S^2-TRANSFORMER FOR MASK-AWARE HYPERSPECTRAL IMAGE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The technology of hyperspectral imaging (HSI) records the visual information upon long-range-distributed spectral wavelengths. A representative hyperspectral image acquisition procedure conducts a 3D-to-2D encoding by the coded aperture snapshot spectral imager (CASSI), and requires a software decoder for the 3D signal reconstruction. By observing this physical encoding procedure, two major challenges may stand in the way of a high-fidelity reconstruction: (i) To obtain 2D measurements, CASSI dislocates multiple channels by disperser-titling and squeezes them onto the same spatial region, yielding an *entangled data loss*. (ii) The physical coded aperture (mask) will lead to a *masked data loss* by selectively blocking the pixel-wise light exposure. To tackle these challenges, we propose a spatial-spectral (S^2-) transformer architecture with a mask-aware learning strategy. Firstly, we simultaneously leverage spatial and spectral attention modelings to disentangle the blended information in the 2D measurement along both two dimensions. A series of Transformer structures across spatial & spectral clues are systematically designed, which considers the information inter-dependency between the two-fold cues. Secondly, the masked pixels will induce higher prediction difficulty and should be treated differently from unmasked ones. Thereby, we adaptively prioritize the loss penalty attributing to the mask structure by inferring the difficulty-level upon the mask-aware prediction. Our proposed method not only sets a new state-of-the-art quantitatively, but also yields a better perceptual quality upon structured areas. Code and pre-trained models are available at `https://anonymous.4open.science/r/S2-transformer-HSI-FEBF`.

1 INTRODUCTION

The technology of hyperspectral imaging (HSI) records pixels of the scene across a wide range of spectrum. The obtained hyperspectral images enable not only high spatial resolutions, but also fine wavelength resolutions. Due to their expressiveness in both domains, hyperspectral images are widely used in applications of biomedicine, remote sensing, astronomy (Fu et al., 2018; 2019; Lu and Fei, 2014; Suo et al., 2021) *etc*. For hyperspectral image acquisition, one of the most popular systems is the coded aperture snapshot compressive imager (CASSI) (Gehm et al., 2007; Wagadarikar et al., 2008). It operates as an optical encoder, which compresses the 3D hyperspectral signals into 2D measurements, and requires a software decoder for the signal reconstruction (Yuan et al., 2021). Concretely, how to precisely reconstruct the hyperspectral images draws lots of research attention.

From iterative-based algorithms (Bioucas-Dias and Figueiredo, 2007b; Liu et al., 2018; Ma et al., 2019) to novel deep learning-based network designs (Cai et al., 2022a; Huang et al., 2021; Meng et al., 2020b; Wang et al., 2021a), previous works made remarkable progresses in tackling the inverse problem of the CASSI-based lossy compression (Yuan et al., 2021). Different from them, in this work, we propose to trace the data loss by directly observing the physical compressive procedure. In light of this, two types of the challenges are presumed. (1) CASSI compresses the information of multiple wavelengths to a single 2D measurement by the channel-wise entangling and addition (implemented by a disperser), which leads to the *entangled data* loss. (2) CASSI blocks the light with a physical binary mask for signal encoding, yielding the *masked data* loss upon some pixels.

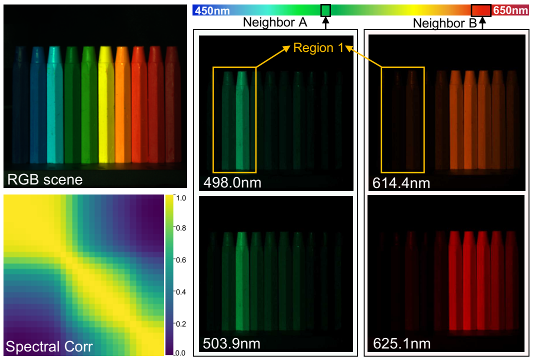


Figure 1: Hyperspectral image characteristics. Spatial correlations vary from adjacent wavelengths (*e.g.*, `neighbor A`, `B`) to distant bands (`region 1`). The bottom-left correlation matrix upon multiple sampled wavelengths (*i.e.*, 28) defines a unique spectral distribution of a hyperspectral image. Spatial and spectral properties are inter-dependent.

For the entangled data loss, we resort to exploit the nature of hyperspectral images for a better 2D-to-3D retrieval. As shown in Fig. 1, hyperspectral data is both *spatial and spectral informative*. Spatially, different wavelengths detect different spatial patterns. Also, spatial textures among adjacent wavelengths (*e.g.*, 498.0nm and 503.9nm in `neighbor A`, 614.4nm and 625.1nm in `B`) are highly correlated. Such a correlation diminishes among more far-between wavelengths as compared by `region 1`. From spectral perspective, each hyperspectral image possesses a unique spectral distribution, owning to its specific spatial content, *e.g.*, `RGB Scene` and `Spectral Corr` in Fig. 1. Thereby, popular off-the-shelf spatial/spectral transformer designs (Dosovitskiy et al., 2020; Cai et al., 2022a) may show limitations in fully exploiting the hyperspectral data: (1) spatial attention only uses a single attention matrix to describe spatial contents at different channels, failing to adapt the matrix to the spectrum variation. Multi-head empowers such a diversity but is hand-crafty defined. (2) Spectral attention considers the channel-wise relationships by the attention matrix, where each entry globally considers a channel but omits the internal pixel-wise relation. Furthermore, both attentions are inter-dependent, which also requests a proper interaction in-between. For the masked data loss, we take the pixel-wise reconstruction difficulty as an indicator. As shown in Fig. 5 (b), a mask pattern appears on the difficulty map by existing method, uncovering different-level of uncertainties between masked/unmasked pixels.

In this study, we propose a spatial-spectral (S^2-) transformer with a mask-aware learning strategy. On the one hand, we present parallel and sequential spatial-spectral attention structures, which could serve as generic solutions for hyperspectral signal modeling. On the other hand, we propose a mask-aware learning strategy to explicitly consider pixel-wise reconstruction difficulties considering the *masked data loss*, which shares the similar spirit of Kendall and Gal (2017); Ning et al. (2021), *i.e.*, pixels with higher "uncertainty" (regression difficulty) should be prioritized for loss penalty. Specifically, we firstly obtain a difficulty-level cube and then adaptively penalize the pixel-wise reconstruction weighted by this cube. The contributions are as follows:

- By observing the optical signal acquisition procedure, we uncover and define two types of data loss that may impede a high-fidelity hyperspectral image reconstruction.
- For the *entangled data* loss, we propose an S^2-Transformer to systematically investigate different self-attention structures adapting to hyperspectral image characteristics.
- For the *masked data loss*, we present a novel mask-aware learning strategy, which evidentially improves the perceptual quality of the reconstruction. Extensive experiments demonstrate the promising performance of the proposed method over the state-of-the-art methods. Besides, a better perceptual quality upon masked regions of the proposed method is also showcased.

2 Related Work

Following the advanced compressive sensing theories (Candès et al., 2006; Donoho, 2006), a series of HSI technologies propose to capture the 3D hyperspectral images with 2D detectors, which enjoys short acquisition time, low-cost consumption, and low-power usage (Yuan et al., 2021). Among existing implementations, CASSI uses a coded aperture and a prism to implement the spectral modulation, which serves as one of the most popular optical designs due to its reliability and simplicity (Yuan et al., 2015). To reconstruct the hyperspectral cube from the 2D measurements by CASSI, various iterative prior-based algorithms (Bioucas-Dias and Figueiredo, 2007a; Figueiredo et al., 2007; Wang et al., 2016; Yang et al., 2015) have been proposed. Among them, the DeSCI (Liu et al., 2018) developed a rank minimization method to tackle the problem. Recent work Wang et al. (2019a) employs data-driven priors to learn the hyperspectral nature. Recently, novel deep learning-based methods (Cai et al., 2022a; Meng et al., 2020c; Miao et al., 2019; Wang et al., 2019a;b) have been introduced for a high-fidelity reconstruction. Among them, the λ-net (Miao et al., 2019) exploits the self-attention in the spatial domain. The TSA-Net (Meng et al., 2020b) introduces the self-attention in the spectral field for the first time. The DGSMP (Huang et al., 2021) formulates the reconstruction as a MAP problem and effectively learns the prior. HDNet (Hu et al., 2022) mainly regularizes the reconstruction in the frequency domain by performing discrete Fourier transform (DFT). It envolves the spatial-spectral learning upon convolutional structure for a better feature extraction but fails to disassemble the spatial and spectral clues for the network design. The MST (Cai et al., 2022a) sets the state-of-the-art by computing the self-attention in the spectral domain. However, it neglects to model spatial dependencies and thus may be limited to fully exploit the hyperspectral characteristic and spatial-spectral inter-dependencies. Besides, by flexibly combining deep networks, deep unfolding methods (Cai et al., 2022b; Wang et al., 2020) provide a promising direction.

3 Method

We firstly give a preliminary knowledge of hyperspectral imaging. We also uncover the potential challenges that may set the bottleneck for the reconstruction performance. Followed by, we propose an S^2-Transformer architecture with mask-aware learning strategy as a corresponding solution.

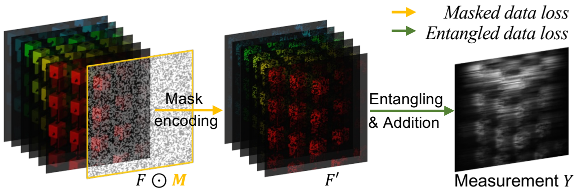


Figure 2: CASSI pipeline (best viewed in color).

