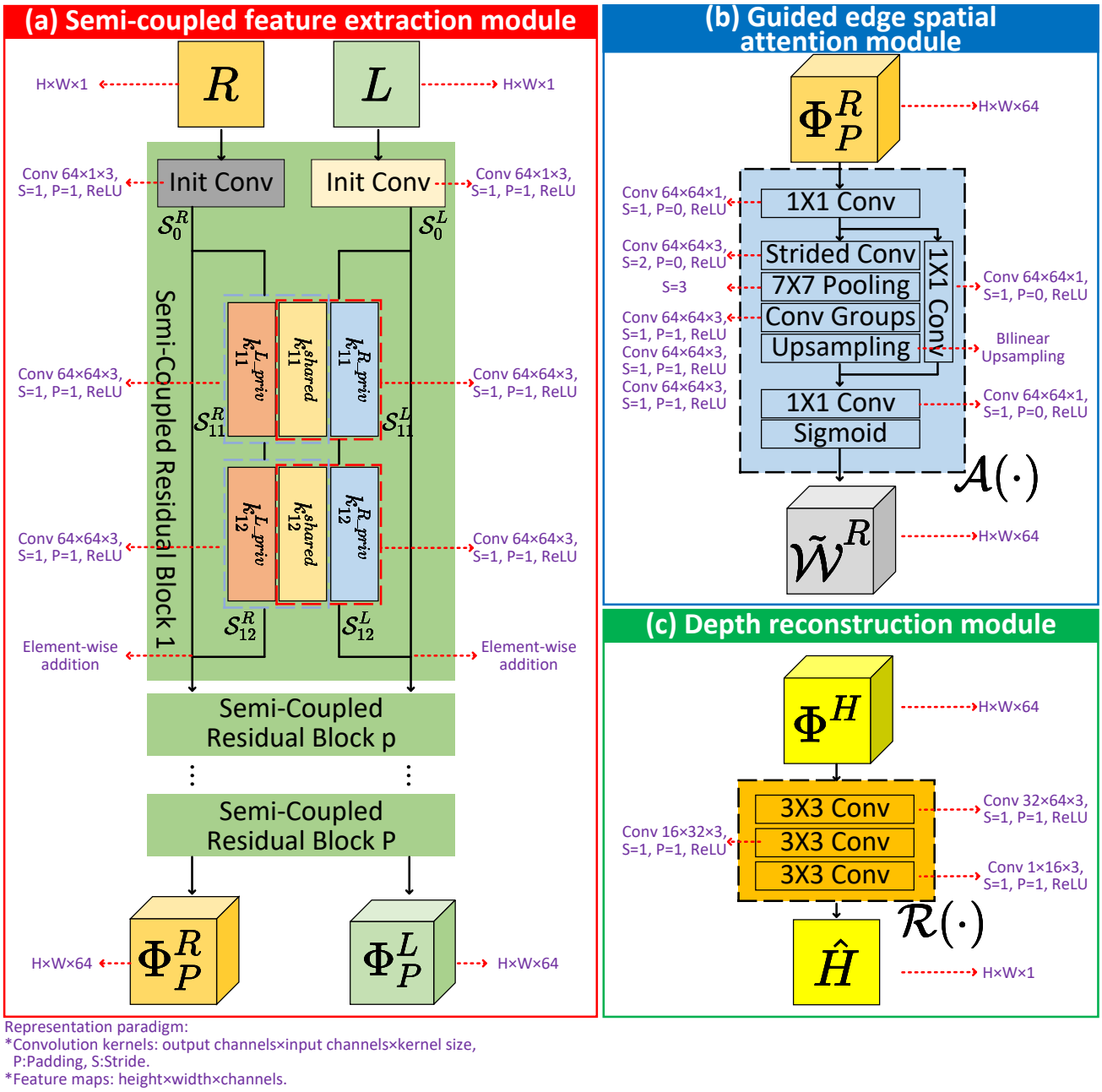
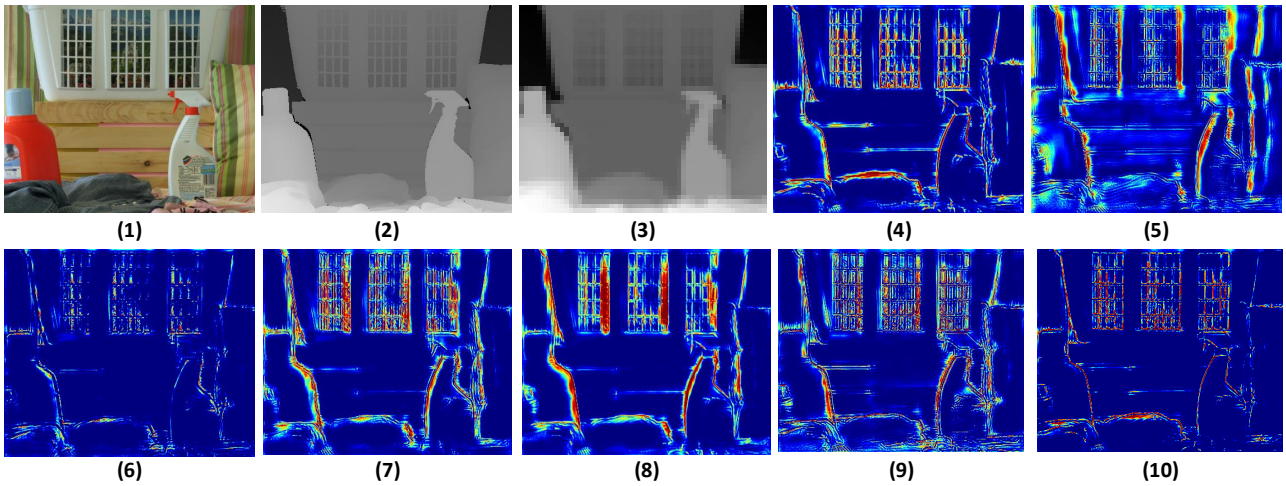


