Supplementary Material: Wavelength- and Depth-Aware Deep Image Prior for Blind Hyperspectral Imagery Deblurring with Coarse Depth Guidance

Jiahuan Li[1,2], Xiaoyu Dong[1,2], Wei He[3,2], and Naoto Yokoya[1,2]
[1]The University of Tokyo, Japan
[2]RIKEN AIP, Japan
[3]Wuhan University, China
li@ms.k.u-tokyo.ac.jp, yokoya@k.u-tokyo.ac.jp

Section 1 details our network structures. Section 2 describes the generation process of the simulation dataset, including the kernel calculation for synthetic blur and coarse depth sampling. Section 3 provides camera settings and calibration for real data collection. Section 4 provides implementation details. Section 5 and Section 6 present additional experimental results for the simulation dataset and real data, respectively. Section 7 discusses the network structure for kernel prior, our considerations for mitigating the intrinsic latency and divergence problems for the DIP-based method, and data acquisition for paired hyperspectral and depth information.

1. Network Architecture

1.1. Depth refinement network

The depth refinement network $\mathcal{G}_D$ follows a U-Net architecture inspired by [1] with an encoder-decoder structure and skip connections, as illustrated in Fig. 1. The encoder consists of five basic blocks, connected by four downsampling steps using 2×2 max pooling. Four basic blocks are used in the decoder. Before each, the output of the previous block is upsampled with a 2×2 transposed convolution and concatenated with the corresponding feature map from the encoder to form a skip connection. Each block contains two 3×3 convolution layers, followed by batch normalization and PReLU. The final decoder layer is 3×3 convolution to generate a single-channel depth map. A sigmoid function is applied to the output, followed by post-processing to rescale the relative depth to an absolute range.

1.2. Multi-head kernel generator

For the multi-head kernel generator $\mathcal{G}_K$, we use a U-Net [12] with multiple output layers as shown in Fig. 2. This structure generates a set of kernels, each with dimensions $C \times 25 \times 25$ (C is the number of channels in the hyperspectral imagery), enabling spectral- and spatial-variant degradation for model-based deblurring.

2. Simulation Dataset

2.1. Wavelength- and depth-variant kernels

To generate blurred hyperspectral imagery (HSI) from the sharp ground truth in HyperSpectral-Depth (HSD) dataset [1], we calculate wavelength- and depth-variant kernels using the geometric optical model of the single-lens imaging system. Specifically, we model blur kernels as Gaussian functions with standard deviations that vary with wavelength and depth. Figure 3 illustrates the imaging process for the single lens. When the lens configuration is fixed, only light rays with a specific combination of wavelength and depth are focused on the sensor plane. Other light rays cause a circle of confusion (CoC) with wavelength- and depth-variant sizes related to the standard deviations of blur kernels. Although real hyperspectral imaging systems are much more complex with multiple optical components, this simplified illustration effectively demonstrates the property of defocus blur caused by lens refraction. We assume a point source of light at scene depth z. After being refracted by the lens, the light is focused at a distance s' that deviates from the sensor plane, resulting in the CoC. The radius of the CoC is calculated as follows:

$$C(\lambda, z) = \frac{1}{2} A (s(\frac{1}{F(\lambda)} - \frac{1}{z}) - 1), \tag{1}$$

where A is the aperture size, s is the sensor distance which depends on the focused distance of the camera, and $F(\lambda)$ is the wavelength-dependent focal length. The corresponding Gaussian blur kernel has a standard deviation $\sigma(\lambda, z)$ which is a constant multiple of $C(\lambda, z)$ [17]. For each sample, blur kernels from three optical configurations with different focal lengths and focused distances are used for data simulation. Figure 4 shows the blur kernels calculated with the focal length of 16mm (for 550nm) and a focused distance of 1.2m. The wavelength-dependent focal length is calculated

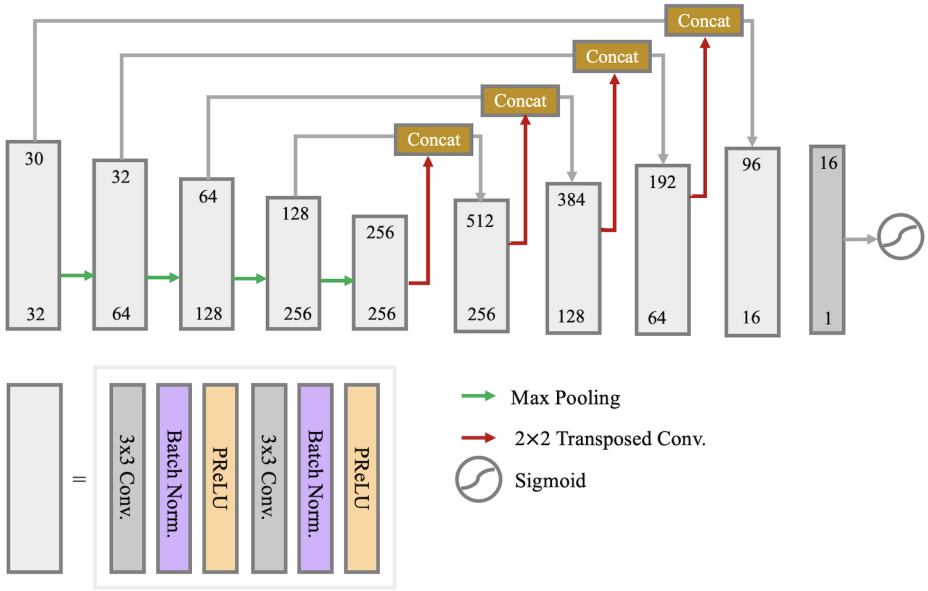

Figure 1. An illustration of the depth refinement network, including the details of the basic block. For each block, the number of input channels is shown at the top, and the number of output channels is shown at the bottom.