3.1 Preliminary Knowledge

The CASSI-based hyperspectral imaging process consists of an optical encoder and a software decoder. Let $\mathbf{F}{\in}\mathbb{R}^{H\times W\times N_\lambda}$ represent the hyperspectral cube with N_λ discrate wavelengths (spectral channels), $\mathbf{M}{\in}\mathbb{R}^{H\times W}$ denotes the physical coded aperture (mask). As shown in Fig. 2, the whole compression procedure could be simplified as two steps. CASSI firstly encodes the signal by

$$\mathbf{F}' = \mathbf{F} \odot \mathbf{M}, \tag{1}$$

where $\odot$ denotes a pixel-wise multiplication with broadcasting, and $\mathbf{F}'$ denotes a "sponge" cube encoded by the mask. Note that the visual information of certain pixels will be erased according to the mask pattern, *i.e.*, where $\mathbf{M}_{ij}{=}0$, or partially disrupted by the noisy mask values, *i.e.*, $\mathbf{M}_{ij}{\in}(0,1)$, both of which lead to the reconstruction challenge of *masked data loss*. After that, the CASSI titles $\mathbf{F}'$ by y-axis-shearing, specifically, $\mathbf{F}'(x,y,n_\lambda) \to \mathbf{F}'(x, y+d(\lambda-\lambda^*), n_\lambda)$, where λ indicates the wavelength of the n_λ-th spectral channel. The λ^* denotes the pre-defined anchor wavelength, and $d(\cdot)$ defines the shifting principle. We conduct a two-pixel shift for each spectral channel following Meng et al. (2020b). The CASSI finally produces the measurement $\mathbf{Y} \in \mathbb{R}^{H\times(W+d(N_\lambda-1))}$ by

$$\mathbf{Y} = \textstyle\sum_{n_\lambda=1}^{n_\lambda=N_\lambda} \mathbf{F}'(:,:,n_\lambda) + \mathbf{\Omega}, \tag{2}$$

where $\mathbf{\Omega}$ denotes the measurement noise. Notably, the pixel-wise addition by Eq. 2 imposes the second reconstruction challenge of *entangled data loss* as one needs to precisely distinguish the spatial details of a specific wavelength from another one given the single 2D measurement.

From a physical encoding perspective, a high-fidelity reconstruction is largely about properly tackling the above challenges. For the *entangled data loss*, we trace back to the nature of the hyperspectral cube, and design a spatial-spectral attention (S^2-attn) mechanism accordingly in expectation to disassemble the spectral signals out of the measurement. We systematically discuss this part in Section 3.3. In Section 3.4, we explicitly measure the pixel-wise difficulty-level owning to the *masked data loss*, and propose a curriculum training strategy (Bengio et al., 2009; Wang et al., 2021b), mask-aware learning, without additional training cost (*i.e.*, time, parameters) introduced.

3.2 Overall Architecture

The proposed spatial-spectral (S^2-) transformer primarily builds upon the self-attention mechanism (Vaswani et al., 2017). Following the merit of Meng et al. (2020a;b); Wang et al. (2021a), we initialize the network input $\mathbf{Y}' \in \mathbb{R}^{H\times W\times N_\lambda}$ by measurement $\mathbf{Y}$ via a `shift` operation for channel expansion and a pixel-wise production with mask

$$\mathbf{Y}'(:,:,n_\lambda) = \mathbf{Y}(:, d(\lambda-\lambda^*) : d(\lambda-\lambda^*)+W) \odot \mathbf{M}, \tag{3}$$

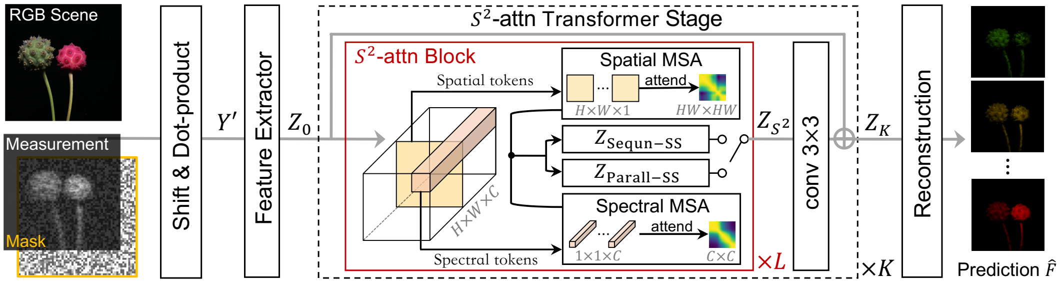


Figure 3: An overview of the proposed spatial-spectral (S^2-) transformer. The network takes the 2D measurement $\mathbf{Y}$ with the mask $\mathbf{M}$ as input and retrieves the hyperspectral image $\widehat{\mathbf{F}}$. It mainly contains K stages, where each consists of L S^2-attn blocks. Both sequential ($\mathbf{Z}_{\texttt{Sequn-SS}}$) and paralleled ($\mathbf{Z}_{\texttt{Parall-SS}}$) blocks employ spatial and spectral multi-head self-attention (MSA).

where $\odot$ denotes a Hadamard product. Next, the whole network is given by $\widehat{\mathbf{F}} = f(\mathbf{\Theta}; \mathbf{Y}')$, where $\mathbf{\Theta}$ denotes the learnable parameters in the network and $\widehat{\mathbf{F}}$ represents the reconstruction result. As shown in Fig. 3, the proposed network is mainly composed of 1) a feature extractor by a CONV3×3 layer, producing $\mathbf{Z_0} \in \mathbb{R}^{H\times W\times C}$. Let C denote the number of embedding channels, which should be large enough to provide the redundancy for the spectrum correlation exploitation. 2) A reconstruction head employs a CONV3×3 layer, which maps the embedded space to the hyperspectral domain. 3) K consecutive S^2-attn transformer stages. Each stage is characterized by a residual structure. Concretely, both L concatenated S^2-attn blocks and one CONV3×3 layer are governed by an identity connection. Let $\mathbf{Z}_k \in \mathbb{R}^{H\times W\times C}$ denote the output of the k-th stage, k=$\{1, ..., K\}$. The underlying mapping function $f_{\texttt{stg}}(\cdot)$ of the stage could be defined as

$$\mathbf{Z}_k = f_{\texttt{stg}}(\mathbf{Z}_{k-1}) = \mathbf{Z}_{k-1} + \texttt{CONV}(f_{\texttt{blks}}(\mathbf{Z}_{k-1})), \tag{4}$$

where $\mathbf{Z}_{k-1}$ denotes the output feature embedding of the (k-1)-th stage. We further expand the mapping of consecutive blocks $f_{\texttt{blks}}(\cdot)$ in Eq. 4 by

$$\mathbf{Z}^{(L)}_{S^2} = f_{\texttt{blks}}(\mathbf{Z}_{k-1}) \text{ where } \mathbf{Z}^{(l)}_{S^2} = f_{\texttt{blk}}(\mathbf{Z}^{(l-1)}_{S^2}), \tag{5}$$

where l=$\{1, ..., L\}$, and $f_{\texttt{blk}}(\cdot)$ expresses the mapping of a single S^2-attn block. Notably, to avoid the quadratic complexity in the traditional self-attention calculation (Vaswani et al., 2017), we conduct the window partition (Liang et al., 2021; Liu et al., 2021) toward the embedded feature cubes.

By observation, hyperspectral images are *both spatially and spectrally informative*. Also, the spatial and spectral characteristics are *inter-dependent*, which might contribute to the 3D signal retrieval hindered by the *entangled data loss*. We efficiently exploit the cues by the S^2-attention blocks.

3.3 S^2-ATTN BLOCK

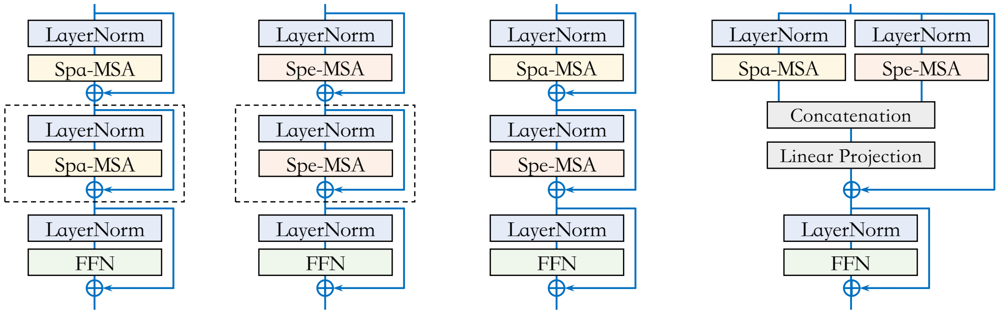
In this section, we firstly give a brief introduce toward both spatial and spectral attention. We also uncover their advantages and potential limitations underlying hyperspectral characteristics, respectively. Furthermore, we propose hybrid spatial-spectral attention structures, all of which constitute a systematic discussion of the attention mechanisms for hyperspectral images.