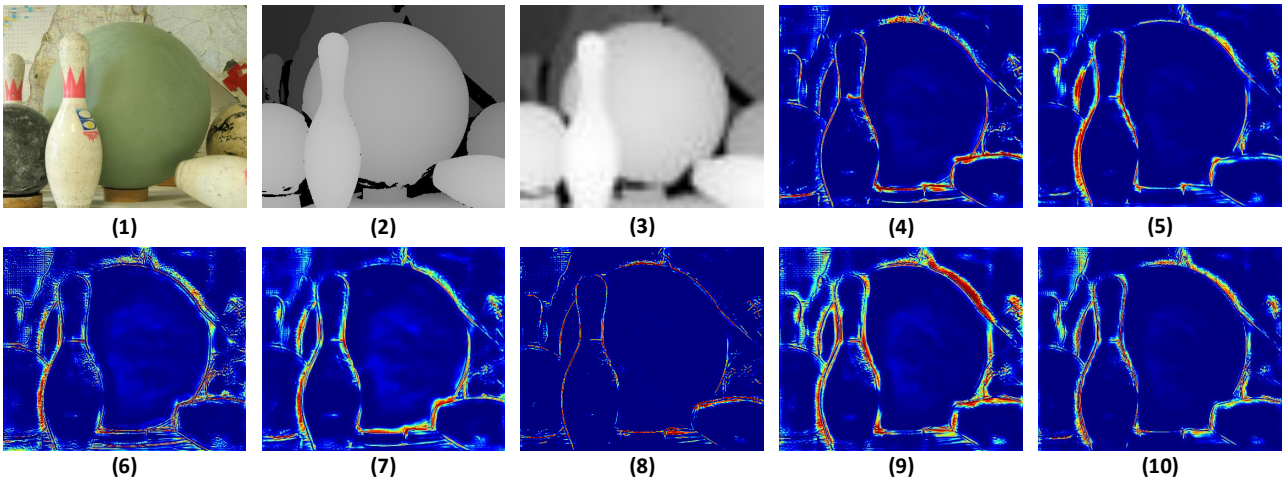
Figure 1. The detail architecture for semi-coupled feature extraction module (a), guided edge spatial attention module (b) and depth reconstruction module (c) in DCTNet framework, which aims to extract cross-modality features, highlight RGB edge information, and reconstruct the HR depth map, respectively.