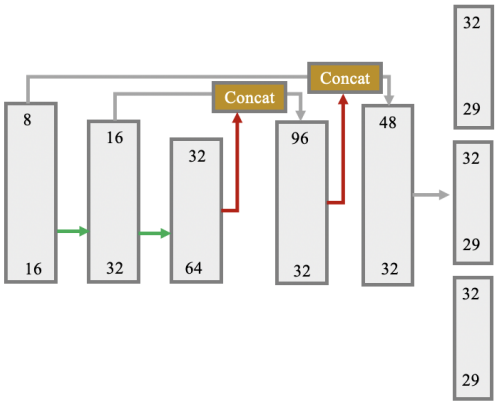

Figure 2. An illustration of the multi-head kernel generator with multiple output layers, showing input channels at the top and output channels at the bottom for each block.

using the refractive index of the lens material NOA61[1]. Corresponding blurring results are shown in Fig. 5. As the wavelength increases, the nearer object exhibits increasing blur, while the farther object becomes progressively sharper.

2.2. Depth processing

To simulate coarse depth in real-world scenarios, we randomly sample 4% of the pixels from the ground truth depth to create sparse depth and add Gaussian noise, as shown in Fig. 6. A simple nearest-neighbor interpolation is applied to the coarse depth to illustrate its degradation. Noise and errors are introduced, particularly at object boundaries, due to the degradation.

[1]https://refractiveindex.info/?shelf=other&book=Norland_NOA-61&page=Norland#google_vignette

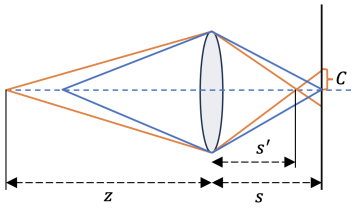

Figure 3. Single-lens imaging system. A light ray (blue line) with specific wavelength and depth is focused on the sensor plane located at a distance s behind the lens. Meanwhile, the light ray (orange line) from a point source at depth z is focused at s', in front of the sensor plane, resulting in CoC with a radius of C.

3. Real Data

3.1. Camera settings

For the hyperspectral camera (EBA JAPAN NH-9), we use a 16mm focal length lens and set the focused distance to 1m for scenes within a depth range of 0.4–1.6m. The aperture is fully opened, with an f-number of 1.4 to ensure sufficient incident light and maintain the spectral accuracy. Additionally, illumination intensity and sensor sensitivity vary with wavelength. To achieve low-noise capturing in general while avoiding over-exposed bands, we capture each scene twice with different exposure times and select the optimal one for each band.

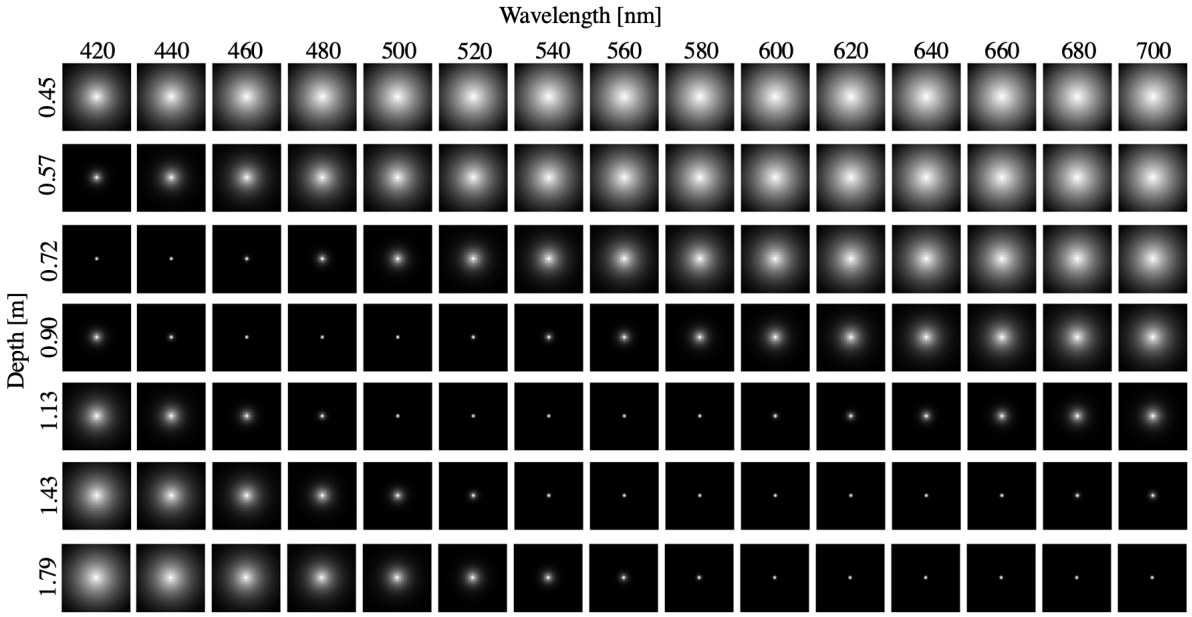

Figure 4. Wavelength- and depth-variant Gaussian blur kernels.

