Computational Spectral Imaging with Unified Encoding Model and Beyond
Supplementary Material

Xinyuan Liu   Lingen Li   Lin Zhu   Lizhi Wang*
Beijing Institute of Technology
{xinyuanliu, lingenli, linzhu, wanglizhi}@bit.edu.cn

Due to the space limit, some formula derivation and additional experiments could not be included in the main paper. This supplementary material provides a comprehensive demonstration of the experimental setup and comparative analysis, structured as follows:

1. PSF Derivation

Here, we describe the details of the PSF derivation for PEM-P by following the general formulation in [1, 3, 6–9].

Scene light generates phase delays $\phi(x, y, \lambda)$ through the DOE of the height map $H(x, y)$:

$$\phi(x, y, \lambda) = \frac{2\pi\Delta n}{\lambda} H(x, y) \tag{1}$$

where Δn is the refractive index difference between the air and the material of the optical element.

The incident point source with coordinates (x, y) at a distance d from the DOE can be expressed as:

$$U_0(x, y, \lambda) = e^{i\frac{2\pi}{\lambda}\frac{x^2+y^2}{d}} \tag{2}$$

The incident light gets phase modulated through the DOE:

$$U_{doe}(x, y, \lambda) = A(x, y) U_0(x, y, \lambda) e^{i\frac{2\pi}{\lambda}\phi(x,y,\lambda)} \tag{3}$$

where $A(x, y)$ is the optical aperture.

Algorithm 1 Deep Unfolding Algorithm

```
Input: Sensor image $I_{rgb}$, optical encoding operation $\mathcal{Q}$,
    trade-off parameter $\eta$
Output: Reconstructed spectral image $Z_K$
Initialize $Z_0$ from $I_{rgb}$
for $k = 1, 2, ..., K$ do
    $I_k = \arg\min_I \|I_{rgb} - \mathcal{Q}(I)\|^2 + \eta \|I - Z_{k-1}\|^2$ ;
    //Solving the encoding Model-drived sub-problem
    $Z_k = f(I_k)$ ;
    //Reconstructing with a Res-U-Net(depth=4)
end for
```

The modulated wavefield passes through the Fresnel diffraction law to reach the sensor plane at a distance z from the DOE:

$$U_{\text{sensor}}(x, y, \lambda) = \mathcal{F}^{-1}\left\{\mathcal{F}\{U_{\text{doe}}(x, y, \lambda)\} e^{i\frac{2\pi}{\lambda}z} e^{-i\pi\lambda z\left(f_x^2+f_y^2\right)}\right\} \tag{4}$$

where f_x and f_y are the frequency variables of x and y, respectively, and $\mathcal{F}$ denotes the Fourier transform.

The PSF $P(x, y, \lambda)$ is the squared intensity of the wavefield:

$$P(x, y, \lambda) \propto |U_{sensor}(x, y, \lambda)|^2 \tag{5}$$

2. Decoding Models

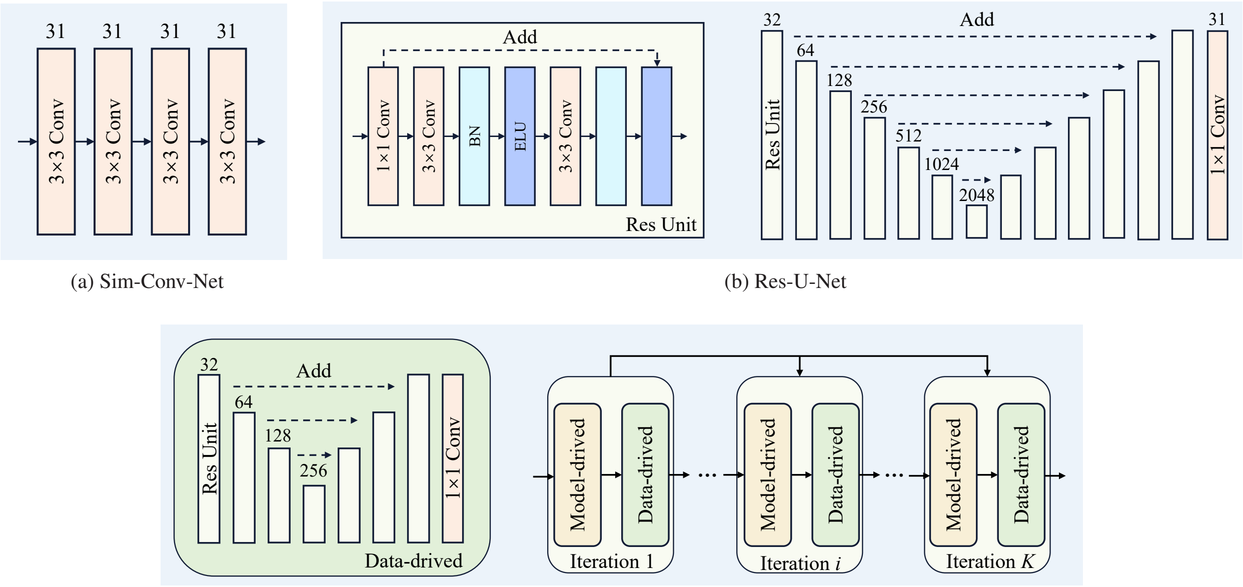
2.1. Structure of Decoding Models

Sim-Conv-Net. The Sim-Conv-Net, shown in Figure 1a, comprises four convolutional layers, and the filter size of each convolutional kernel is 31.

Res-U-Net. The Res-U-Net, shown in Figure 1b, consists of Res Units, the filter size indicated by the top number. Res-U-Net integrates these concepts by incorporating residual blocks into the U-Net architecture. This combination aims to leverage the benefits of both residual connections for gradient flow and U-Net for precise segmentation.

Unfolding-Net. The Unfolding-Net, shown in Figure 1c, is based on the deep unfolding algorithm [4, 10–13] and consists of a model-driven module and a data-driven module.

*Corresponding author: Lizhi Wang.



(a) Sim-Conv-Net

(b) Res-U-Net

(c) Unfolding-Net

Figure 1. The structure of decoding models.

The algorithmic process is shown in Algorithm 1, and the final reconstruction result is obtained by iterating these two modules several times. In the experiment, the number of iterations K is 4.

2.2. Deep Unfolding Algorithm

In general, spectral image reconstruction aims at recovering the potential spectral image I from the RGB sensor observation $I_{rgb} = \mathcal{Q}(I) + n$, where $\mathcal{Q}$ is a noise-independent optical encoding operation and n is assumed to be additive white Gaussian noise with a standard deviation of σ.

From a Bayesian perspective [2], the solution $\hat{I}$ can be obtained by solving a Maximum A Posteriori (MAP) estimation problem.

$$\hat{I} = \arg\max_{I} \log p(I_{rgb} \mid I) + \log p(I) \tag{6}$$

where $\log p(I_{rgb} \mid I)$ represents the log-likelihood of observation I_{rgb}, $\log p(I)$ delivers the prior of clean image I and is independent of degraded image I_{rgb}. Eq. 6 can be reformulated as

$$\hat{I} = argmin_I ||I_{rgb} - \mathcal{Q}(I)||^2 + \lambda R(I) \tag{7}$$

where λ is a balance parameter. The data term enforces alignment with the observation model, while the regularization term enforces the desired spectral image prior $R(\cdot)$.

To separate non-differentiable regularization terms from the data term in Eq. 7, the variable splitting technique is often used, introducing an auxiliary variable Z and reformulating the equation as a constrained optimization problem.

$$\hat{I} = argmin_I ||I_{rgb} - \mathcal{Q}(I)||^2 + \lambda R(Z), s.t. Z = I \tag{8}$$

Afterward, the constrained optimization problem can be transformed into a nonconstrained optimization problem using the half quadratic splitting (HQS) method.