- Lu dataset[3] [8]: this dataset consists of 6 RGBD image pairs acquired by ASUS Xtion Pro camera. We use it for testing.
- RGBDD dataset[4] [2]: a new RGBD dataset benchmark proposed in CVPR 2021 [2] with four main categories: portraits, models, plants, and lights. The RGB images and LR depth maps are collected by Huawei P30 Pro and the HR depth

[3] http://web.cecs.pdx.edu/ fliu/project/depth-enhance/
[4] http://mepro.bjtu.edu.cn/resource.html



(a) Visual exhibition of highlighting edge attention weights. (1)-(3): Input R, ground truth H and input L of "05-Laundry", respectively. (4)-(10): Representative highlighting edge weights output by GESA module.



(b) Visual exhibition of highlighting edge attention weights. (1)-(3): Input R, ground truth H and input L of "06-Bowling2", respectively. (4)-(10): Representative highlighting edge weights output by GESA module.

Figure 2. Visual exhibitions for the highlighting edge attention weights.

maps are captured by Helios ToF camera[5] produced by LUCID vision labs. In our experiments, 297 portraits, 68 plants, 40 models are utilized for testing. For the *real-world branch*, 1586 portraits, 380 plants, 249 models are for training and the testset is the same as above.

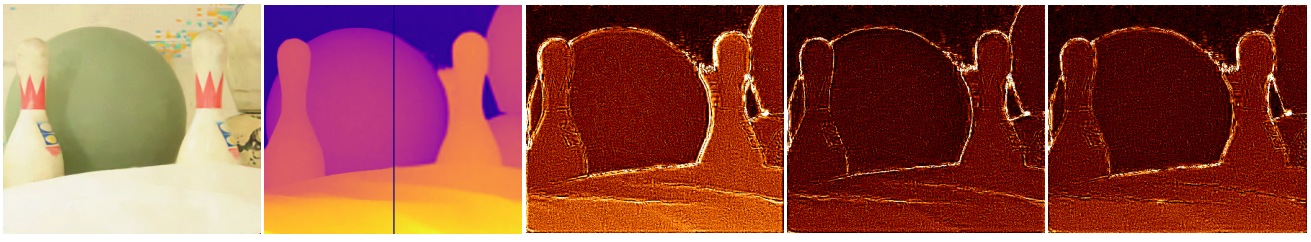
4. Visual exhibitions for the highlighting edge attention weights

We show more visual exhibitions for the highlighting edge attention weights in Fig. 2 (a) and (b), which are the outputs of the GESA module. Obviously, after the attention weight purification of the GESA module, the edge contour of the object is effectively highlighted and the texture information inside the object is eliminated, which can alleviate the issue for texture over-transferred and benefit the GDSR operation.