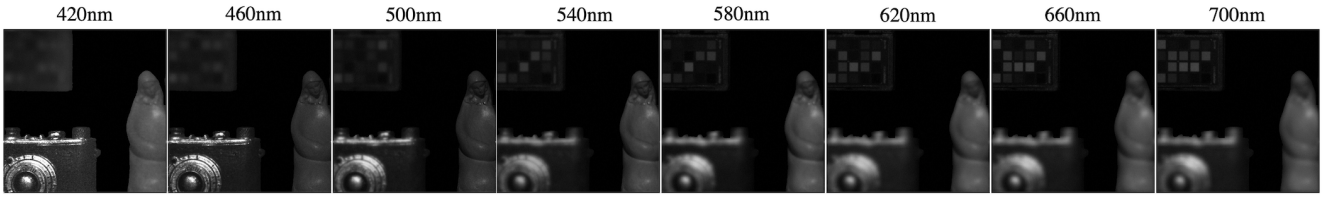

Figure 5. Degraded spectral images with the blur kernels illustrated in Fig. 4.

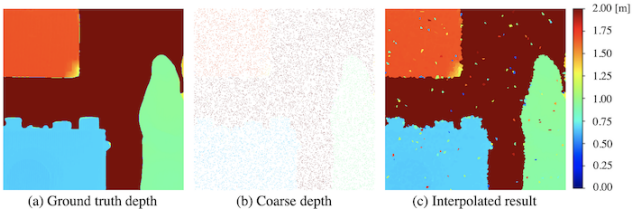

Figure 6. Coarse depth (b) is generated by sampling 4% pixels in ground truth (a) and adding Gaussian noise. The interpolation result (c) of the coarse depth illustrates the degree of degradation.

3.2. Calibration

The depth maps are captured using a separate RGB-depth camera (Azure Kinect DK), which has a different field of view and resolution compared to the hyperspectral camera, resulting in misalignment with the captured HSIs, as shown in Fig. 7(a-c). To address this issue, we perform careful calibration to register the depth maps to the coordinate system of HSIs. Specifically, we first extract the intrinsic matrix and distortion coefficients for both cameras. Then, we compute the rotation matrix and the translation vector that project 3D points from the RGB-depth camera to the hyperspectral camera. All calibrations are performed using the OpenCV API. For each pixel in the depth map, we project it onto the 2D coordinate system of the HSIs using the aforementioned parameters and the depth value. Occlusion filtering is applied to the projected depth to handle the pixels that are visible in the RBG-depth camera but occluded in the hyperspectral camera[2]. Figure 7(d) shows the registered depth, which is well-aligned with the spectral imagery as shown in (e).

4. Implementation Details

4.1. Proposed method

For the simulation experiments, we apply nearest-neighbor interpolation to the coarse depth before concatenating it with the blurred HSI to form the input for the refinement network. For real data, we use the coarse depth