$$(\hat{I}, \hat{Z}) = \arg\min_{f,h} \|y - \mathcal{Q}(I)\|^2 + \eta \|Z - I\|^2 + \lambda R(Z) \tag{9}$$

where η is a penalty parameter. Eq. 9 can be split into two subproblems as

$$I_k = argmin_I ||I_{rgb} - \mathcal{Q}(I)||^2 + \eta ||I - Z_{k-1}||^2 \tag{10}$$

$$Z_k = argmin_Z \mu ||Z - I_k||^2 + R(Z) \tag{11}$$

The I_k-problem in Eq. 10 is a quadratic regularized least-squares problem that ensures the data fidelity. The direct solution can be given according to the specific encoding model.

$$I_k = \arg\min_{I} \|I_{rgb} - \mathcal{Q}(I)\|^2 + \eta \, \|I - Z_{k-1}\|^2 \tag{12}$$

The regularization Z_k-problem in Eq. 11 is generally solved using a data-driven module with a deep neural network. This process can be expressed as:

$$Z_k = f\,(I_k) \tag{13}$$

Concretely, we adopt the Res-U-Net(depth=4) as the data-driven module.

AEM. Rewrite Eq.12 for AEM:

$$I_k = \arg\min_{I} \left\|I_{rgb} - AID\right\|^2 + \eta \left\|I - Z_{k-1}\right\|^2 \qquad (14)$$

where A is the Mask of AEM and D represents the integration operation of the sensor. The solution is obtained using the gradient descent method:

$$I_k = I_{k-1} - \alpha(-2AD^T(I_{rgb} - AI_{k-1}) + 2\eta(I_{k-1} - Z_{k-1})) \qquad (15)$$

where α is the step length of the gradient descent.

PEM. Rewrite Eq.12 for PEM:

$$I_k = \arg\min_{I} \left\|I_{rgb} - I \otimes PD\right\|^2 + \eta \left\|I - Z_{k-1}\right\|^2 \qquad (16)$$

where P is the PSF of PEM and D represents the integration operation of the sensor.

The solution is obtained using the gradient descent method:

$$I_k = I_{k-1} - \alpha(-2PD \otimes (I_{rgb} - I_{k-1} \otimes PD) + 2\eta(I_{k-1} - Z_{k-1})) \qquad (17)$$

where α is the step length of the gradient descent.

WEM. Rewrite Eq.12 for WEM:

$$I_k = \arg\min_{I} \left\|I_{rgb} - ID\right\|^2 + \eta \left\|I - Z_{k-1}\right\|^2 \qquad (18)$$

where D represents the integration operation of the sensor. The solution is obtained using the gradient descent method:

$$I_k = I_{k-1} - \alpha(-2D^T(I_{rgb} - I_{k-1}D) + 2\eta(I_{k-1} - Z_{k-1})) \qquad (19)$$

where α is the step length of the gradient descent.

3. Additional Experiments

3.1. Response Selection of WEM-P

We compare the response dataset [5], which includes 28 camera responses, to the response of the FLIR BFS_U3_04S2C_C camera in the WEM-P. According to Table 1, the latter experiment performs the best. Therefore, we select the response curve of the FLIR BFS_U3_04S2C_C camera as the response curve for the WEM-P and other experiments with fixed response curves.

3.2. Initialization of Ideal Encoding Models

The ideal model encoding possesses significant flexibility, and the initialization considerably influences the experimental outcomes. Therefore, we conduct experiments on initializing AEM-I, PEM-I, and WEM-I encodings to determine the optimal initialization for the experimental setup. To this end, we employ two types of initialization: constant initialization and random initialization.

Initialization of AEM-I. Two constant initialization types, all-0.5 and all-1, are used for AEM-I encoding. Two random initialization types are applied by setting the mask to a number between 0 and 1: uniform and Gaussian.

Table 1. Comparison of different response curves of WEM-P, with the best results in bold.