Spatial Attention. Since proposed by Vaswani et al. (2017) and developed in Dosovitskiy et al. (2020), the frequently-used spatial attention (Spa) has been evolved with key components of multi-head self-attention (MSA), layer normalization (Ba et al., 2016), and feed-forward net (FFN). As shown in Fig. 4 (a), given the input $\mathbf{Z} \in \mathbb{R}^{H\times W\times C}$, a typical structure firstly performs tokenization and obtains $\mathbf{Z} \in \mathbb{R}^{\frac{HW}{M^2}\times M^2\times C}$, where M is the window size. Then the mapping is conducted by

$$\mathbf{Z} = f_{\texttt{Spa-MSA}}(\texttt{LN}(\mathbf{Z})) + \mathbf{Z}, \mathbf{Z} = f_{\texttt{FFN}}(\texttt{LN}(\mathbf{Z})) + \mathbf{Z}, \tag{6}$$

where $f_{\texttt{FFN}}(\cdot)$ is instantiated by Linear-GELU-Linear-GELU structure, and $f_{\texttt{Spa-MSA}}(\cdot)$ denotes the typical multi-head self-attention (MSA) module, where each head is allocated with C_h=$\lfloor\frac{C}{T}\rfloor$ channels. Let T be the number of heads. We use $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$$\in$$\mathbb{R}^{\frac{HW}{M^2}\times M^2\times C_h}$ to express the query, key, and value upon the feature embedding $\mathbf{Z}$ after the linear projection

$$\mathbf{K} = \mathbf{W^K Z},\ \mathbf{Q} = \mathbf{W^Q Z},\ \mathbf{V} = \mathbf{W^V Z}, \tag{7}$$



(a) Spatial Attn (b) Spectral Attn (c) Sequn-SS Attn (c) Parall-SS Attn

Figure 4: Different spatial-spectral (S^2-) attention blocks, which are readily wrapped by off-the-shelf components, *i.e.*, layer normalization (LN), and multi-head self-attention (MSA). Single-typed attentions (a, b) are enhanced by additional LN-MSA (in dashed boxes) for a comparable complexity.

where $\mathbf{W^K}$, $\mathbf{W^Q}$, and $\mathbf{W^V}$ are learnable parameters. The output of spatial attention $\mathbf{Z}_{\texttt{spa}}$ is

$$\mathbf{A}_{\texttt{spa}} = \texttt{softmax}(\mathbf{KQ}^T/\beta + \mathbf{B}),\ \mathbf{Z}_{\texttt{spa}} = \mathbf{A}_{\texttt{spa}}\mathbf{V}, \tag{8}$$

where β is a learnable scaling factor to counteract for the overlarge values, initialized as $\sqrt{C_h}$, $\mathbf{B}$ represents a learnable position bias matrix following Bao et al. (2020); Hu et al. (2018), and $\mathbf{A}_{\texttt{spa}} \in \mathbb{R}^{\frac{HW}{M^2} \times M^2 \times M^2}$ stands for $\frac{HW}{M^2}$ partitioned self-alignment matrices (Tay et al., 2021). The outputs of T heads will be concatenated afterward.

Concretely, given the spatial tokens $\mathbf{Z}_{\texttt{tok}} \in \mathbb{R}^{1\times 1\times C_h}$, the spatial attention benefits the reconstruction by 1) taking well advantage of semantic clues of each pixel, 2) enabling interactions among neighbored embedding within each head, which meets with the nature that spectrum-adjacent spatial contents are highly correlated (neighbor A, B in Fig. 1). However, it fails to describe the spectral variations in each head. Besides, manually dividing the embedding into T heads is quite inflexible.

Spectral Attention. We also leverage the self-attention in spectral domain (Spe). The spectral attention module shares a similar architecture, *i.e.*, Eq. 6, but utilizes $f_{\texttt{Spe-MSA}}(\cdot)$, where multi-head mechanism is still employed to facilitate the spectral dependency modeling. We compute query, key, and value upon Eq. 7 and followed by, the spectral attention $\mathbf{Z}_{\texttt{spe}} \in \mathbb{R}^{\frac{HW}{M^2} \times M^2 \times C_h}$ is

$$\mathbf{A}_{\texttt{spe}} = \texttt{softmax}(\mathbf{K}^T\mathbf{Q}/\beta + \mathbf{B}),\ \mathbf{Z}_{\texttt{spe}} = \mathbf{V}\mathbf{A}_{\texttt{spe}}, \tag{9}$$

where $\mathbf{A}_{\texttt{spe}} \in \mathbb{R}^{\frac{HW}{M^2} \times C_h \times C_h}$ stands for $\frac{HW}{M^2}$ partitioned self alignment matrices upon the spectral domain. We scale the matrix multiplication by a learnable scalar β. $\mathbf{B}$ represents a relative position bias matrix (Shaw et al., 2018). According to the tokenization, the spectral attention differs from the spatial one by using $\mathbf{Z}_{\texttt{tok}} \in \mathbb{R}^{M\times M\times 1}$. On the one hand, it captures the data-dependent spectral correlations varying to the $H \times W$ spatial content, *i.e.*, concatenation of $M \times M$ windows. On the other hand, by observing all pixels from a single channel, it can better abstract the inherent spectrum principle underlying discrete wavelengths. However, each token $\mathbf{Z}_{\texttt{tok}}$ only attends to a scalar value within $\mathbf{A}_{\texttt{spe}}$, indicating an undesirable data loss for correlation modeling, especially when M becomes large. Moreover, long-range dependencies among pixels fail to be modeled.

Considering the limitations of both attention types, it might be inadequate to solely employ either one for a high-fidelity reconstruction. In the following, we provide two types of hybrid structures, which not only own two-fold advantages, but also explore the spatial-spectral inter-dependencies.

Sequential Spa-Spe Attention. We perform sequential attention (Sequn-SS) by cascading $f_{\texttt{Spa-MSA}}(\cdot)$ and $f_{\texttt{Spe-MSA}}(\cdot)$, with layer normalization and residual strategy keeping. An LN-FFN structure follows the attention module As shown in Fig. 4 (c), the feedforward pass upon input $\mathbf{Z}$ is

$$\mathbf{Z} = f_{\texttt{Spe-MSA}}(\texttt{LN}(\mathbf{Z})) + \mathbf{Z},\ \ \mathbf{Z} = f_{\texttt{Spa-MSA}}(\texttt{LN}(\mathbf{Z})) + \mathbf{Z},\ \ \mathbf{Z}_{\texttt{Sequn-SS}} = f_{\texttt{FFN}}(\texttt{LN}(\mathbf{Z})) + \mathbf{Z}, \tag{10}$$

where $f_{\texttt{Spa-MSA}}(\cdot)$ is given by Eq. 8 and $f_{\texttt{Spe-MSA}}(\cdot)$ is computed by Eq. 9. The Sequn-SS explores the data correlations from spatial and spectral perspectives, and allows interaction between both attention mechanisms. However, the behaviors of both attention types are highly interrelated and either one lacks independence. Notably, the potential superiority of Sequn-SS might attribute to a large model size and higher complexity in comparison to previous network structures with the same block and stage numbers. Accordingly, we empower the Spa and Spe by plugging in LN-MSA modules for a comparable model size and complexity, as shown by dashed boxes in Fig. 4 (a, b).

Paralleled Spa-Spe Attention. Besides the Sequn-SS, concurrently conducting spatial and spectral attention is another straightforward hybrid schema (Parall-SS). The outputs of both attentions will be combined in a learnable mode. Given the initial input $\mathbf{Z}$, we have

$$\mathbf{Z} = \mathbf{W}^{\texttt{cat}}[f_{\texttt{Spe-MSA}}(\texttt{LN}(\mathbf{Z})), f_{\texttt{Spe-MSA}}(\texttt{LN}(\mathbf{Z}))] + \mathbf{Z},\ \mathbf{Z}_{\texttt{Parall-SS}} = f_{\texttt{FFN}}(\texttt{LN}(\mathbf{Z})) + \mathbf{Z}, \tag{11}$$

where $[\cdot]$ denotes the concatenation and we use $\mathbf{W}^{\texttt{cat}}$ to perform the linear projection. The underlying advantages of Parall-SS are (1) both spatial and spectral attentions take effect upon the same input, during which their mutual interference existing in Sequn-SS is minimized. (2) The learnable combination is more expressive in feature fusion. As a result, the proposed Parall-SS enjoys a structural superiority with negligible increase in parameters, in comparison to the other structures.

3.4 Mask-aware Learning

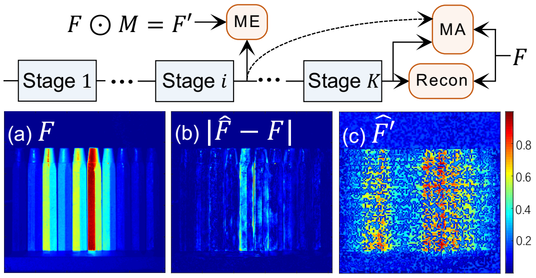


Figure 5: Mask-aware learning strategy. (a) Reference. (b) Reconstruction difficulty by the SOTA method, MST (Cai et al., 2022a). (c) Estimation of encoded signal. Zoom in for a better visualization.

Previous works have well explored different types of learning objectives for a better optimization (Johnson et al., 2016), i.e., $\mathcal{L}_1$ loss (Huang et al., 2021), RMSE loss (Meng et al., 2020b), Spectrum Constancy loss (Cai et al., 2022a), and perceptual loss (Meng and Yuan, 2021), *etc*. By taking advantage of the semantic representations or spectral correlations, they enable promising texture retrieval and perceptual superiority. However, there still exist content-irrelevant artifacts in predictions and unexpected reconstruction difficulty in smooth areas (compared by (a), (b) in Fig. 5), as existing learning objectives neglect the reconstruction difficulty between the masked pixels and unmasked ones and generally treat them equally. Mask encoding yields the spatial data loss. Corresponding pixels should be predicted with higher uncertainty, while the unmasked regions potentially allow an easier retrieval with higher confidence.