[2]https://github.com/eugeniu1994/Stereo-Camera-LiDAR-calibration.git

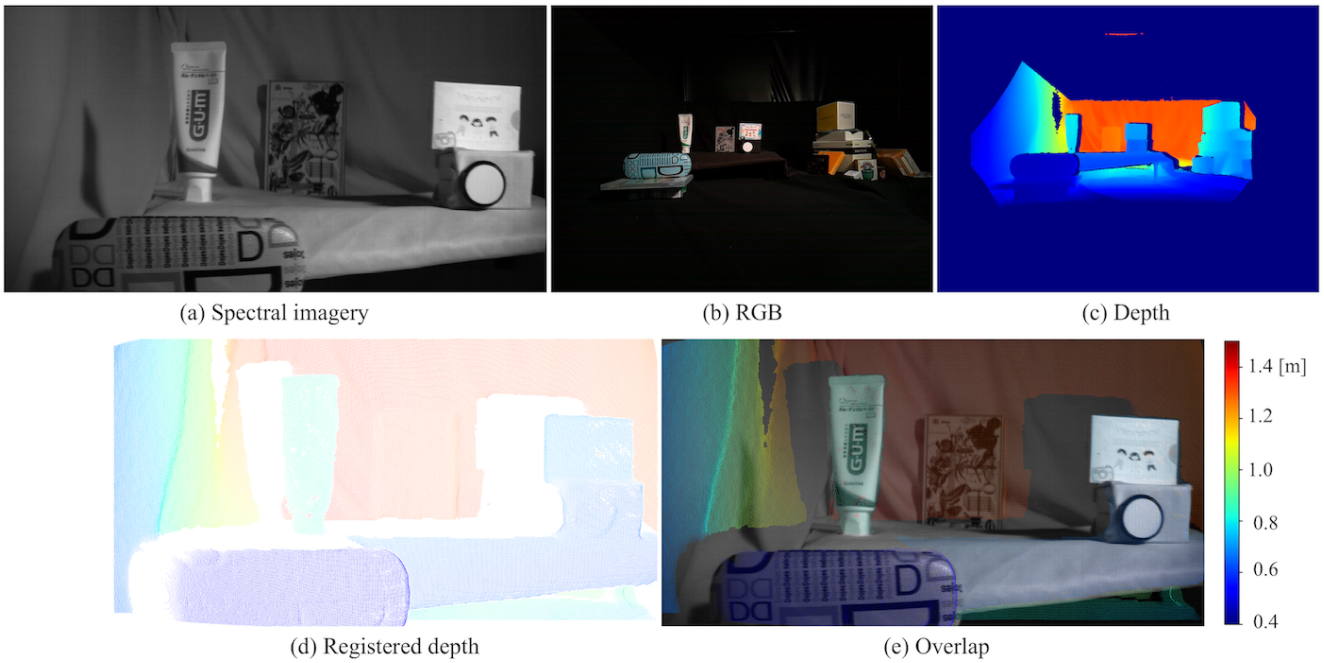

Figure 7. The RGB image (b) and depth map (c) captured by the RGB-depth camera are misaligned with the spectral imagery (a) captured by the hyperspectral camera. We perform calibration to register the depth map to the HSI coordinate system, resulting in (d). The overlap (e) between the registered depth and spectral imagery demonstrates the accuracy of registration.

directly without interpolation. The center depth for generating the soft-weight map is adaptively determined for each sample by dividing the input depth into m levels using a spacing-increasing discretization strategy [6]. We set the standard deviation of the Gaussian function to 0.1 in the soft-weight map and 6 in the kernel regularization term. The experiments are conducted on a PC with an NVIDIA RTX A5000 GPU. The total model size is approximately 31MB for the synthetic experiments and 33MB for the real experiments. For the simulated data with dimensions $29 \times 512 \times 512$, the processing time is 806 seconds for 2,000 iterations. For the real data with 121 channels, the processing time is 912 seconds for 800 iterations.

4.2. Comparison methods

All comparison methods, except for PnP [14], take RGB or grayscale images as the input. These methods are applied to HSI deblurring in a channel-wise manner, with deblurring conducted separately for each wavelength. For the referenced-based ABGI [9], we use the channel with the highest PSNR as the reference for the simulation dataset. For real data, where ground truth is unavailable for calculating PSNR, we use the $60{-}th$ channel (700nm) as the reference since our captured HSIs are typically less blurry and noisy around the center of the wavelength range and more degraded at the edges. For the non-blind method PnP [14] with a deep denoiser as the prior, we use its pre-trained prior along with blur kernels estimated by DCP [10] for the simulated data and PMP [16] for the real data.

5. Additional Results on Simulation Dataset

Figure 8 provides qualitative comparisons between the proposed method and all comparison methods. Model-based methods such as (c) DCP [10] and (e) SelfDeblur [11] exhibit significant distortion and artifacts because they estimate spatially uniform kernels without considering depth-variant blurriness. Although other comparison methods have fewer artifacts, they are less effective when encountering large-degree blurriness, producing results that are still over-smoothed. This is particularly evident in the learning-based methods (f) IFAN [8] and (g) NAFNet [4]. Our method achieves superior performance with better texture restoration and fewer artifacts. The error maps for three HSIs, calculated between the ground truth and deblurred results, further demonstrate the effectiveness of our method, as shown in Fig. 9. We visualize the results of the ablation study in Fig. 10. Without considering depth variation, *i.e.*, estimating spatial-invariant blur kernels, the texture of the closer object is successfully restored while the farther object exhibits severe artifacts (c). Integrating depth guidance through either binary masks or soft-weight maps reduces these artifacts (d-e). The final model, which uses soft-weight maps, outperforms the mask-based approach, delivering sharper textures at 700nm and improved color fidelity.