Response curves	PSNR↑	PSNR-SI↑	SAM↓	ERGAS↓
FLIR_BFS-U3-04S2C-C	**44.96**	**38.81**	**0.03988**	**6.23**
Canon_1DMarkIII	43.54	38.47	0.0461	7.88
Canon_20D	43.26	38.73	0.0472	8.38
Canon_300D	43.93	38.58	0.0438	7.23
Canon_40D	43.43	38.45	0.0481	8.09
Canon_500D	44.01	38.41	0.0443	7.25
Canon_50D	43.88	38.27	0.0445	7.34
Canon_5DMarkII	44.04	38.57	0.0435	7.37
Canon_600D	43.81	38.40	0.0452	7.66
Canon_60D	43.98	38.41	0.0437	7.35
Hasselblad_H2	44.18	39.00	0.0416	7.35
Nikon_D3X	43.85	38.75	0.0439	7.50
Nikon_D200	43.74	38.72	0.0444	7.64
Nikon_D3	43.47	38.55	0.0462	7.97
Nikon_D300s	43.65	38.68	0.0452	7.73
Nikon_D40	44.29	38.63	0.0426	6.69
Nikon_D50	44.31	38.76	0.0431	6.93
Nikon_D5100	43.33	38.57	0.0459	8.21
Nikon_D700	43.70	38.66	0.0444	7.90
Nikon_D80	43.24	38.55	0.0463	8.35
Nikon_D90	43.76	38.71	0.0441	7.63
Nokia_N900	44.05	38.52	0.0441	6.83
Olympus_E_PL2	43.84	38.63	0.0436	7.37
Pentax_K_5	43.93	38.63	0.0436	7.63
Pentax_Q	44.30	38.65	0.0437	6.97
GS3-U3-50S5C-C	42.85	38.76	0.0542	9.30
GS2-FW-14S5C-C	44.19	38.94	0.0425	7.35
Phase_One	43.41	38.09	0.0465	7.56
SONY_NEX_5N	44.18	38.69	0.0439	7.05

Table 2. Comparison of the different initializations of the AEM-I, with the best results in bold.

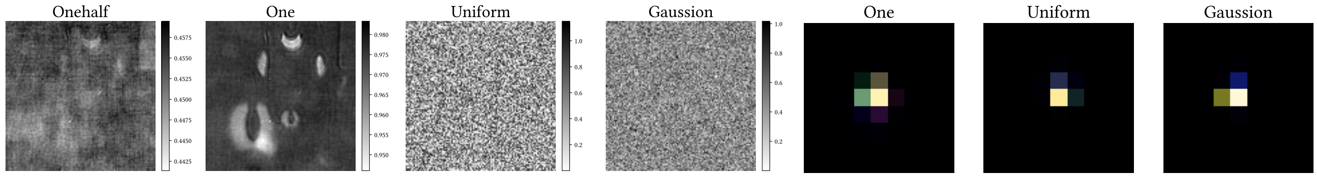
Initialization	PSNR↑	PSNR-SI↑	SAM↓	ERGAS↓
One	44.64	38.58	0.0411	6.85
Onehalf	**45.08**	**38.95**	0.0405	6.24
Uniform	44.57	37.77	**0.0394**	5.84
Gaussian	44.97	38.30	0.0407	**5.79**

Table 3. Comparison of the different initializations of the PEM-I, with the best results in bold.

Initialization	PSNR↑	PSNR-SI↑	SAM↓	ERGAS↓
One	44.49	38.23	0.0419	6.67
Uniform	44.87	38.78	0.0412	6.39
Gaussian	**44.90**	**38.68**	**0.0398**	**6.28**

As shown in Table 2, the all-0.5 initialization setting has the highest PSNR and PSNR-SI test results, so we use the all-0.5 initialization as the experimental setup for the AEM-I. As no apparent features are observed after compositing the RGB image from the 31-channel mask, we select a channel mask for cropping to show the central region, as shown in Figure 2a. The constant-initialized mask is distinguishable from channel to channel, and the random-initialized mask remains morphologically randomly distributed.

Initialization of PEM-I. We utilize a constant initialization



(a) Mask

(b) PSF

Figure 2. Visual comparison of masks of AEM-I and PSFs of PEM-I. The mask is zoomed in for better visualization.

Table 4. Comparison of the different initializations of the WEM-I w/ P.C., with the best results in bold.