Following this intuition, we propose a novel mask-aware learning strategy. The main idea is to firstly determine both encoded and high-frequency regions of original hyperspectral signal, then adaptively emphasize these regions with a higher loss penalty. Our solution is to observe the reconstruction difficulty conditioned by the mask, yielding a novel mask-encoding loss term ME. Besides, the original learning objective $||\widehat{\mathbf{F}} - \mathbf{F}||_1$ is kept to stabilize the training

$$\mathcal{L}_{\texttt{ME}} = \alpha \times \underbrace{||f_{\texttt{head}}(\mathbf{Z}_i) - \mathbf{F}'||_1}_{\texttt{ME}} + \underbrace{||\widehat{\mathbf{F}} - \mathbf{F}||_1}_{\texttt{Recon}}, \tag{12}$$

where $i \in \{1, ..., K-1\}$, and $f_{\texttt{head}}(\cdot)$ is implemented by an LN-CONV. $\mathbf{F}'$ is given by Eq. 1. We use α to balance between two terms. The network not only regresses to the ground truth $\mathbf{F}$, but also allows the masked signal $\mathbf{F}'$ reconstruction upon part of the network structure. According to (c) in Fig. 5, the difference cube $||f_{\texttt{head}}(\mathbf{Z}_i) - \mathbf{F}'||_1$ provides clues for a further pixel prioritization: (1) Masked pattern can not be faithfully reconstructed – the masked regions are zeros, which loses the underlying spatial smoothness and disrupts the semantic meaning. (2) Original texture patterns mainly appear on the unmasked regions, indicating a meaningful and easier information retrieval.

Given different difficulties, ME firstly prioritizes the unmasked region retrieval at pre-training period. Then, the masked regions gain more attention and mainly contributes to the ME term convergence. Meanwhile, an emphasized loss weight is expected when we take it as a denominator, leading to $\mathcal{L}_{\texttt{MA}}$

$$\mathcal{L}_{\texttt{MA}} = \alpha \times \underbrace{||f_{\texttt{head}}(\mathbf{Z}_i) - \mathbf{F}'||_1}_{\texttt{ME}} + \underbrace{\frac{\beta}{||f_{\texttt{head}}(\mathbf{Z}_i) - \mathbf{F}'||_1} \times ||\widehat{\mathbf{F}} - \mathbf{F}||_1}_{\texttt{MA}} + \underbrace{||\widehat{\mathbf{F}} - \mathbf{F}||_1}_{\texttt{Recon}}, \tag{13}$$

where we attenuate the α and signify the masked areas by enlarging β. The $||\widehat{\mathbf{F}} - \mathbf{F}||_1$ is kept to stabilize the training. Finally, a performance boost especially perceptual quality upgrade is expected.

Table 1: PSNR (dB) values by different methods on the benchmark simulation dataset.

Methods	Scene1	Scene2	Scene3	Scene4	Scene5	Scene6	Scene7	Scene8	Scene9	Scene10	*Avg.*
GAP-TV (Yuan, 2016)	26.82	22.89	26.31	30.65	23.64	21.85	23.76	21.98	22.63	23.10	24.36
DeSCI (Liu et al., 2018)	27.13	23.04	26.62	34.96	23.94	22.38	24.45	22.03	24.56	23.59	25.27
TSA-Net (Meng et al., 2020b)	32.03	31.00	32.25	39.19	29.39	31.44	30.32	29.35	30.01	29.59	31.46
DGSMP (Huang et al., 2021)	33.26	32.09	33.06	40.54	28.86	33.08	30.74	31.55	31.66	31.44	32.63
SRN (Wang et al., 2021a)	34.96	35.46	36.18	41.60	32.70	34.70	33.83	32.88	35.09	32.31	35.07
HDNet (Hu et al., 2022)	35.14	35.67	36.03	42.30	32.69	34.50	33.67	32.48	34.89	32.38	34.97
MST (Cai et al., 2022a)	35.40	35.87	36.51	42.27	32.77	34.80	33.66	32.67	35.39	32.50	35.18
S^2-Transformer	**36.17**	**37.57**	**37.29**	**42.96**	**34.40**	**36.44**	**35.41**	**34.50**	**36.54**	**33.57**	**36.48**

Table 2: SSIM values by different methods on the benchmark simulation dataset.

Methods	Scene1	Scene2	Scene3	Scene4	Scene5	Scene6	Scene7	Scene8	Scene9	Scene10	*Avg.*
GAP-TV (Yuan, 2016)	0.7544	0.6103	0.8024	0.8522	0.7033	0.6625	0.6881	0.6547	0.6815	0.5839	0.6993
DeSCI (Liu et al., 2018)	0.7479	0.6198	0.8182	0.8966	0.7057	0.6834	0.7433	0.6725	0.7320	0.5874	0.7207
TSA-Net (Meng et al., 2020b)	0.8920	0.8583	0.9145	0.9528	0.8835	0.9076	0.8782	0.8884	0.8901	0.8740	0.8939
DGSMP (Huang et al., 2021)	0.9152	0.8977	0.9251	0.9636	0.8820	0.9372	0.8860	0.9234	0.9110	0.9247	0.9166
SRN (Wang et al., 2021a)	0.9345	0.9373	0.9476	0.9703	0.9444	0.9512	0.9241	0.9443	0.9414	0.9348	0.9430
HDNet (Hu et al., 2022)	0.9352	0.9404	0.9434	0.9694	0.9460	0.9518	0.9263	0.9406	0.9415	0.9365	0.9431
MST (Cai et al., 2022a)	0.9405	0.9440	0.9525	0.9734	0.9471	0.9553	0.9254	0.9479	0.9491	0.9408	0.9476
S^2-Transformer	**0.9490**	**0.9582**	**0.9567**	**0.9754**	**0.9596**	**0.9654**	**0.9461**	**0.9625**	**0.9592**	**0.9517**	**0.9584**

4 Experiment

4.1 Experimental Settings

Dataset. Following existing methods, we adopt the simulation dataset leveraged by TSA-Net (Meng et al., 2020b). Specifically, the training data origins from CAVE dataset (Yasuma et al., 2010) and a benchmark testing dataset is collected from the KAIST dataset (Choi et al., 2017). The network deals with hyperspectral cubes of the spatial size 256×256 and 28 spectral channels ranging from 450nm∼650nm [1]. The 2D measurements are computationally obtained by following the mathematical model of CASSI in Eq. 1 and Eq. 2. Besides, we also evaluate the proposed method with several 660×714 real-captured measurements collected by Meng et al. (2020b). The corresponding hyperspectral cubes are 660×660×28. Same as Cai et al. (2022a); Meng et al. (2020b); Wang et al. (2021a), we use the same augmented dataset for training. We inject the Gaussian noise to the simulated measurements during training, with an intuition to consider the real measurement noise $\mathbf{\Omega}$.

Implementation Details. Our proposed S^2-Transformer contains K=4 stages, where each consists of L=6 S^2-attn blocks for a high-fidelity reconstruction performance. We let the number of embedding channels C to be 60 and split them into T=6 heads. For partition and cyclic shifting, we employ 8×8 windows. The model is trained for 300 epochs with Adam optimizer (Kingma and Ba, 2014) (β_1=0.9, β_2=0.999), The initial learning rate is 4×10^{-4} and halved every 50 epochs. Alternatively, the total amount of training epochs remains the same when using mask-aware learning strategy. Specifically, the model is pre-trained upon $\mathcal{L}_{\text{ME}}$ for 150 epochs to get a relatively precise approximation for encoded signal $\widehat{\mathbf{F}'}$, to differentiate the difficulties between the masked/unmasked regions. Then we train the model upon $\mathcal{L}_{\text{MA}}$ for the other 150 epochs. The loss weight α is 1.5 and then attenuated to 1.0, and β is 10. No additional computational cost is introduced. Our method is trained on 2 NVIDIA RTX3090 GPUs. All the other settings are kept the same as compared methods for a fair comparison.

Compared Methods. We compare with seven state-of-the-art methods, including iterative-based methods, *e.g.*, DeSCI (Liu et al., 2018), and GAP-TV (Yuan, 2016), CNN-based methods, *i.e.*, TSA-Net (Meng et al., 2020b), SRN (Wang et al., 2021a), DGSMP (Huang et al., 2021), HDNet (Hu et al., 2022), and MST (Cai et al., 2022a). Following previous works, we report the best performance of most compared methods.[2] Notably, we choose the best-performed variants of the compared methods, *i.e.*, `SRN(v1)`, `MST-L`. PSNR and SSIM (Wang et al., 2004) are employed as metric comparisons.

[1] Detailed dataset determination process could be found in Meng et al. (2020b).

[2] Due to the different metric calculations, we re-train the SRN and provide the results in the supplementary.

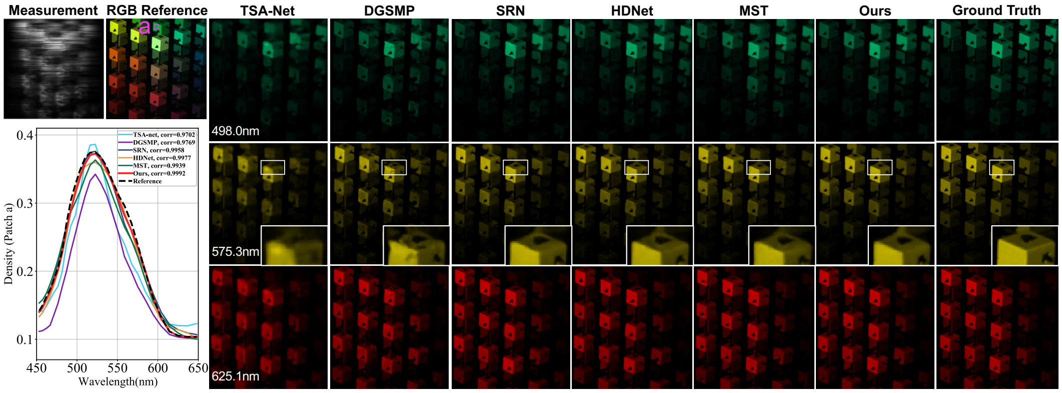


Figure 6: Reconstruction results for a simulation hyperspectral image. Five state-of-the-art methods and the proposed method (second to the right) are presented on 3 out of 28 wavelengths. The RGB reference is shown to demonstrate the color (top-left). The density-vs-wavelength curves (bottom-left) corresponding to the chosen patch (*i.e.*, `patch a`) are plotted to demonstrate the **spectral fidelity**.