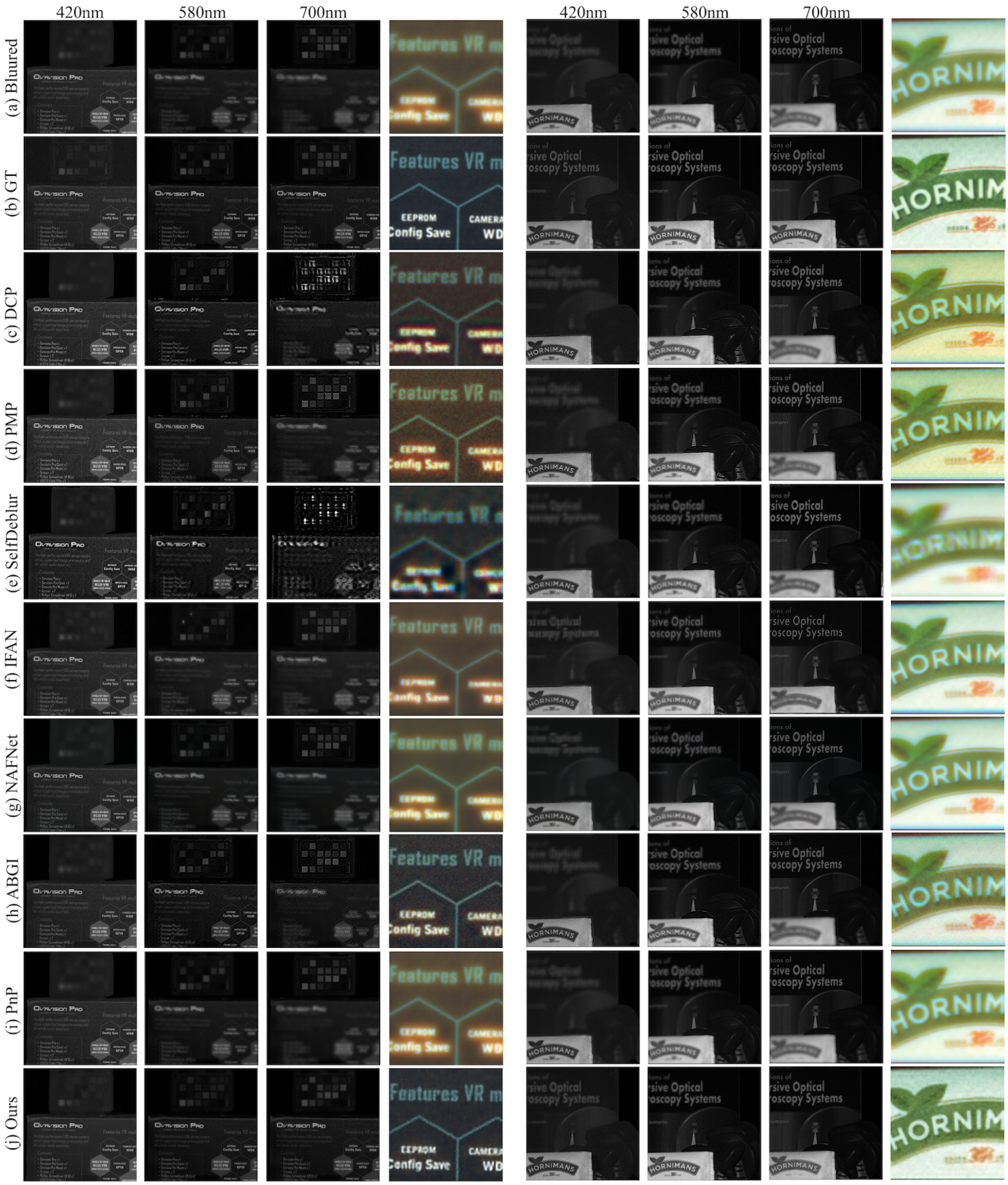

Figure 8. Spectral images at 420nm, 580nm, and 700nm from two HSIs. The fourth and eighth columns display synthetic RGB images created by uniformly combining wavelengths of 420-500nm for the blue channel, 510-610nm for the green channel, and 620-700nm for the red channel.

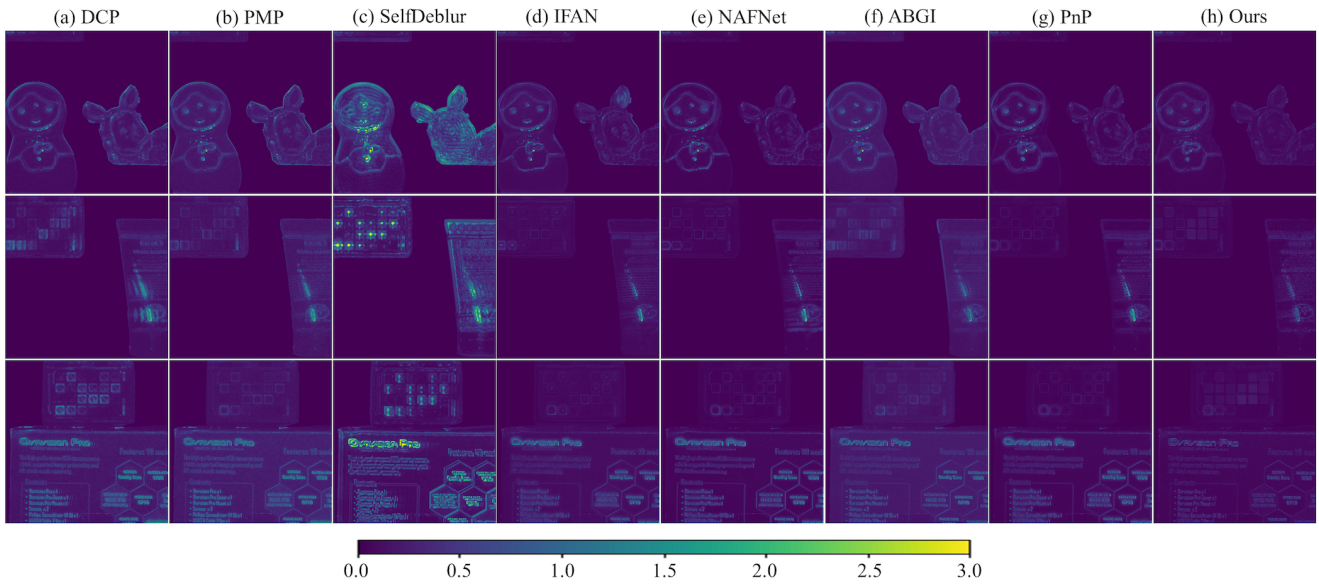

Figure 9. Error maps between ground truth HSIs and deblurred results. For channel-wise methods such as DCP [10], PMP [16], SelfDeblur [11], IFAN [8], NAFNet [4], and ABGI [9], we concatenate the single channel results to form the deblurred HSIs used for error calculation.

