Supplementary Materials of A high-resolution video-rate hyperspectral camera with fusion architecture

Renwei Dian[1†], Yuanye Liu[1†], Lishan Tan[2], Anjing Guo[1],
Shutao Li[1,2*]

[1]the School of Robotics and the Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, Hunan University, Lushan South Road, Changsha, 410082, Hunan Province, China.
[2]the College of Electrical and Information Engineering and the Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, Hunan University, Lushan South Road, Changsha, 410082, Hunan Province, China.

*Corresponding author(s). E-mail(s): shutao_li@hnu.edu.cn;
Contributing authors: drw@hnu.edu.cn; yuanye_liu@hnu.edu.cn;
LishanTan@hnu.edu.cn; anjing_guo@hnu.edu.cn;
†These authors contributed equally to this work.

1 Simulation experiment setup

The ICVL dataset consists of 200 HS images, all of which are outdoor scenes. Each image has 1392×1300 pixels and 31 spectral bands, with a spectral range from 400 nm to 700 nm and a spectral resolution of 10 nm. We randomly select 160 images as the training data and the remaining 40 images as the test data. We anticipate that the proposed fusion architecture is also applicable to remote sensing satellites, so we test our algorithm on the Houston2013 dataset, which consists of a single HS image captured by the satellite. The HS image consists of 144 spectral bands covering wavelengths from 380 nm to 1050 nm, with a resolution of 349×1905 pixels. We select the left 349×400 area as the test set, and the rest as the training set.

We treat the original HS image as the high-resolution HS image, which is only used for quantitative evaluation and does not participate in training. For the ICVL dataset, we apply a 5×5 gaussian blur operation followed by a downsampling operation with a factor of 4 to obtain low-resolution HS images. MS images are then generated using the spectral response function of the Nikon D700 camera. For the Houston dataset, following the procedure in work [1, 2], we average every 36 bands to obtain the MS images. The low-resolution HS image is generated by applying a 7×7 gaussian blur operation followed by a downsampling operation with a factor of 4. Before generating the data, the original images are normalized to a range between 0 and 1.

2 Modeling and Solving the Imaging System

The relative observation model refer to the relationships among the corresponding frames of the low-resolution HS video, the PAN video, and the high-resolution HS video, which have been extensively explored in numerous studies [1, 3]. The high-resolution HS video frame to be estimated is denoted

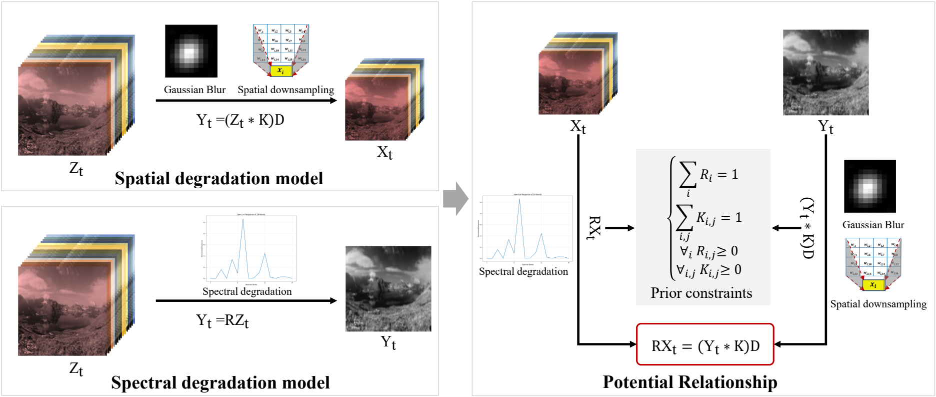

Fig. 1 The relationship among the observed HS image, the observed PAN image, and the ideal high-resolution HS image.

as Z_t, while the low-resolution HS video frame and PAN video frame are represented as X_t and Y_t, respectively. Figure 1 illustrates their relationships.

Without considering the influence of mosaic coding, the spatial degradation model from Z_t to X_t can be regarded as gaussian blur and downsampling according to previous studies [3, 4]. The process is expressed as:

$$X_t = (Z_t * K)D + \varepsilon_x \tag{1}$$

where $*$ denotes the convolution operation along the spatial dimensions, K represents the spatial blur kernel, D denotes the mean downsampling operation, which corresponds to pixel binning in our cameras, and ε_x represents the noise. The spectral degradation model from Z_t to Y_t can be written as:

$$Y_t = (RZ_t) + \varepsilon_y \tag{2}$$

where $R \in R^{L \times l}$ is the spectral response operator, and ε_y is the noise. Equation (1) and (2) not only reveal the relationship between the ideal high-resolution HS image and the two observed images, but also provide a theoretical foundation for HS fusion imaging.