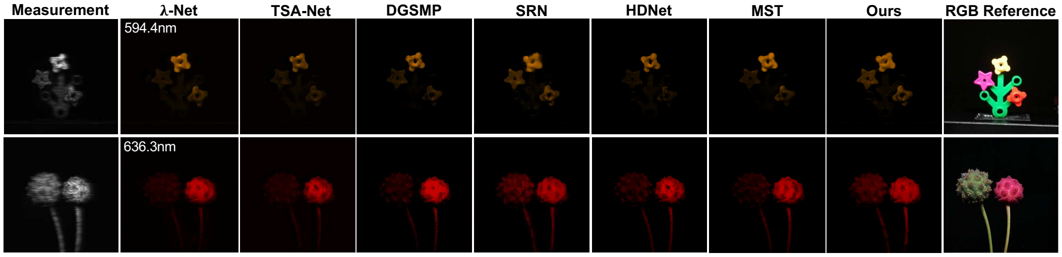


Figure 7: Reconstruction results on real-captured measurements. **Zoom in for a better visualization**.

4.2 HSI Reconstruction Performance

First of all, we metrically compare the proposed S^2-Transformer with popular methods by Tab. 1 and Tab. 2. Our method outperforms MST (Cai et al., 2022a) by 1.30dB/0.0108 in terms of PSNR/SSIM, and achieves 1.51dB/0.0153 improvement in contrast to HDNet (Hu et al., 2022). These comparisons demonstrate the effectiveness of the proposed spatial-spectral attention structure. Besides, we achieve the best SSIM on ten testing hyperspectral images, which indicates a better reconstruction ability on structured regions. One of the potential reasons might be that we explicitly consider the variant reconstruction difficulty-levels underlying the mask structure by the mask-aware learning strategy. We discuss more in Section 4.3. Moreover, we visually compare the proposed method with others upon simulation testing hyperspectral images. As shown in Fig. 6, the proposed method achieves competitive performance with the most recent methods, also enjoys fine-grained retrieval on semantic area (*e.g.*, less crooked edges by the enlarged window). Furthermore, we provide the visual comparison upon real-captured measurements shown in Fig. 7. By observation, the proposed method not only enables a precise reconstruction on matched channels (*e.g.*, red contents recorded by 620nm∼650nm), but also provides clear across-channel content retrievals (*e.g.*, green contents recorded by 620nm∼650nm), owning to the joint modeling upon spatial and spectral domain[3].

4.3 Model Discussion

Spatial-spectral attentions. We conduct ablation studies upon the proposed S^2-Transformer with full model size *i.e.*, K=4 stages, where each contains L=6 blocks. Specifically, we compare the model size (`#params (M)`), computational complexity by floating point operations (FLOPs) and the reconstruction performances. As discussed in the Section 3, directly compare the `Spa` or `Spe` with hybrid ones may lead to unfair analysis. Therefore, we pay more attention to their enhanced versions, *i.e.*, `SpaSpa` and `SpeSpe` upon additional `LN-MSA` modules. As shown in Tab. 3a, the proposed S^2-attention structures achieve relatively higher performances than both `SpaSpa` and `SpeSpe`

[3] We employ a no reference image quality assessment to quantitatively evaluate real data, please see Section B

Table 4: Model size and complexity of different methods.

Types	PSNR	SSIM	#params (M)	FLOPs (G)
TSA-Net (Meng et al., 2020b)	31.46	0.8939	44.25	110.06
DGSMP (Huang et al., 2021)	32.63	0.9166	3.76	646.65
SRN (Wang et al., 2021a)	35.07	0.9430	**1.25**	81.84
HDNet (Hu et al., 2022)	34.97	0.9431	2.37	154.76
MST (Cai et al., 2022a)	35.18	0.9476	2.03	28.15
S^2-Transformer	**36.48**	**0.9584**	1.80	**27.21**

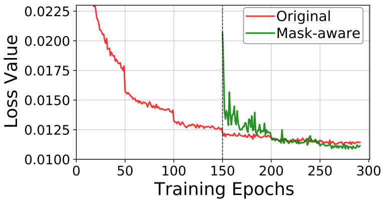


Figure 8: Mask-aware learning curve.

attention mechanisms at cost of the minimum computational overhead, indicating the necessity of jointly modeling from spatial and spectral domains. Specifically, the `Parall-SS` outperforms the `Sequn-SS` by allowing more flexible interactions between the spatial and spectral domain with the learnable concatenation module employed. Consequently, the proposed S^2-attentions achieve better trade-off between the performance and the computational burden. Moreover, we not only discuss among ablated models, but also compare with most recent methods in Tab. 4. Our method requires the smallest model size,and relatively small FLOPs, but enables a better reconstruction performance.

Table 3: Ablation studies conducted upon large model sizes as previous, *i.e.*, K=4, L=6.

(a) Ablation study of different self-attention structures.

Types	`SpaSpa`	`SpeSpe`	`Sequn-SS`	`Parall-SS`
PSNR	35.98	35.18	36.04	**36.31**
SSIM	0.9563	0.9536	0.9565	**0.9569**
FLOPs (G)	37.75	13.52	27.21	27.21
#params (M)	1.60	1.65	1.62	1.80

(b) Loss ablation for mask-aware learning.

`Recon`	`ME`	`MA`	pre-train	PSNR(dB)	SSIM
✓	✗	✗	✗	36.31	0.9569
✓	✓	✗	✗	36.30	0.9569
✓	✓	✓	✗	34.88	0.9496
✓	✓	✓	✓	**36.48**	**0.9584**

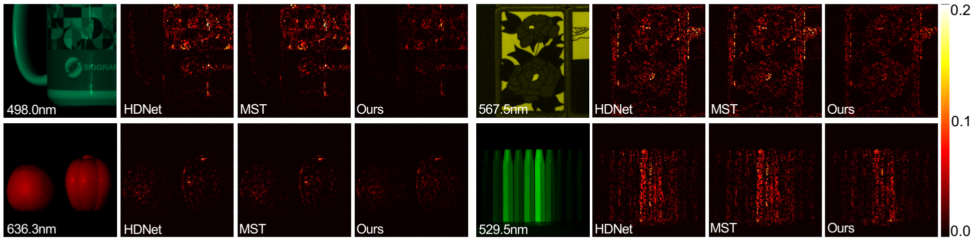


Figure 9: Evidenced reconstruction difficulty upon masked regions. Lower intensity is better.

Mask-aware learning. We firstly discuss the mask-aware learning strategy via a loss ablation study in Tab. 3b. The network settings are kept the same as in Tab. 3a with `Parall-SS` attention. A complete mask-aware learning treatment enables 0.17dB/0.0015 improvement over baseline method. Besides, the perceptual quality improvement is also apparent, considering that mask-aware learning adaptively emphasizes the masked regions with larger penalty. This could be further evidenced by the absolute difference between the predictions and the references (evidenced reconstruction difficulty) upon masked areas. In Fig. 9, we compare among MST (Cai et al., 2022a), HDNet (Hu et al., 2022), and the proposed method. Our method enables a more reliable reconstruction on the masked regions. Besides, we compare the learning curves of `Recon` term under both scenarios in Fig. 8. The red curve corresponds to the traditional $||\widehat{\mathbf{F}} - \mathbf{F}||_1$ objective, while the green curve is jointly affected by `MA` and `ME` terms. By the proposed strategy, not only does the model achieve a lower mean absolute error, but also appears to converge faster, owning to the mask-aware curriculum learning schedule.

5 Conclusion

In this work, we firstly observed and presumed two-fold challenges with regard to data loss owing to CASSI optical encoding procedure. For the first challenge, *i.e.*, entangled data loss, we resorted to exploit the characteristics underlying hyperspectral images to better disentangle the 2D signal. We introduced the S^2-Transformer by systematically discussing several self-attention structures for the hyperspectral image. For the other challenge, *i.e.*, masked data loss, we proposed to model the pixel-wise reconstruction difficulty given the consideration that masked pixels are harder to be reconstructed and thus should be emphasized by the loss penalty. Our proposed method achieves promising performance compared with existing SOTA methods quantitatively and perceptually.

6 Reproducibility Statement

In this section, we summarize the efforts that have been made for the reproducibility. In Section 3.2, we outline the overall architecture of the S^2-Transformer. In Section 3.3, we systematically discuss the proposed spatial-spectral attention mechanisms. In Section 3.4, we introduce the proposed mask-aware learning strategy. Besides, our implementation details and the employed dataset are provided in Section 4.1. Our code, pre-trained models, and results are open-sourced at the anonymous GitHub repostitory: https://anonymous.4open.science/r/S2-transformer-HSI-FEBF/. To facilitate the comparison, the simulation reconstruction results of the SRN, the real results of the MST and HDNet are also available. For the simulation data, we provide the code for full reference image quality assessments, *i.e.*, PSNR and SSIM. For the real data, we provide the code on no reference image quality assessment, *i.e.*, NIQE. Please refer to the website for detailed instructions.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, pages 642–652. PMLR, 2020.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

J.M. Bioucas-Dias and M.A.T. Figueiredo. A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16(12):2992–3004, December 2007a.