Initialization	**PSNR↑**	**PSNR-SI↑**	**SAM↓**	**ERGAS↓**
Random	**44.26**	**38.84**	**0.0432**	**5.37**
Constant	42.19	35.62	0.0522	8.67

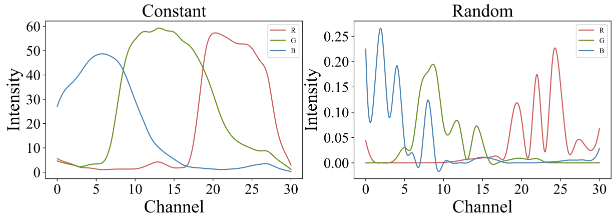


Figure 3. Visual comparison of WEM-I w/ P.C. response curves with different initializations.

of all ones and uniform and Gaussian random initialization for PEM-I.

Table 3 concludes that the PEM-I model performs best with Gaussian random initialization for all evaluation metrics. Therefore, we select the Gaussian random initialization as the experimental setup for the PEM-I. Figure 2b shows a composite RGB image of the 31-channel PSFs obtained using this initialization, which exhibits a central convergence pattern that leads to the best imaging quality.

Initialization of WEM-I. In WEM-I, a linear layer is employed for encoding instead of a fixed RGB filter response function. We try constant initialization of the response curve of the FLIR BFS_U3_04S2C_C camera and random initialization to initialize the weights of the convolution kernel in the WEM-I with positive constraints (WEM-I w/ P.C.).

Table 4 illustrates that random initialization yields better imaging performance than constant initialization. Moreover, the response curve for constant initialization remains unchanged from the initialization settings, while the response curve for random initialization has a distinctive shape, as shown in Figure 3. Consequently, random initialization is used in WEM-I and WEM-I w/ P.C experiments.

3.3. Selection of Loss Functions

In addition to the MAE loss function used in the main text experiments, we also compare the effects of two additional loss functions on the results: MSE loss and ERGAS loss.

Table 5 demonstrates that switching the loss function has no impact on the consistency of individual systems. MAE loss is the most appropriate loss function for training joint encoder-decoder optimization computational spectral imaging systems, compared to MSE loss and ERGAS loss. Therefore, we select the MAE loss as the loss function for the experiments in the main text. When MSE loss is used as the loss function, the PEM-I encounters convergence problems. This is because MSE loss is not well-suited for models with a high degree of freedom. ERGAS loss provides only a slight improvement in the ERGAS metric.

3.4. PSF Sizes and Kernel Sizes of PEM-I

We conduct experiments with varying PSF sizes and convolutional kernel sizes of the decoding model. The comparison results are shown in Table 6 and Figure 4. The results indicate that smaller learnable PSF and kernel sizes in the decoding model can lead to greater convergence of PSF and improve performance.

In summary, to achieve clear imaging quality in the PEM, a concentrated PSF must be designed while maintaining a slight variance of the PSF across the channels to preserve spectral information.

3.5. Validation of ICVL Dataset

During training, we use the common practice of cropping each image into multiple 512 × 512 patches, with sufficient data to cover the training requirements. We conduct three additional experiments under the WEM-P setting, reducing the training set by 10% and 20%, and repartitioning the training and validation sets. As shown in Table 7, the fluctuation of the test results is within 1 dB for both reducing the amount of data in the training set and repartitioning the dataset, proving the validity of the training data.

Table 5. Spectral imaging performance of systems using UEMs on different loss functions, with the best results in bold.

	Physical Model					Ideal Model				
Loss	Encoding Model	PSNR↑	PSNR-SI↑	SAM↓	ERGAS↓	Encoding Model	PSNR↑	PSNR-SI↑	SAM↓	ERGAS↓
MAE	**WEM-P**	44.96	38.81	0.0399	6.23	**WEM-I**	**45.25**	**38.51**	**0.0328**	**5.23**
	AEM-P	40.21	32.66	0.0491	8.96	**AEM-I**	44.64	38.58	0.0411	6.85
	PEM-P	33.75	26.02	0.0655	16.62	**PEM-I**	44.9	38.68	0.0398	6.28
MSE	**WEM-P**	**43.94**	**37.74**	0.0465	6.73	**WEM-I**	43.52	36.77	**0.0448**	**6.63**
	AEM-P	38.41	30.78	0.0611	10.47	**AEM-I**	43.48	37.2	0.0485	7.03
	PEM-P	33.71	25.99	0.0689	16.7	**PEM-I**	43.16	36.75	0.0498	7.15
ERGAS	**WEM-P**	**44.49**	**38.24**	**0.0427**	6.3	**WEM-I**	44.11	38.29	0.0428	**5.14**
	AEM-P	39.9	32.37	0.0523	8.98	**AEM-I**	43.77	38.12	0.0448	5.4
	PEM-P	34.14	26.77	0.0690	15.99	**PEM-I**	44.43	38.27	0.0430	6.43