According to Wald protocol [5], the relationship between X_t and Y_t can be further obtained as:

$$(RX_t) = (Y_t * K)D + \varepsilon \tag{3}$$

where ε is the noise, which is ignored in the subsequent solving process. To estimate both spatial and spectral degradation models, we developed a relative observation model learning network that models spatial degradation using learnable blur kernels and spatial downsampling operations, while representing the spectral degradation with a bias-free fully connected layer. In addition, we applied a series of physics-based sensor priors, including: $\sum_i R_i = 1$, $\sum_{i,j} K_{i,j} = 1$, $\forall_i \ R_i \geq 0$, and $\forall_{i,j} \ K_{i,j} \geq 0$.

The relative observation model learning network can implicitly capture the noise present in the real imaging process and accurately estimate the spatial blur kernel K and spectral response operator R in the observation model described by Eqs. (1) and (2). As demonstrated by simulation experiments, even in the presence of noise, the proposed reconstruction algorithm still achieves excellent results.

3 Efficient fusion network

Our cameras support video-rate imaging, which raises the efficiency requirements for the fusion algorithm. Thus, we develop an efficient fusion network to generate high-resolution HS video by fusing the HS video and the PAN video frame by frame. As shown in Fig. 2, the network primarily consists of two parts: first, a feature extraction module (FEM) based on shared convolutional layers is used to extract image features at multiple scales. Second, a reconstruction module (RM) with efficient spatial-spectral attention (ESSA) blocks is employed to reconstruct the corresponding video frame from these features.

In the FEM, the HS video frame is spatially upsampled by a factor of 4 using the PixelShuffle operation, while the PAN video frame is spatially downsampled by a factor of 2 using the PixelUnshuffle operation. Then the features are then stacked and fed into the lightweight convolution (LC) block to obtain low-resolution feature. Next, the low-resolution feature is spatially upsampled by a factor of 2 using the PixelShuffle operation to be stacked with the original PAN video frame. Finally,

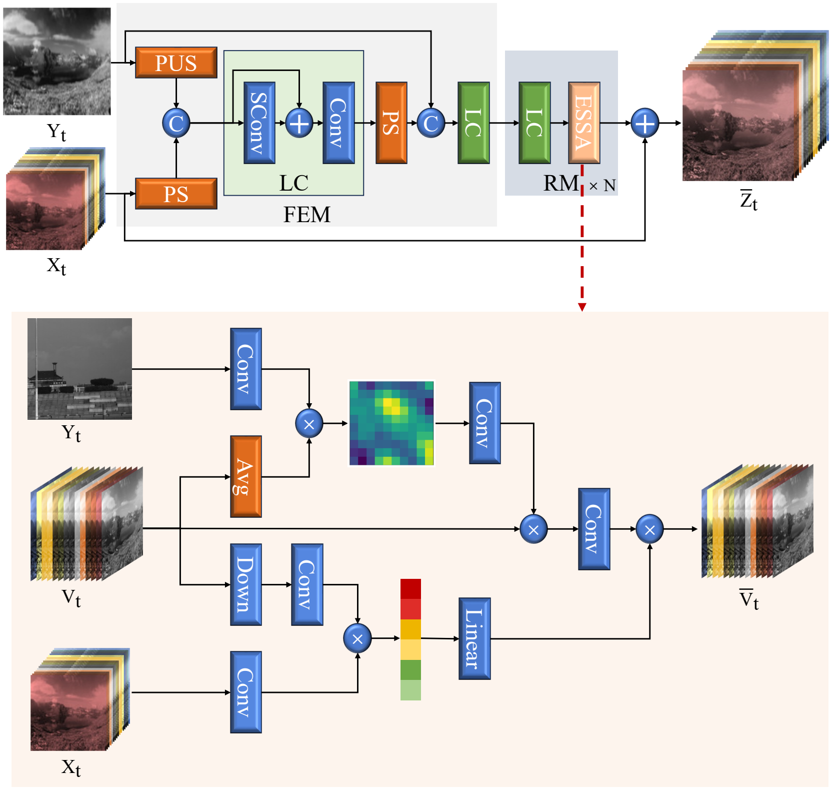

Fig. 2 Schematic of efficient fusion network.

the stacked feature is processed by an LC block to obtain the final feature. The LC block comprises a shared convolutional layer and a standard convolutional layer.