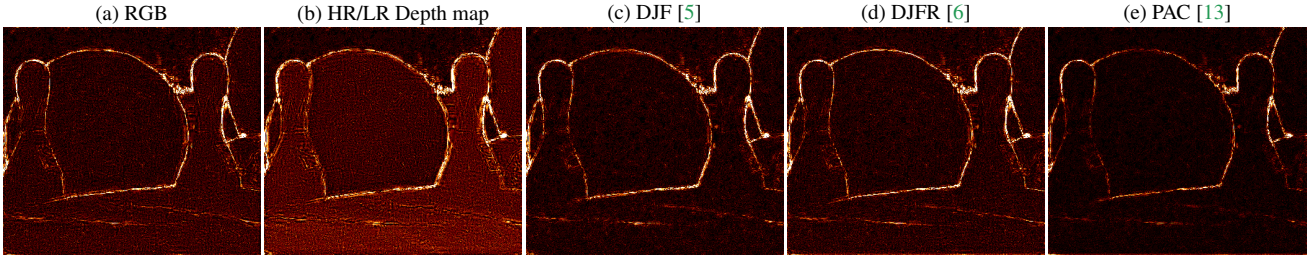
5. More qualitative comparison results

More qualitative comparison results are displayed in Fig. 3 - Fig. 5. Our method has excellent performance under multiple datasets and different downsampling scales, showing that our method is suitable for different objects and imaging conditions,

[5] https://thinklucid.com/helios-time-of-flight-tof-camera/

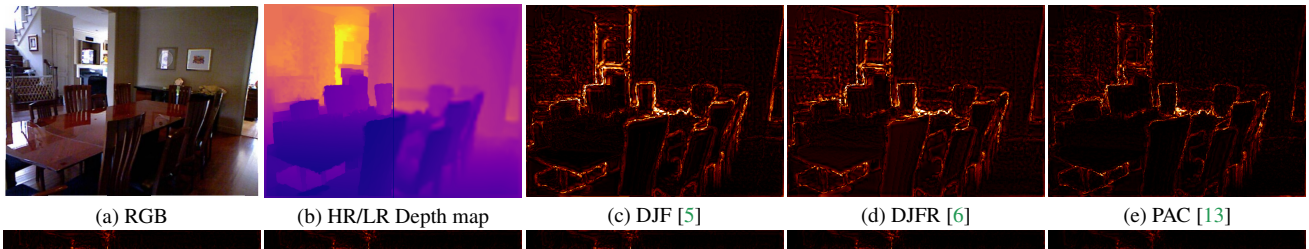


(a) RGB (b) HR/LR Depth map (c) DJF [5] (d) DJFR [6] (e) PAC [13]

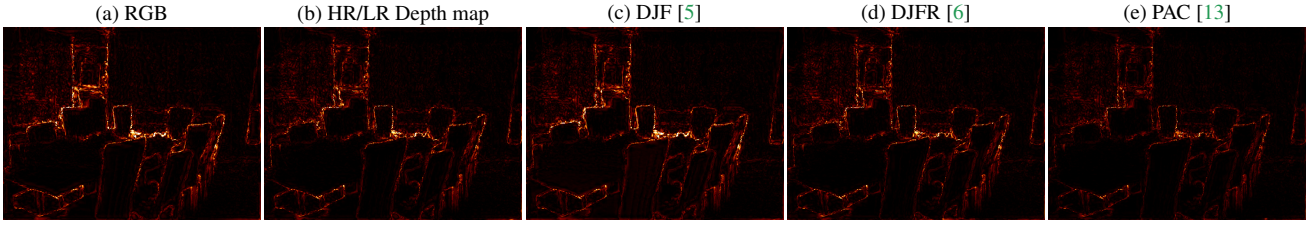


(f) CUNet [1] (g) DKN [4] (h) FDKN [4] (i) FDSR [2] (j) DCTNet (Ours)

Figure 3. Error maps for visual comparisons of "Image_07" of Middlebury dataset in 4× upscaling.



(a) RGB (b) HR/LR Depth map (c) DJF [5] (d) DJFR [6] (e) PAC [13]



(f) CUNet [1] (g) DKN [4] (h) FDKN [4] (i) FDSR [2] (j) DCTNet (Ours)

Figure 4. Error maps for visual comparisons of "Image_1432" of NYU v2 dataset in 8× upscaling.

and can surpass SOTA methods.