Table 6. Comparison of the imaging quality of systems using the PEM-I and the Sim-Conv-Net with different PSF and kernel size. The best result is marked in bold, and the second-best result is underlined in each column.

PSF Size	Kernel Size	PSNR↑	PSNR-SI↑	SAM↓
3	3	**44.75**	**38.69**	**0.0409**
3	5	44	37.54	0.0421
3	7	42.43	35.57	0.0471
3	9	41.54	34.35	0.0483
9	3	44.55	38.32	0.0417
9	5	44.01	37.47	0.0421
9	9	40.28	32.91	0.0512
9	15	34.14	26.52	0.0776
16	3	44.4	38.05	0.0419
16	5	40.43	33.33	0.0568
64	3	40.6	33.24	0.0509

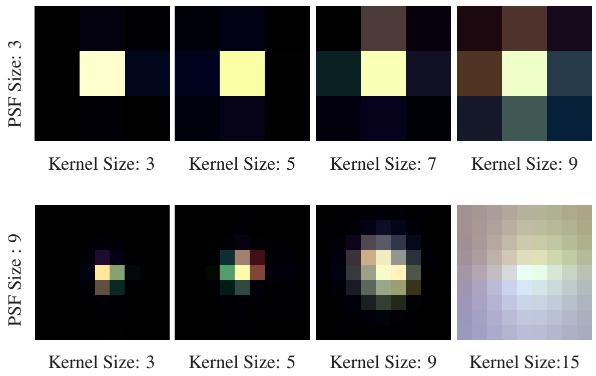


Figure 4. Visual comparison of RGB PSFs of PEM-I with different sizes of learnable PSF and different convolutional kernel sizes of the decoding model.

Table 7. Spectral imaging performance of systems using the WEM-P on ICVL datasets.

Dataset	PSNR↑	PSNR-SI↑	SAM↓	ERGAS↓
ICVL	44.96	38.81	0.0399	6.23
ICVL-90%Training set	44.77	38.76	0.0402	6.36
ICVL-80%Training set	44.18	38.36	0.0437	6.59
ICVL-Redividing	44.65	38.65	0.0413	6.49

4. Visualization of Ideal Encoding Patterns

Due to the limited space, we demonstrate a few channels of PSF and mask in the main paper. We further provide the full version here. Figure 5 shows the full visualization of the 31 channels of the AEM-I masks. Figure 6 shows the full visualization of the 31 channels of the PEM-I PSFs.

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 (ICCV)*, pages 2651–2660, 2021. 1

[2] BAYES. An essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3-4):296–315, 1958. 2

[3] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10193–10202, 2019. 1

[4] Weisheng Dong, Peiyao Wang, Wotao Yin, Guangming Shi, Fangfang Wu, and Xiaotong Lu. Denoising prior driven deep neural network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 41(10): 2305–2318, 2018. 1

[5] Jun Jiang, Dengyu Liu, Jinwei Gu, and Sabine Süsstrunk. What is the space of spectral sensitivity functions for digital color cameras? In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 168–179, 2013. 3

[6] Lingen Li, Lizhi Wang, Weitao Song, Lei Zhang, Zhiwei Xiong, and Hua Huang. Quantization-aware deep optics for diffractive snapshot hyperspectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19780–19789, 2022. 1

[7] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1375–1385, 2020.

[8] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018.

[9] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1386–1396, 2020. 1

[10] Lizhi Wang, Chen Sun, Ying Fu, Min H Kim, and Hua Huang. Hyperspectral image reconstruction using a deep spatial-spectral prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8032–8041, 2019. 1

[11] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3929–3938, 2017.