The RM consists of several LC blocks and ESSA blocks. The spatial and spectral attention mechanisms have shown advantages in many image reconstruction tasks, particularly global attention based on Transformers. However, Transformers typically require substantial computational resources, rendering them unsuitable for high-resolution image reconstruction. Furthermore, their inference speed is insufficient to meet the demands of real-time fusion imaging. To address this challenge, we propose the ESSA block, which models global features with lower computational cost. In particular, the ESSA block is mainly composed of two branches: spatial attention and spectral attention. Let the input feature be denoted as $V_t \in R^{c \times w \times h}$. For spatial attention, the PAN video frame Y_t is passed through a convolutional layer to expand the number of channels, yielding the feature $F_y \in R^{c \times w \times h}$. Meanwhile, the input feature V_t undergoes global average pooling to generate the feature $F_{v1} \in R^{c \times 1 \times 1}$. The feature F_y is then transposed and reshaped into $wh \times c$, and F_{v1} is reshaped into $c \times 1$. Subsequently, F_y is multiplied by F_{v1} and the resulting tensor is passed through a convolutional layer to generate the final spatial attention matrix $M_{spa} \in R^{w \times h}$. The above processes are formulated as follows:

$$M_{spa} = f_{spa}(V_t, X_t, Y_t) \tag{4}$$

where $f_{spa}(\cdot)$ is the function of spatial attention block. For spectral attention, the low-resolution HS video frame X_t undergoes a convolutional layer to to expand the number of channels, resulting in $F_x \in R^{c \times w_x \times h_x}$. The input feature V_t is processed through two convolutional layers for spatial downsampling and one convolutional layer for channel downsampling, yielding $F_{v2} \in R^{1 \times w_x \times h_x}$. The feature F_{v2} is then transposed and reshaped into $wh \times 1$, while F_x is reshaped into $c \times wh$. Finally, F_x and F_{v2} are multiplied, and the resulting tensor is passed through a linear layer to obtain the spectral attention vector $M_{spe} \in R^{c \times 1}$. The processes described above are formulated as follows:

$$M_{spe} = f_{spe}(V_t, X_t, Y_t) \tag{5}$$

where $f_{spe}(\cdot)$ is the function of spectral attention block. Finally, the input feature V_t is sequentially multiplied by the spatial attention matrix and the channel attention vector to obtain the refined feature $\bar{V}_t$.

$$\bar{V}_t = Conv(V_t \cdot M_{spa}) \cdot M_{spe} \tag{6}$$

4 Model Training Details

We first train the relative observation model learning network, and its objective function can be obtained from Eq.(3) as:

$$L_{bkl} = 1 - SSIM(X_t * R, (Y_t * K)D) \tag{7}$$

where $SSIM(\cdot)$ represents structural similarity index (SSIM) [6]. Figure 3 depicts the training and inference processes. In the real data experiments, we employ the trained blur kernel learning network to generate training data by spatially downsampling the observed HS and PAN images. Similarly, for the simulated data, spatial downsampling is applied to the MS and low-resolution HS images. The resampled data is then used as training data to train our efficient fusion network.

5 Evaluation metrics

Peak Signal-to-Noise Ratio (PSNR): PSNR is calculated based on the ratio between the maximum possible pixel value of the image and the error.

Error Relative Global Dimensional Synthesis (ERGAS): ERGAS measures the overall similarity between the reconstructed image and the reference image by quantifying the error of the reconstructed image relative to the reference image to evaluate the fidelity of the image.

Spectral Angle Mapper (SAM): SAM evaluates the similarity between the spectral vectors of corresponding pixels in the reference image and the reconstructed image by calculating the angle between them. The smaller the angle, the higher the similarity between the two spectral vectors, indicating better preservation of spectral information.

Structural Similarity Index (SSIM): SSIM compares the images in terms of luminance, contrast, and structure, aligning more closely with the human visual system's perception of image quality.

6 Application Experiments

To evaluate the application potential of the developed VIS-HS and NIR-HS, we conducted a series of experiments, including the recognition of real and artificial objects for material discrimination, monitoring of land desertification for environmental protection, identification of drug components for pharmaceutical safety, and dynamic target tracking. The experimental environment is divided into outdoor and indoor. The outdoor light source is natural light, while the indoor lighting is provided by a full-spectrum lamp in a dark box, as shown in Fig. 4. The spectrum of the full-spectrum lamp light covers 400nm-2500nm.

Segmentation for sand and soil To assess the potential of the developed HS cameras for remote sensing applications, we acquire HS images of sand and soil in an outdoor environment via VIS-HS. A central region of 1100 × 1100 pixels is selected for analysis. Within this region, two 100 × 100 pixel areas—one from the sand region and one from the soil region—are extracted as training data. A