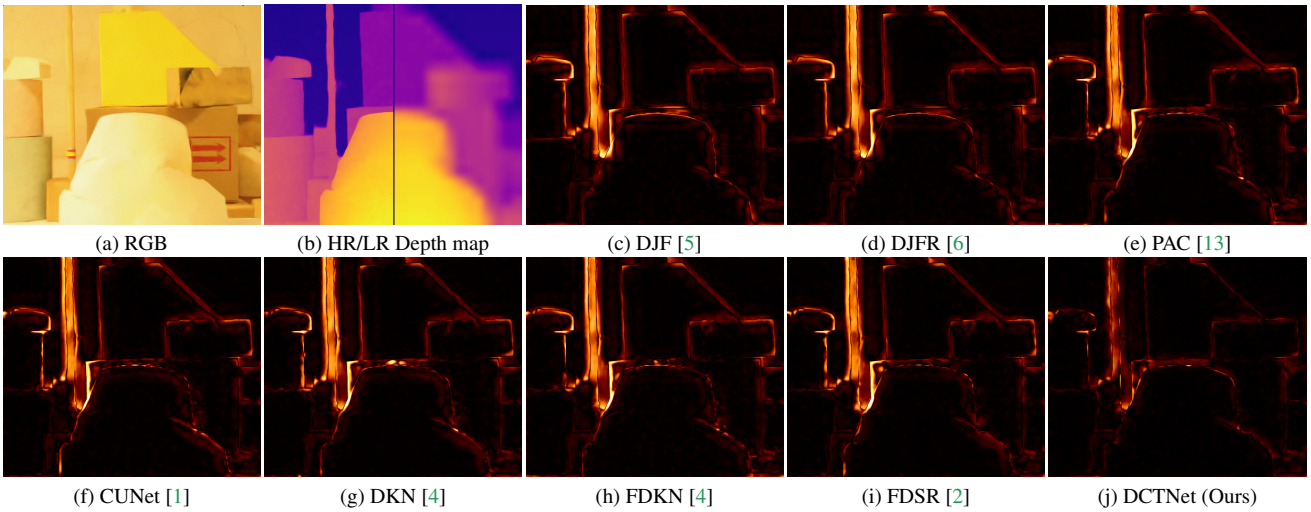
6. Architectures for ablation experiments

The network architectures for *semi-coupled residual blocks* (SCRBs) in ablation experiments Exp. I, II, DCT module in Exp. III and SCRB in Exp. V are shown in Fig. 6 (a)-(d), respectively. The experiments are referred in Sec. 4.4 of the original paper. Notably, for the setting of SCRB in Exp. I, II and V, we take the first SCRB in semi-coupled feature extraction module as an example.

7. Additional ablation experiments

Ablation study of $\tilde{\lambda}$ in Eq. (11). Besides $\tilde{\lambda} = e^{0.1}$ in the original ablation experiment IV (Tab. 4 in original paper), we conduct more experiments with $\tilde{\lambda} \in \{e^{-0.5}, e^{-0.1}, e^{0.5}\}$ to understand its behavior. Exp. VI-VIII in Tab. 2 show that different fixed $\tilde{\lambda}$ values lead to similar results, while a fixed $\tilde{\lambda}$ significantly reduces the flexibility compared to our final model with a learnable $\tilde{\lambda}$.

Ablation study on guided edge spatial attention. We first verify the necessity of the GESA module. Exp. IX in Tab. 2 shows that if we remove the GESA module and directly take the RGB features into the DCT module, the super-resolution performance is significantly degraded, demonstrating the necessity of the GESA module. We then verify the rationality of the



(a) RGB (b) HR/LR Depth map (c) DJF [5] (d) DJFR [6] (e) PAC [13]

(f) CUNet [1] (g) DKN [4] (h) FDKN [4] (i) FDSR [2] (j) DCTNet (Ours)

Figure 5. Error maps for visual comparisons of "Image_16" of Middlebury dataset in 16× upscaling.