[12] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3217–3226, 2020.

[13] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 44(10):6360–6376, 2021. 1

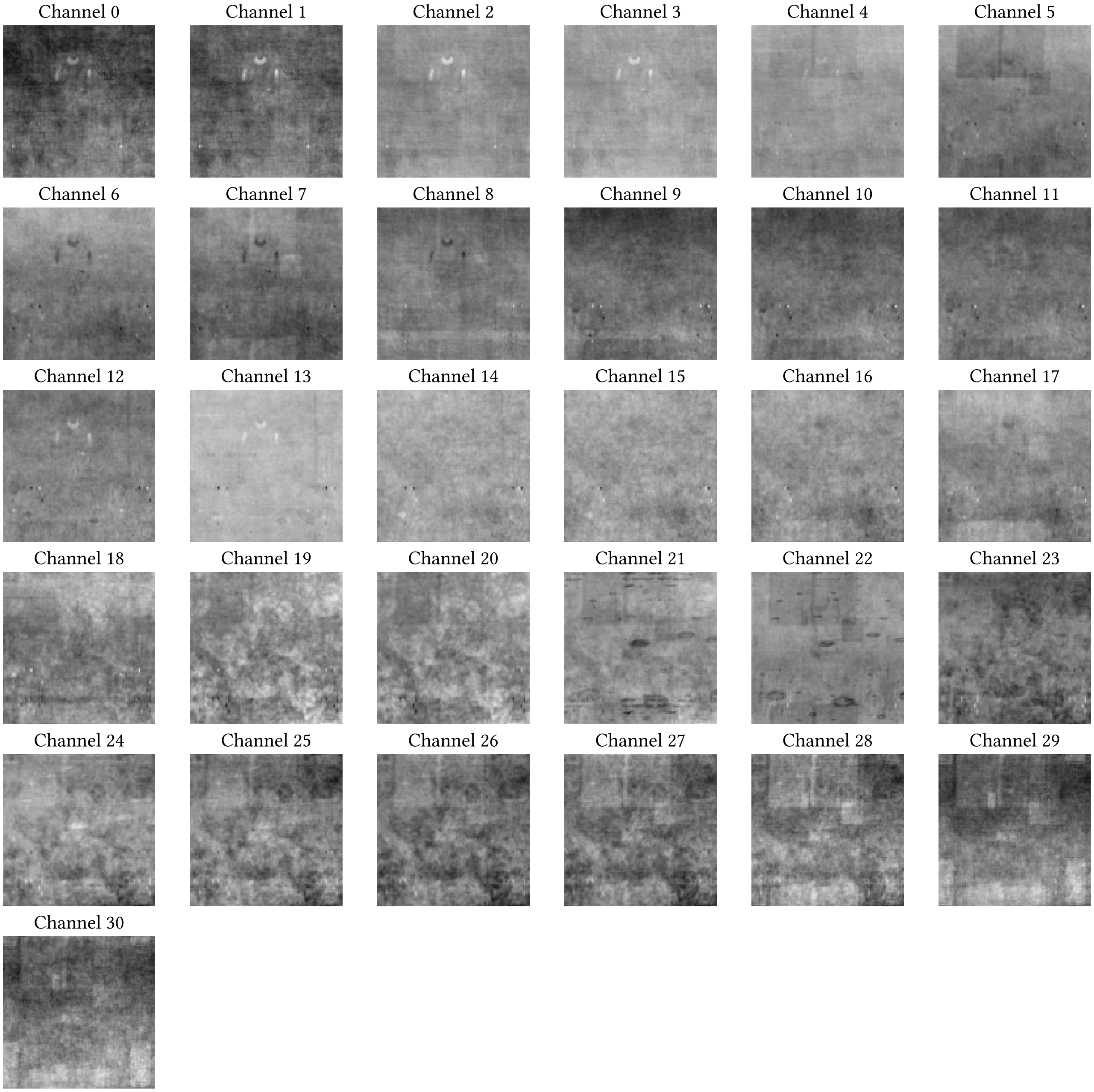


Figure 5. Visualization of the AEM-I masks.