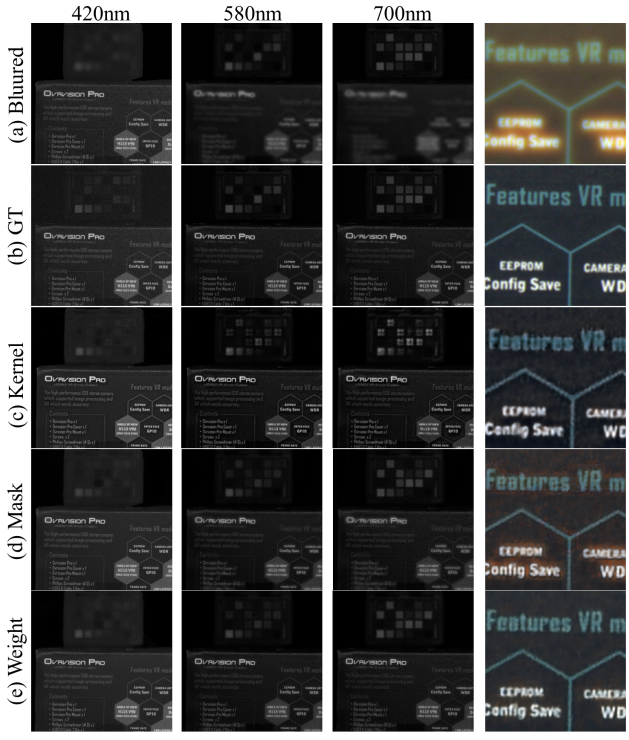

Figure 10. Qualitative results of ablation study. (c) Wavelength-aware deblurring without considering depth. (d) Wavelength- and Depth-aware deblurring with binary masks. (e) Wavelength- and Depth-aware deblurring with soft-weight maps (our final model).

6. Additional Results on Real Data

Figure 11 present additional deblurred results on real blurred HSIs captured by our set-up, corresponding to samples (4-6) in the Tab. 3 of main paper. We also show detailed deblurred results of an extremely degraded channel (400nm) in Fig. 12. Our method produces sharper textures with fewer color fringes and artifacts compared to others.

7. Discussions

7.1. Network structures for kernel prior

For the kernel prior in DIP-based deblurring methods, Wang *et al*. [15] represent the kernel with a convolutional neural network (CNN) to leverage DIP for blur kernels. While, Ren *et al*. [11] argue that DIP is designed for capturing prior information in natural imagery and may not be favorable for kernel prior. Given the small parameter amount of the 2D blur kernel in their problem setting, they utilize a fully-connected network (FCN) to model the kernel prior, which demonstrates superior performance compared to Double-DIP [7] using CNNs for both image and kernel prior. In contrast to the use of neural networks, whether CNNs or FCNs, a recent study uses a normalized array to directly represent the blur kernel, achieving comparable deblurring performance with that of FCN [2]. Consequently, the optimal network structure for capturing the prior information of the degradation kernel in deblurring and its relationship with kernel features remain open questions.