José M Bioucas-Dias and Mário AT Figueiredo. A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. *TIP*, 2007b.

Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17502–17511, 2022a.

Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. *arXiv preprint arXiv:2205.10102*, 2022b.

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

Inchang Choi, MH Kim, D Gutierrez, DS Jeon, and G Nam. High-quality hyperspectral reconstruction using a spectral prior. Technical report, 2017.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Mário AT Figueiredo, Robert D Nowak, and Stephen J Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of selected topics in signal processing*, 1(4):586–597, 2007.

Ying Fu, Yinqiang Zheng, Hua Huang, Imari Sato, and Yoichi Sato. Hyperspectral image super-resolution with a mosaic rgb image. *IEEE Transactions on Image Processing*, 27(11):5539–5552, 2018.

Ying Fu, Tao Zhang, Yinqiang Zheng, Debing Zhang, and Hua Huang. Hyperspectral image super-resolution with optimized rgb guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11661–11670, 2019.

Michael E Gehm, Renu John, David J Brady, Rebecca M Willett, and Timothy J Schulz. Single-shot compressive spectral imaging with a dual-disperser architecture. *Optics express*, 15(21): 14013–14027, 2007.

Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3588–3597, 2018.

Xiaowan Hu, Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17542–17551, 2022.

Tao Huang, Weisheng Dong, Xin Yuan, Jinjian Wu, and Guangming Shi. Deep gaussian scale mixture prior for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16216–16225, 2021.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision (ECCV)*, pages 694–711. Springer, 2016.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems (NeurIPS)*, 30, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1833–1844, 2021.

Yang Liu, Xin Yuan, Jinli Suo, David J Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *IEEE transactions on pattern analysis and machine intelligence*, 41(12): 2990–3006, 2018.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 10012–10022, 2021.

Guolan Lu and Baowei Fei. Medical hyperspectral imaging: a review. *Journal of biomedical optics*, 19(1):010901, 2014.

Jiawei Ma, Xiao-Yang Liu, Zheng Shou, and Xin Yuan. Deep tensor admm-net for snapshot compressive imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10223–10232, 2019.

Ziyi Meng and Xin Yuan. Perception inspired deep neural networks for spectral snapshot compressive imaging. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2813–2817. IEEE, 2021.

Ziyi Meng, Shirin Jalali, and Xin Yuan. Gap-net for snapshot compressive imaging. *arXiv preprint arXiv:2012.08364*, 2020a.

Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In *European Conference on Computer Vision (ECCV)*, August 2020b.

Ziyi Meng, Mu Qiao, Jiawei Ma, Zhenming Yu, Kun Xu, and Xin Yuan. Snapshot multispectral endomicroscopy. *Optics Letters*, 45(14):3897–3900, 2020c.

Xin Miao, Xin Yuan, Yunchen Pu, and Vassilis Athitsos. λ-net: Reconstruct hyperspectral images from a snapshot measurement. In *IEEE/CVF Conference on Computer Vision (ICCV)*, 2019.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

Qian Ning, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. Uncertainty-driven loss for single image super-resolution. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Jinli Suo, Weihang Zhang, Jin Gong, Xin Yuan, David J Brady, and Qionghai Dai. Computational imaging and artificial intelligence: The next revolution of mobile vision. *arXiv preprint arXiv:2109.08880*, 2021.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention for transformer models. In *International Conference on Machine Learning (ICML)*, pages 10183–10192. PMLR, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30, 2017.

Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady. Single disperser design for coded aperture snapshot spectral imaging. *Applied optics*, 47(10):B44–B51, 2008.

Jiamian Wang, Yulun Zhang, Xin Yuan, Yun Fu, and Zhiqiang Tao. A new backbone for hyperspectral image reconstruction. *arXiv preprint arXiv:2108.07739*, 2021a.

Lizhi Wang, Zhiwei Xiong, Guangming Shi, Feng Wu, and Wenjun Zeng. Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging. *IEEE transactions on pattern analysis and machine intelligence*, 39(10):2104–2111, 2016.

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)*, 2019a.

Lizhi Wang, Tao Zhang, Ying Fu, and Hua Huang. Hyperreconnet: Joint coded aperture optimization and image reconstruction for compressive hyperspectral imaging. *IEEE Transactions on Image Processing*, 28(5):2257–2270, May 2019b. ISSN 1057-7149. doi: 10.1109/TIP.2018.2884076.

Lizhi Wang, Chen Sun, Maoqing Zhang, Ying Fu, and Hua Huang. Dnu: Deep non-local unrolling for computational spectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Jianbo Yang, Xuejun Liao, Xin Yuan, Patrick Llull, David J Brady, Guillermo Sapiro, and Lawrence Carin. Compressive sensing by learning a Gaussian mixture model from measurements. *IEEE Transaction on Image Processing*, 24(1):106–119, January 2015.

Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K Nayar. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE transactions on image processing*, 19(9):2241–2253, 2010.

Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543. IEEE, 2016.

Xin Yuan, Tsung-Han Tsai, Ruoyu Zhu, Patrick Llull, David Brady, and Lawrence Carin. Compressive hyperspectral imaging with side information. *IEEE Journal of selected topics in Signal Processing*, 9(6):964–976, 2015.

Xin Yuan, David J Brady, and Aggelos K Katsaggelos. Snapshot compressive imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 38(2):65–88, 2021.

SUPPLEMENTARY MATERIAL

In this section, we present the additional results and analyses of the proposed method as follows.

- Additional visual evidence of the mask-aware learning (Section A)
- Quantitative evaluation of real data (Section B)
- Spectral fidelity analysis (Section C)
- High-frequency spectral detail (Section D)
- Additional reconstruction results (Section E)

A MASK-AWARE LEARNING

In this section, we firstly visualize the reconstruction difficulty upon the simulation data. Specifically, the reconstruction difficulty, *i.e.*, Δ, is defined as

$$\Delta = |\widehat{\mathbf{F}} - \mathbf{F}| \odot (\mathbf{1} - \mathbf{M}), \tag{14}$$

where $\odot$ denotes the pixel-wise production, and the mask pixel $\mathbf{M}_{ij} \sim [0, 1]$ with the mask noise considered. The lower mask value $\mathbf{M}_{ij}$ is, the more data losses on this pixel, the reconstruction difficulty is supposed to be large. On the other hand, the lower evidenced reconstruction difficulty observed, the better data retrieval is achieved, and thus the more reliable the prediction is. As compared in Fig. 10 and Fig. 11, mask-aware learning enables a lower reconstruction difficulty.

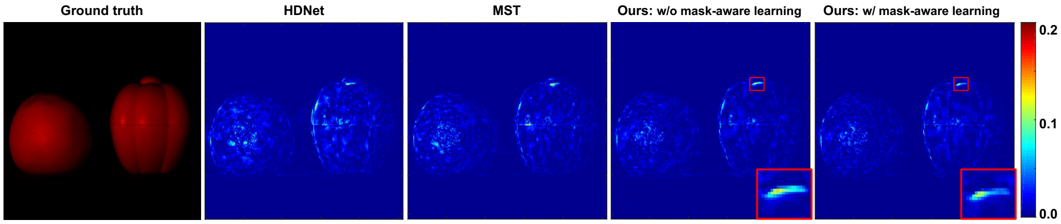


Figure 10: Reconstruction difficulty of the simulation data. Lower is better.

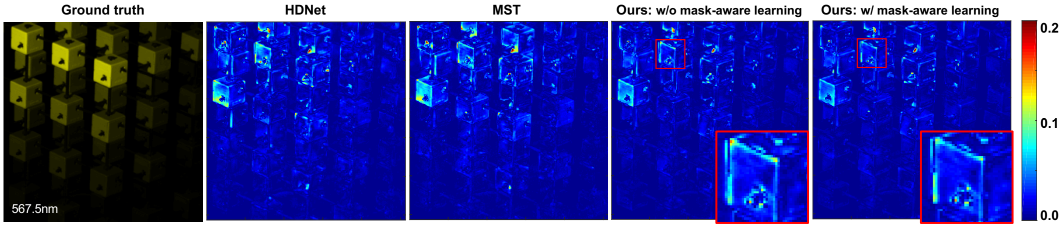


Figure 11: Reconstruction difficulty of the simulation data. Lower is better.

Secondly, we demonstrate the superiority of the mask-aware learning by comparing between the ablated models. In Fig. 12, we visualize upon the simulation data. In Fig. 13, we visualize upon the real data. Please zoom in for a better perceptual comparison.

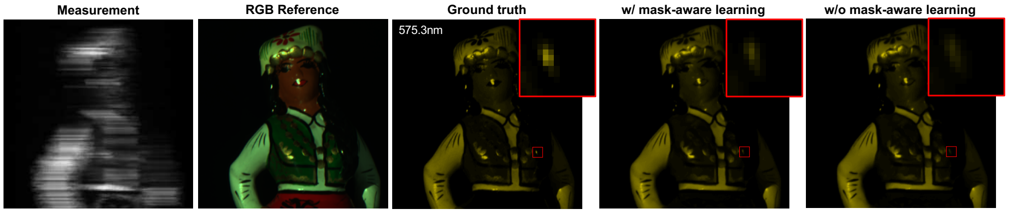


Figure 12: Visual comparison of the mask-aware learning upon simulation data.