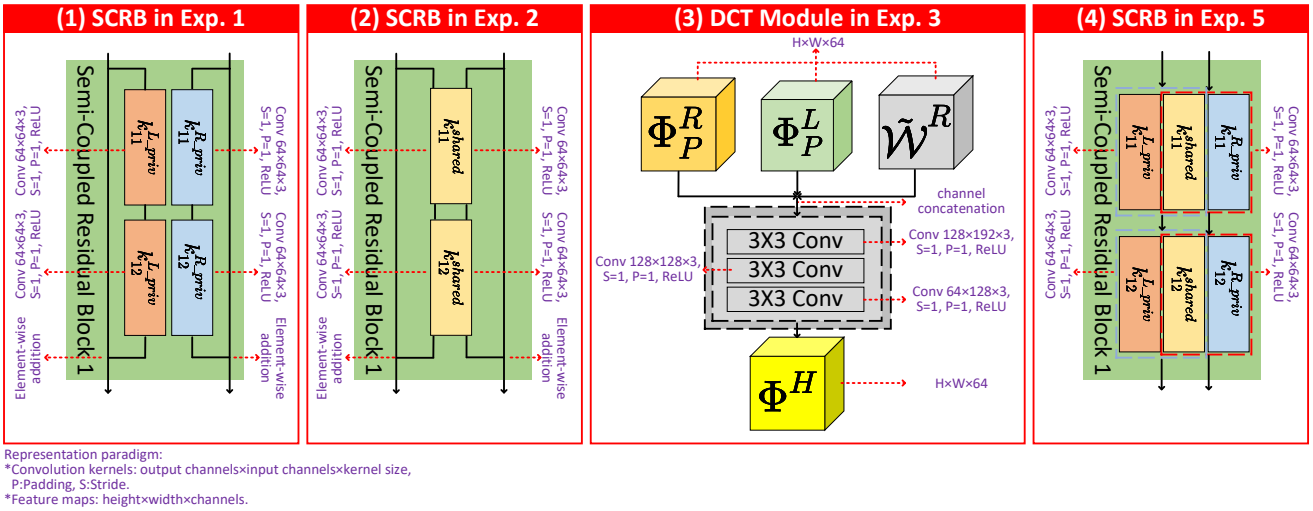


Figure 6. The detail architecture for semi-coupled residual blocks (SCRBs) in ablation experiments Exp. I, II, DCT module in Exp. III and SCRB in Exp. V, respectively. **Note that in the original paper, $\{\Phi_P^R$ and $\Phi_P^L\}$ can be abbreviated as $\{\Phi^R$ and $\Phi^L\}$.**

selected ESA [7] block in GESA by changing ESA to the CBAM [14] module, another commonly used attention mechanism. Exp. X in Tab. 2 shows that although using CBAM is better than w/o GESA, it is still worse than ours. The experiments validate that the ESA block cannot be simply eliminated or replaced by other modules.

8. Common/modality-specific features visualization

The shared/private features are visualized in Fig. 7. Obviously, the shared kernels extract the common properties of the depth/RGB image pair, mainly the edges (or contours). Modality-specific features, such as texture details and smooth depth regions, are extracted by the private kernels. The visualization is consistent with our motivation.

9. Limitation of DCTNet

We show representative failure cases in Fig. 8 and 9. If the illumination of the RGB image is insufficient, it may result in the ineffective guidance.

Intuitively, compared with the qualitative results shown in the Sec. 5, the prediction errors of failure cases are significantly larger. For the RMSE value, although DCTNet can surpass other SOTA methods, there is still a small gap between DCTNet

	Configurations	×4	×8	×16
VI	fixed $\tilde{\lambda} = e^{-0.5}$	1.80	3.60	6.73
VII	fixed $\tilde{\lambda} = e^{-0.1}$	1.81	3.63	6.65
VIII	fixed $\tilde{\lambda} = e^{0.5}$	1.85	3.64	6.71
IX	w/o GESA module	1.97	3.93	7.05
X	ESA block ⇒ CBAM	1.66	3.40	6.43
	Ours	**1.59**	**3.16**	**5.84**

Table 2. Results additional requested ablation experiments.