We adopt a CNN with multiple branches at the out-

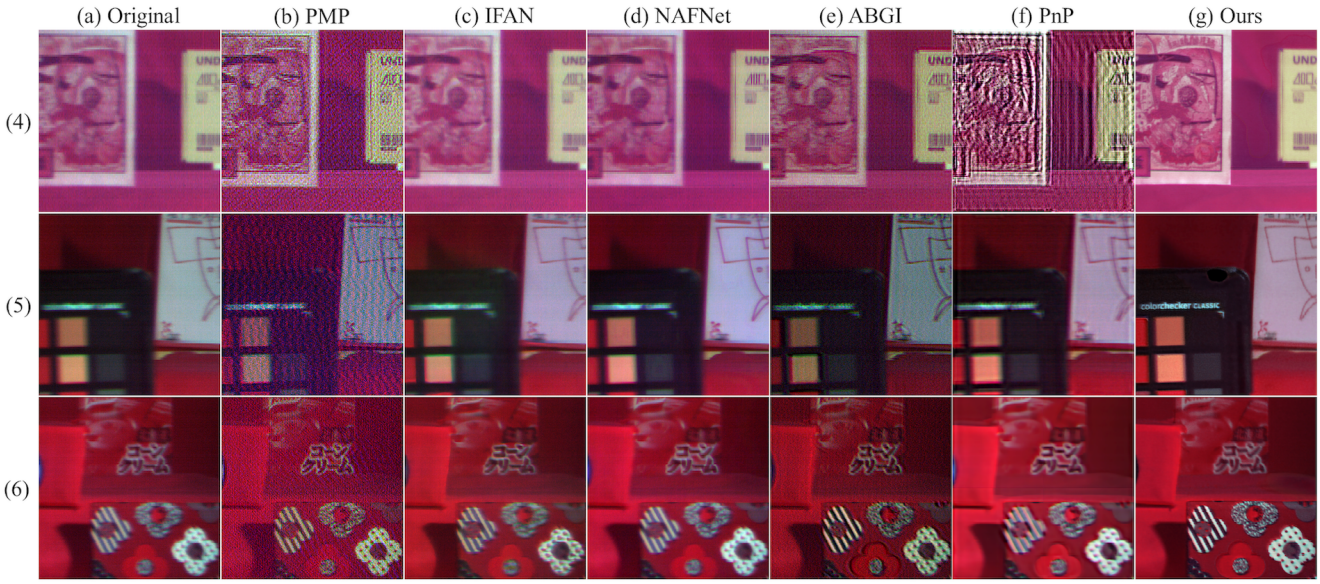

Figure 11. Original captures and deblurred results for real HSIs. The false color images are generated by assigning 400nm, 420nm, and 1000nm to blue, green, and red channels, respectively.

put layer to model the depth-dependent multi-channel blur kernel from the considerations of the kernel structure and model size. The CNN is more suitable for outputting multiple 3D kernels than the FCN which needs reshaping. Also, the kernels for different wavelengths and depths usually share similar spatial shape determined by the optics while having different spatial sizes, which makes the CNN more favorable with its ability to capture spatial features. On the other hand, the CNN is more efficient for 3D kernels. For example, to generate a 3D kernel with the size of $29 \times 25 \times 25$, the CNN (three-stage U-Net [12]) has about 0.15 million parameters, significantly less than the FCN with a single hidden layer (479 million).

7.2. Considerations for latency and divergence problems

Although DIP-based methods offer the advantage of requiring no training process or dataset, they suffer from high latency due to iterative forward and backward network processes. In our model, we estimate multiple multi-channel kernels, which involves more parameters and thus increases processing time. To reduce the latency, we employ a CNN with shared parameters for depth-variant kernels. Additionally, we adaptively divide the depth range into a few levels for each sample instead of estimating kernels for the entire depth range, thereby alleviating the computational burden. Another concern is that DIP-based methods are sensitive to hyperparameter tuning, and inappropriate parameters even lead to divergence. To address this, we apply an additional regularization term to encourage the kernel to have relatively large values in the central region, reflecting the property of defocus blur. There is also a smoothness regularization term to prevent the network from converging to delta results. Further improvements can be achieved by incorporating additional image priors, such as total variation [3]. A diffusion prior on the kernel can also be considered, as it can be trained on synthetic kernels without the need for HSI data [5].

7.3. HSD data acquisition

Our work focuses on the defocus blur problem in wide-range HSI imaging, and first integrates coarse depth guidance into HSI deblurring. We also simultaneously refine depth, providing a practical approach for high-quality paired HSD data acquisition. This multi-modality with both semantic and geometric information is essential for scene understanding. Existing acquisition strategies can be divided into two types: combination-based and reconstruction-based. The former combines two different imaging mechanisms for two modalities, such as coded aperture-based HSI imaging and time-of-flight system for depth imaging [13]. This strategy usually suffers from a large form factor, which limits its application in laboratory environments. In contrast, the reconstruction-based method uses compact devices with learnable optics to capture RGB images and reconstruct HSI and depth information afterwards [1]. This strategy relies heavily on the dataset to optimize the reconstruction algorithm and optics. Currently, the HSD dataset is limited in both data amount and wavelength range (visible range only). Our approach captures