B QUANTITATIVE EVALUATION OF REAL DATA

The ground truth of the real hyperspectral reconstruction results are unavailable. The full-reference quantitative evaluation metrics like PSNR and SSIM are not accessible. Therefore, real data are gener-

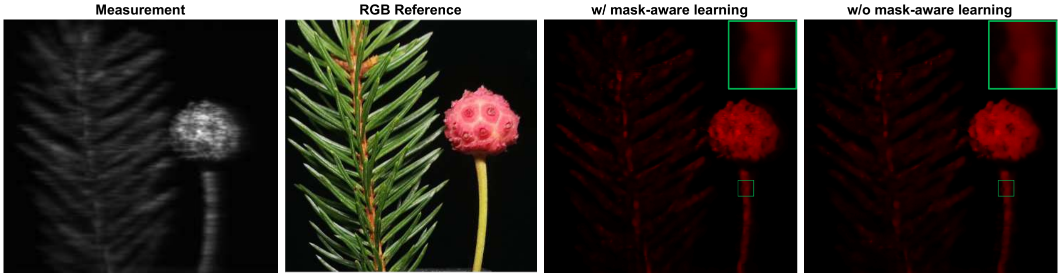


Figure 13: Visual comparison of the mask-aware learning upon real data.

Table 5: Naturalness Image Quality Evaluator (NIQE) evaluation of real hyperspectral images.

Methods	MST (Cai et al., 2022a)	HDNet (Hu et al., 2022)	S^2-Transformer w/o mask-aware	S^2-Transformer w/ mask-aware
NIQE (↓)	6.9219	5.9207	6.0950	**5.8833**

ally visually compared. However, to better evaluate the performance of the proposed method under the practical scenarios, We compute the averaged Naturalness Image Quality Evaluator (NIQE) (Mittal et al., 2012) among the real reconstruction results. As compared in Table 5 below, the proposed method outperforms HDNet and MST. Note that a smaller value indicates a better perceptual quality. The real reconstruction results and evaluation code can be found in the anonymous link.

C Spectral Fidelity Analysis

In this section, we analyze the spectral fidelity by visualizing the spectral correlation matrices (Benesty et al., 2009) of the reconstruction results by the proposed method. As compared in Fig. 14, the spectral correlation matrices of the predictions are highly similar to those of the ground truth.

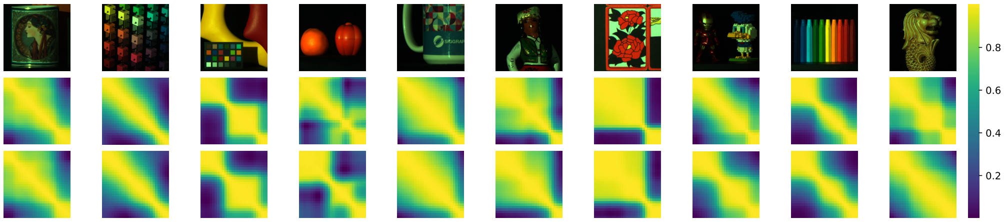


Figure 14: RGB scene of ten benchmark test data (top line), spectral correlation coefficient (Pearson correlation) visualizations by the ground truth (middle line), and the proposed S^2-Transformer (bottom line). Each correlation coefficient matrix is 28×28.

D High-frequency Spectral Detail

In this section, we provide the visualizations based on two simulated (Fig. 16) and one real hyperspectral data (Fig. 15) for the high-frequency spectral detail comparison. By the zoomed-in windows, we found that the proposed method gains advantages by: (1) better sharpness, (2) better luminance, (3) less halo artifact, and (4) more complete content retrieval.

E Additional Reconstruction Results

In this section, we visualize more reconstruction results of the proposed method. In Fig. 17, we present a simulation reconstruction result by comparing with popular methods upon the zoom-in windows, including TSA-Net (Meng et al., 2020b), DGSMP (Huang et al., 2021), SRN (Wang et al., 2021a), HDNet (Hu et al., 2022), and MST (Cai et al., 2022a). In Figs. 18∼23, we demonstrate the reconstruction results upon total 28 spectral channels. More details about the spectral fidelity curve (lower-left in Fig. 17) computation could be found in Meng et al. (2020b).

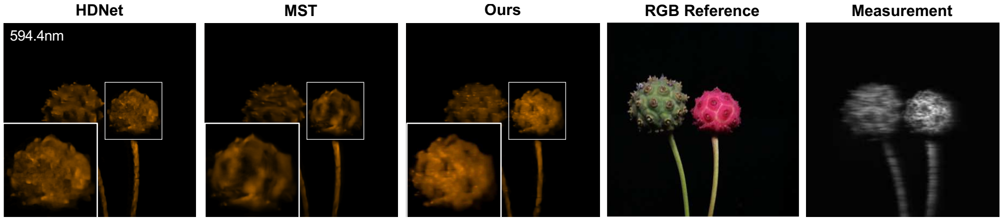


Figure 15: High-frequency spectral detail comparison on real hyperspectral data.

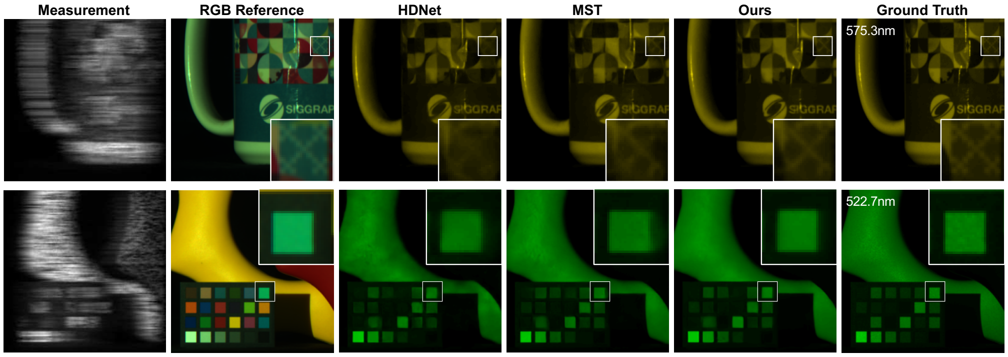


Figure 16: High-frequency spectral detail comparison on two simulation data.

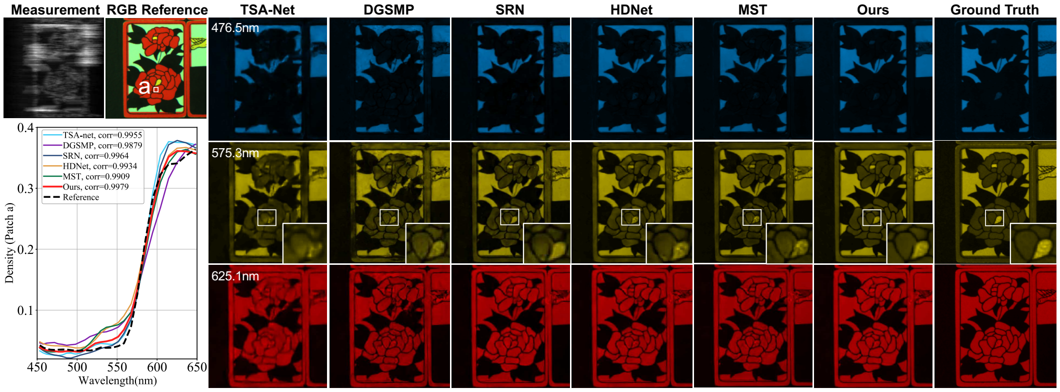


Figure 17: Reconstruction results for a simulation hyperspectral image. Five state-of-the-art methods and the proposed method (second to the right) are presented on 3 out of 28 wavelengths. The RGB reference is shown to demonstrate the color (top-left). The density-vs-wavelength curves (bottom-left) corresponding to the chosen patch (*i.e.*, `patch a`) are plotted to demonstrate the **spectral fidelity**.

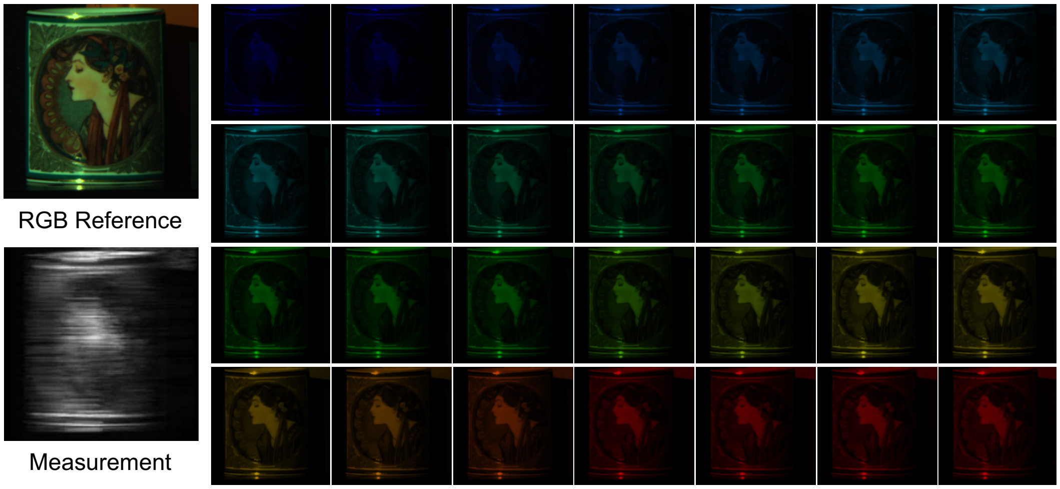


Figure 18: Reconstruction results by the proposed method (second to the right).The total 28 wavelengths are presented. The RGB reference is shown to demonstrate the color (top-left).