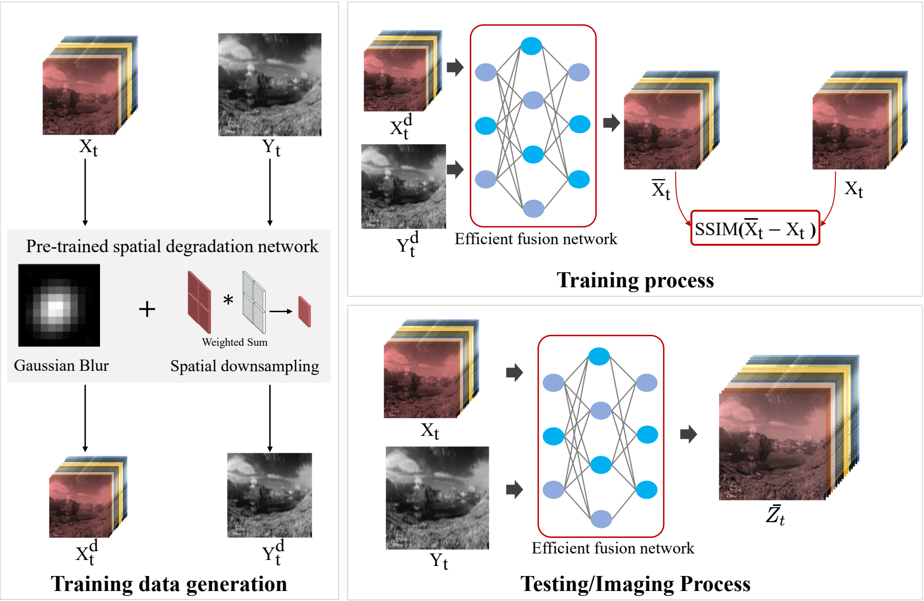

Fig. 3 Overview of the zero-shot training strategy.

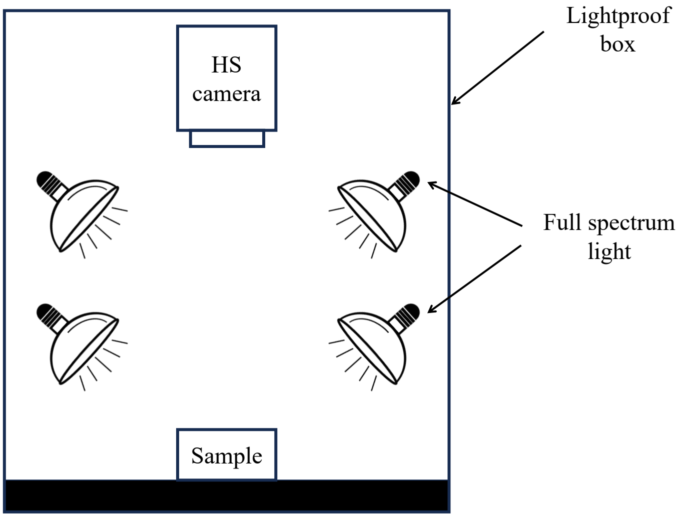

Fig. 4 Indoor illumination setup.

5 × 5 sliding window is used to crop the training data, generating training samples. These samples are used to train an SVM classifier, which is subsequently applied to perform segmentation across the entire 1100 × 1100 pixel region.

Classification for medicinal powder To assess the advantages of HS imaging over RGB imaging, we conduct a classification experiment using medicinal powders. Since ensuring that RGB images from different devices maintain the same spatial resolution as HS images is challenging, we create pseudo-RGB images by selecting the 536 nm, 550 nm, and 663 nm bands from the HS data. This approach eliminates the effect of resolution differences, ensuring that variations in classification results are solely caused by spectral information. We choose four medicinal powders with very similar visible colors: Dioscorea powder, Poria powder, lily powder, and starch. From each sample, we extract a 1200 × 1200 pixel region and apply non-overlapping 16 × 16 pixel windows to crop the data, generating the required samples. The SVM algorithm is then used, with half of the samples randomly selected as the training set and the remaining half as the test set.

References

[1] Dian, R., Guo, A., Li, S.: Zero-shot hyperspectral sharpening. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12650–12666 (2023) https://doi.org/10.1109/TPAMI.2023.3279050

[2] Liu, Y., Dian, R., Li, S.: Low-rank transformer for high-resolution hyperspectral computational imaging. International Journal of Computer Vision (2024)

[3] Guo, A., Dian, R., Li, S.: Unsupervised blur kernel learning for pansharpening. In: IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, pp. 633–636 (2020). https://doi.org/10.1109/IGARSS39084.2020.9324543

[4] Guo, A., Dian, R., Li, S.: A deep framework for hyperspectral image fusion between different satellites. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 7939–7954 (2023) https://doi.org/10.1109/TPAMI.2022.3229433

[5] Zeng, Y., Huang, W., Liu, M., Zhang, H., Zou, B.: Fusion of satellite images in urban area: Assessing the quality of resulting images. In: 2010 18th International Conference on Geoinformatics, pp. 1–4 (2010). https://doi.org/10.1109/GEOINFORMATICS.2010.5568105

[6] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004) https://doi.org/10.1109/TIP.2003.819861