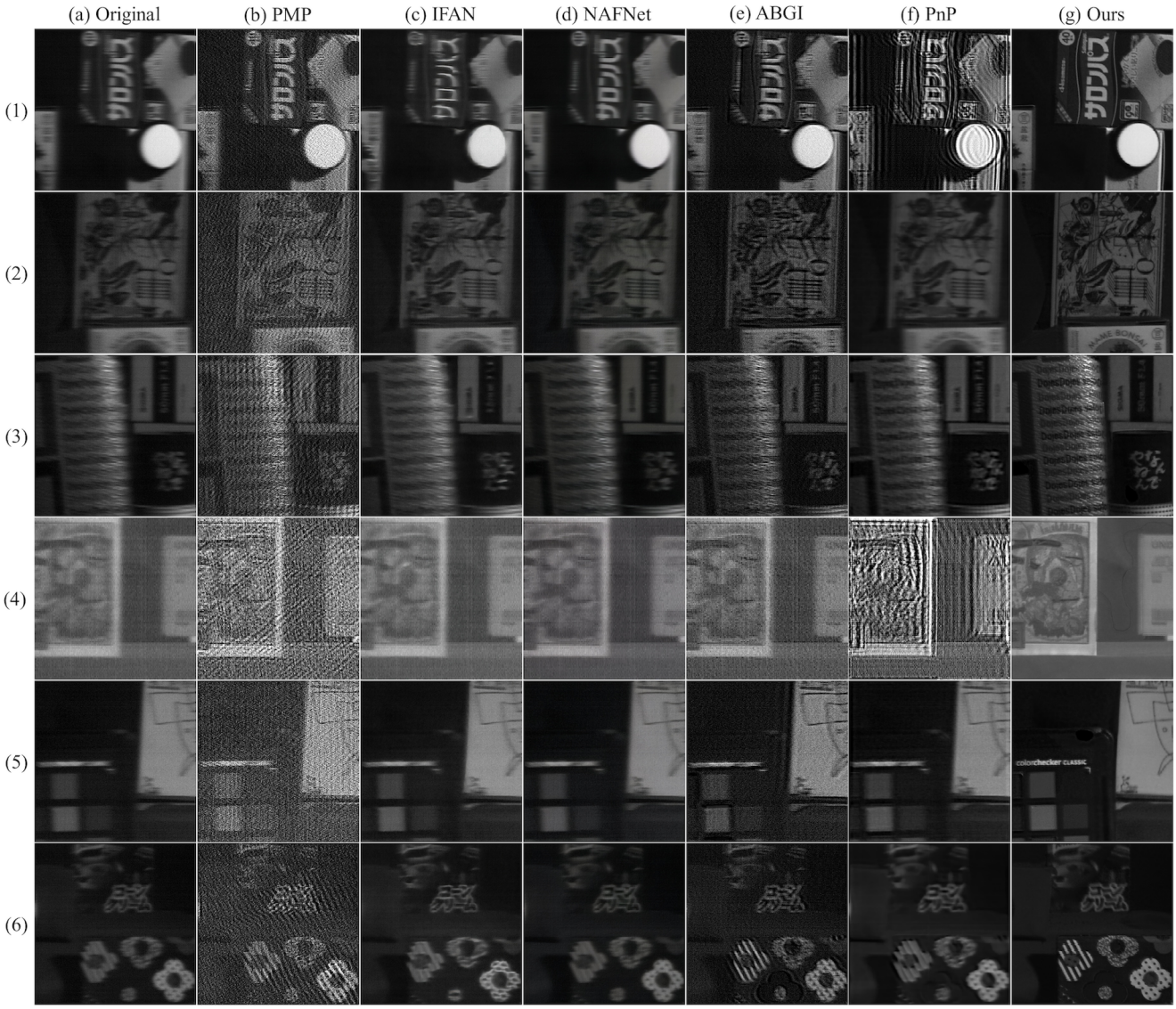

Figure 12. Original captures and deblurred results of an extremely degraded channel (400nm) in HSIs.

wide-range HSI and depth with separate cameras followed by registration and restoration, which is more practical for HSD data acquisition in various environments, promoting future developments for HSD imaging devices and multi-modal scene understanding.

References

[1] Seung-Hwan Baek, Hayato Ikoma, Daniel S Jeon, Yuqi Li, Wolfgang Heidrich, Gordon Wetzstein, and Min H Kim. Single-shot hyperspectral-depth imaging with learned diffractive optics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2660, 2021. 1, 7

[2] Antonie Brozová and Václav Šmıdl. Avoiding undesirable solutions of deep blind image deconvolution. *Proceedings Copyright*, 559:566, 2024. 6

[3] Tony F Chan and Chiu-Kwong Wong. Total variation blind deconvolution. *IEEE transactions on Image Processing*, 7(3):370–375, 1998. 7

[4] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 4, 6

[5] Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. Parallel diffusion models of operator and image for blind inverse problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6059–6069, 2023. 7

[6] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 2002–2011, 2018. 4

[7] Yosef Gandelsman, Assaf Shocher, and Michal Irani. " double-dip": unsupervised image decomposition via coupled deep-image-priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11026–11035, 2019. 6

[8] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2034–2042, 2021. 4, 6

[9] Zhuangtianyu Liao, Wenyi Zhang, Qingwei Chu, Hao Ding, and Yuxin Hu. Multispectral remote sensing image deblurring using auxiliary band gradient information. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023. 4, 6

[10] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1628–1636, 2016. 4, 6

[11] Dongwei Ren, Kai Zhang, Qilong Wang, Qinghua Hu, and Wangmeng Zuo. Neural blind deconvolution using deep priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3341–3350, 2020. 4, 6

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1, 7

[13] Hoover Rueda-Chacon, Juan F Florez-Ospina, Daniel L Lau, and Gonzalo R Arce. Snapshot compressive tof+ spectral imaging via optimized color-coded apertures. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2346–2360, 2019. 7

[14] Xiuheng Wang, Jie Chen, and Cédric Richard. Tuning-free plug-and-play hyperspectral image deconvolution with deep priors. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 4

[15] Zhunxuan Wang, Zipei Wang, Qiqi Li, and Hakan Bilen. Image deconvolution with deep image and kernel priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 6

[16] Fei Wen, Rendong Ying, Yipeng Liu, Peilin Liu, and Trieu-Kien Truong. A simple local minimal intensity prior and an improved algorithm for blind image deblurring. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):2923–2937, 2020. 4, 6

[17] Guodong Xu, Yuhui Quan, and Hui Ji. Estimating defocus blur via rank of local patches. In *Proceedings of the IEEE international conference on computer vision*, pages 5371–5379, 2017. 1