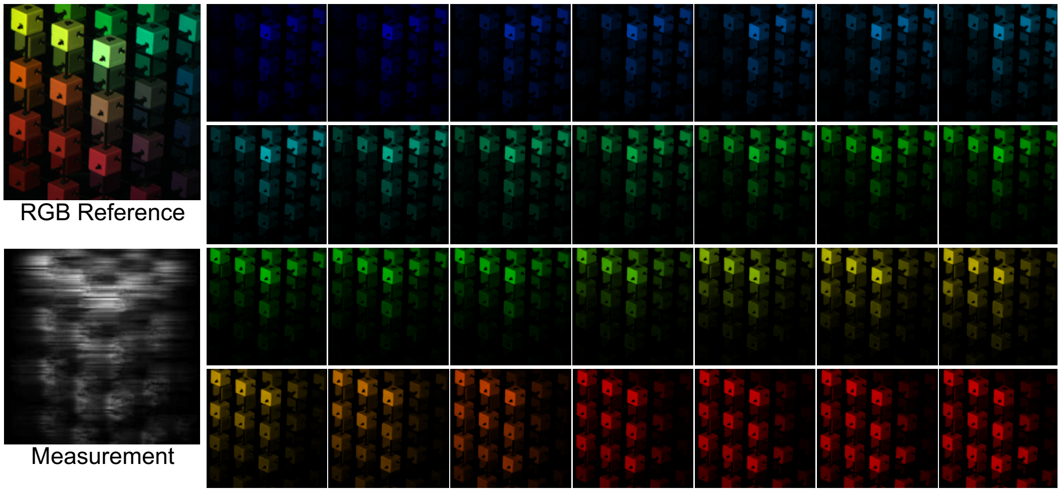


Figure 19: Reconstruction results by the proposed method (second to the right).The total 28 wavelengths are presented. The RGB reference is shown to demonstrate the color (top-left).

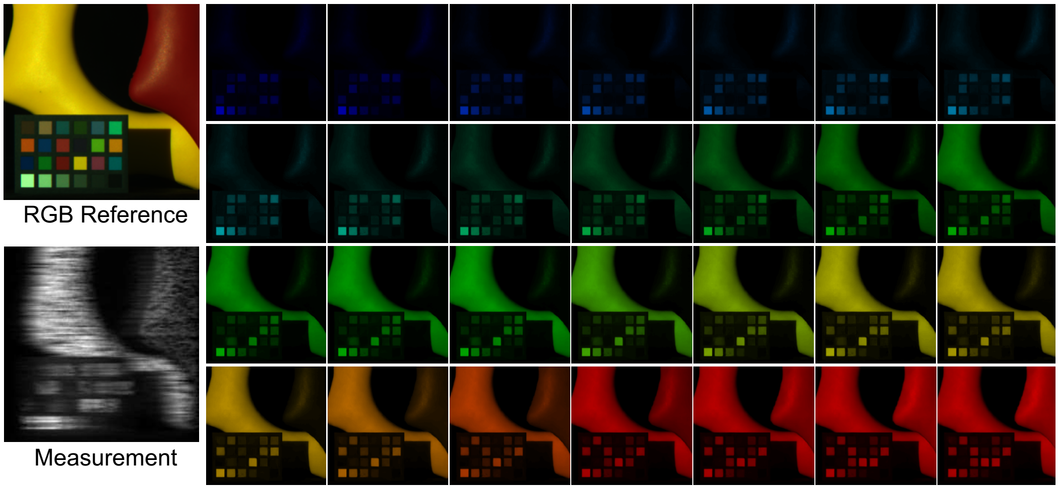


Figure 20: Reconstruction results by the proposed method (second to the right).The total 28 wavelengths are presented. The RGB reference is shown to demonstrate the color (top-left).

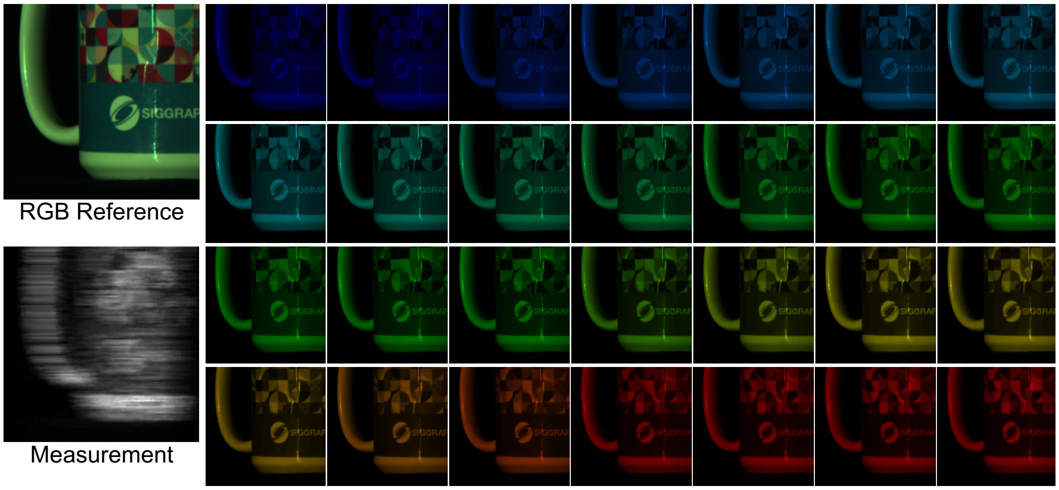


Figure 21: Reconstruction results by the proposed method (second to the right).The total 28 wavelengths are presented. The RGB reference is shown to demonstrate the color (top-left).

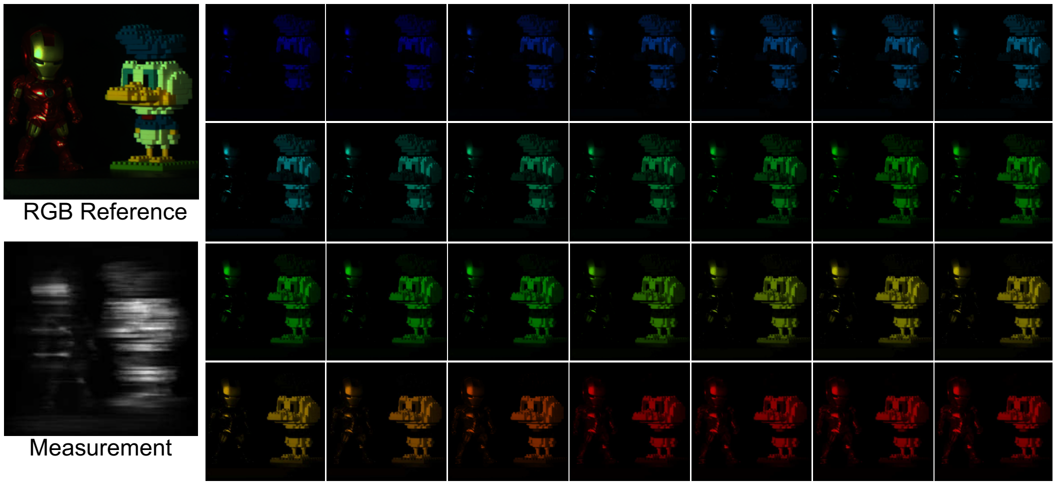


Figure 22: Reconstruction results by the proposed method (second to the right).The total 28 wavelengths are presented. The RGB reference is shown to demonstrate the color (top-left).

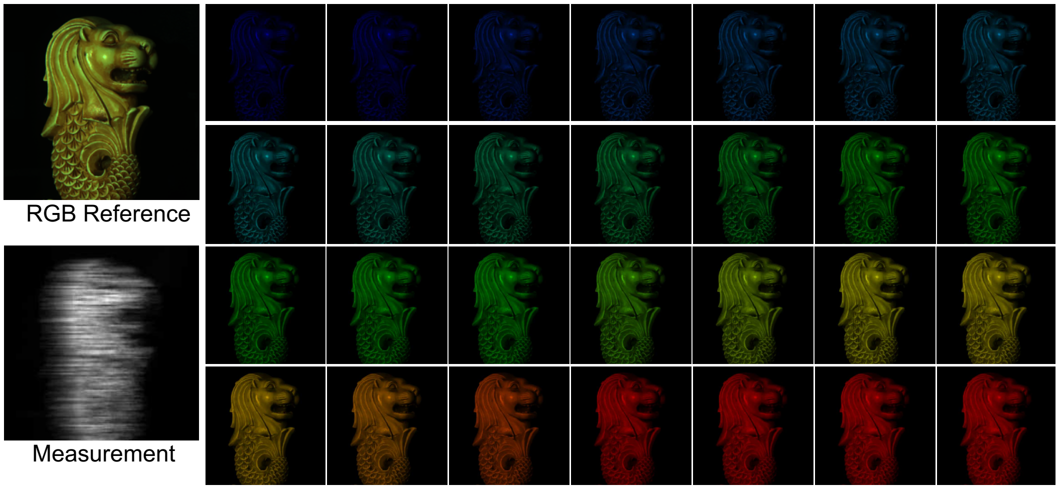


Figure 23: Reconstruction results by the proposed method (second to the right).The total 28 wavelengths are presented. The RGB reference is shown to demonstrate the color (top-left).