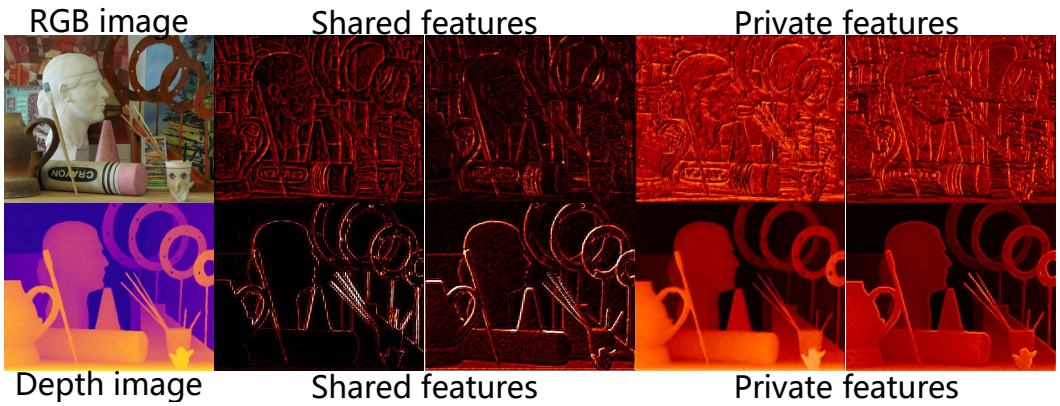
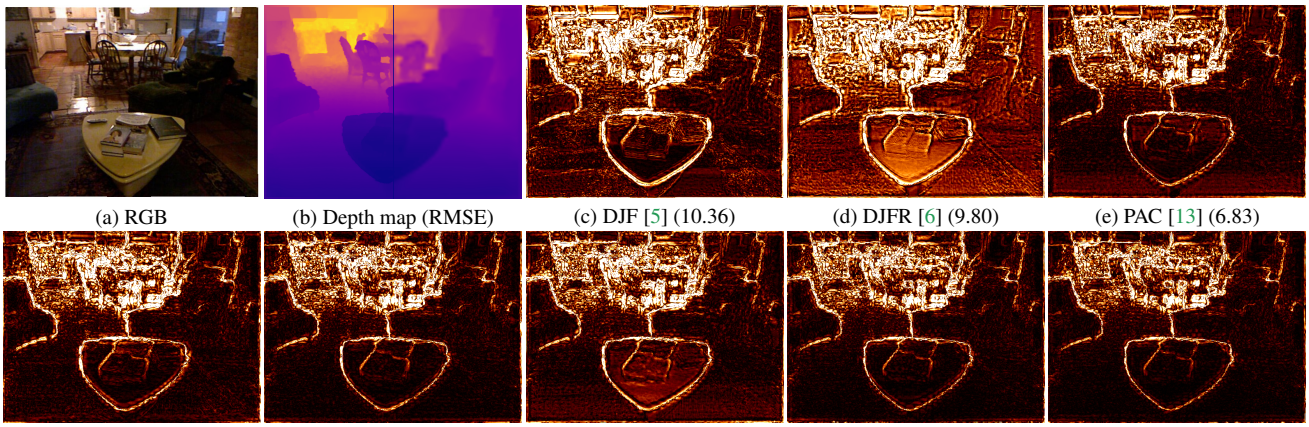


Figure 7. Visulization of the shared/private features.



(a) RGB (b) Depth map (RMSE) (c) DJF [5] (10.36) (d) DJFR [6] (9.80) (e) PAC [13] (6.83)

(f) CUNet [1] (7.72) (g) DKN [4] (6.40) (h) FDKN [4] (7.20) (i) FDSR [2] (6.06) (j) DCTNet (Ours) (6.02)