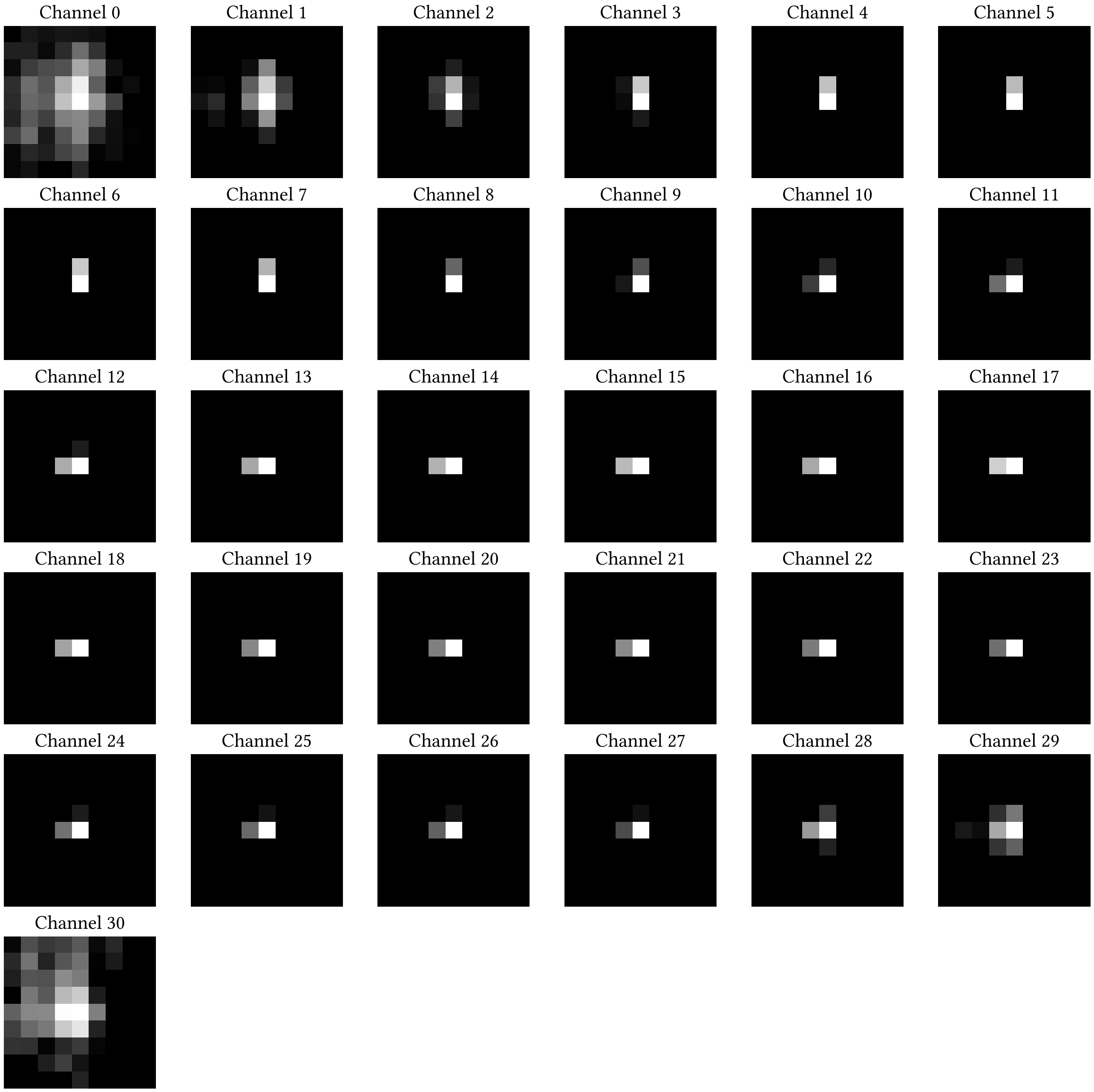


Figure 6. Visualization of the PEM-I PSFs.