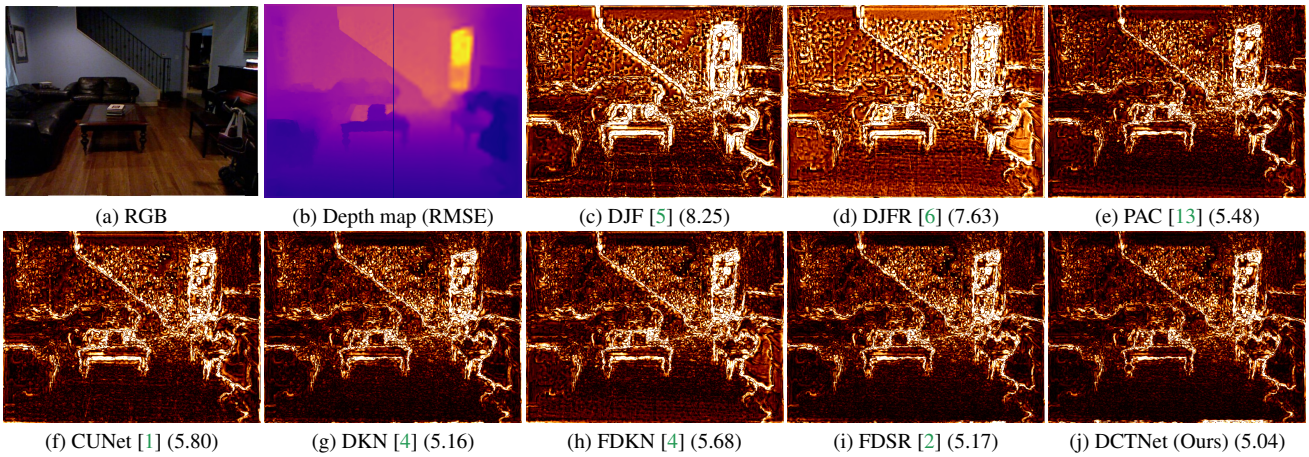
Figure 8. Error maps and the RMSE value of representative failure case from NYU v2 dataset in 8× upscaling.

and the ground-truth. The RMSE values of two failure cases (6.02 and 5.04) are also much larger than the average RMSE value (3.16) on the NYU v2 dataset in scaling factor 8×.

References

[1] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3333–3348, 2021. 5, 6, 7, 8

[2] Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In *CVPR*, pages 9229–9238. IEEE Computer Society, 2021. 2, 3, 5, 6, 7, 8



(a) RGB (b) Depth map (RMSE) (c) DJF [5] (8.25) (d) DJFR [6] (7.63) (e) PAC [13] (5.48)

(f) CUNet [1] (5.80) (g) DKN [4] (5.16) (h) FDKN [4] (5.68) (i) FDSR [2] (5.17) (j) DCTNet (Ours) (5.04)

Figure 9. Error maps and the RMSE value of representative failure case from NYU v2 dataset in 8× upscaling.

[3] Heiko Hirschmüller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *CVPR*. IEEE Computer Society, 2007. 2

[4] Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *Int. J. Comput. Vis.*, 129(2):579–600, 2021. 2, 5, 6, 7, 8

[5] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *ECCV*, pages 154–169. Springer, 2016. 2, 5, 6, 7, 8

[6] 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. 2, 5, 6, 7, 8

[7] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, pages 2356–2365. IEEE Computer Society, 2020. 6

[8] Si Lu, Xiaofeng Ren, and Feng Liu. Depth enhancement via low-rank matrix completion. In *CVPR*, pages 3390–3397. IEEE Computer Society, 2014. 2, 3

[9] William H Press, H William, Saul A Teukolsky, A Saul, William T Vetterling, and Brian P Flannery. *Numerical recipes in C, 2nd Edition: The art of scientific computing*. Cambridge university press, 1992. 2

[10] Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *CVPR*. IEEE Computer Society, 2007. 2

[11] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, pages 746–760. Springer, 2012. 2

[12] Gilbert Strang. The discrete cosine transform. *SIAM review*, 41(1):135–147, 1999. 2

[13] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik G. Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *CVPR*, pages 11166–11175. IEEE Computer Society, 2019. 2, 5, 6, 7, 8

[14] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In *ECCV*, pages 3–19. Springer, 2018. 6

[15] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE Multim.*, 19(2):4–10, 2012. 2