ESSAformer: Efficient Transformer for Hyperspectral Image Super-resolution

Mingjin Zhang[1], Chi Zhang[1*], Qiming Zhang[2*], Jie Guo[1], Xinbo Gao[3], Jing Zhang[2]
[1] Xidian University, China
[2] The University of Sydney, Australia
[3] Chongqing University of Posts and Telecommunications, China
mjinzhang@xidian.edu.cn, ch_zhang@stu.xidian.edu.cn, qzha2506@uni.sydney.edu.au,
jguo@mail.xidian.edu.cn, gaoxb@cqupt.edu.cn, jing.zhang1@sydney.edu.au

Abstract

Single hyperspectral image super-resolution (single-HSI-SR) aims to restore a high-resolution hyperspectral image from a low-resolution observation. However, the prevailing CNN-based approaches have shown limitations in building long-range dependencies and capturing interaction information between spectral features. This results in inadequate utilization of spectral information and artifacts after upsampling. To address this issue, we propose ESSAformer, an ESSA attention-embedded Transformer network for single-HSI-SR with an iterative refining structure. Specifically, we first introduce a robust and spectral-friendly similarity metric, i.e.*, the spectral correlation coefficient of the spectrum (SCC), to replace the original attention matrix and incorporates inductive biases into the model to facilitate training. Built upon it, we further utilize the kernelizable attention technique with theoretical support to form a novel efficient SCC-kernel-based self-attention (ESSA) and reduce attention computation to linear complexity. ESSA enlarges the receptive field for features after upsampling without bringing much computation and allows the model to effectively utilize spatial-spectral information from different scales, resulting in the generation of more natural high-resolution images. Without the need for pretraining on large-scale datasets, our experiments demonstrate ESSA's effectiveness in both visual quality and quantitative results. The code will be released at ESSAformer.*

1. Introduction

Hyperspectral imaging (HSI) involves densely sampling spectral features with many narrow bands to encode rich spectral and spatial structure information for material differentiation. It has been widely used in various applications. Hyperspectral image super-resolution (HSI-SR) aims to generate high-resolution HSI from low-resolution HSI and can be categorized into two approaches: single-HSI-SR [25, 32, 19, 54] and fusion-based HSI-SR [30, 64, 34, 24]. This paper focuses on the challenging single-HSI-SR task, which aims to restore high-resolution HSI from a single low-resolution HSI without auxiliary images.

Conventional methods for single-HSI-SR involve designing a mapping function between low-resolution and high-resolution HSI using hand-crafted priors such as low-rank approximation and sparse coding [18, 40, 17, 13]. However, with the fast development of deep learning, powerful convolutional neural networks (CNNs) have led to significant progress in the single-HSI-SR task [25, 27, 32, 48, 23, 39, 19]. These CNN-based approaches usually use deep neural networks to formulate and learn the mapping function in an end-to-end manner using abundant training data pairs. As a result, they achieve significant improvements in both visual quality and quantitative metrics.

However, the CNNs methods show limitations in solving the single-HSI-SR task. There exists a significant amount of long-range information in high-dimensional data of HSI, while the most prevailing CNNs focus on local features captured by the convolutional kernels [25, 23, 39, 19, 27, 32, 48, 57, 55, 56]. The limited receptive field in the network can thus hinder the models' representation ability. Consequently, unwished artifacts, such as the blocking ones, may appear and affect the model's generation quality. To address this issue, we make an attempt to propose a Transformer model for the single-HSI-SR task. The attention mechanism introduced in Vision Transformers allows them to capture long-range dependencies and provide powerful representations, leading to superior performance compared to CNNs in many vision tasks [7, 66, 63, 59, 47, 9, 65, 60].

While the long-range dependency advantage of Vision Transformers can potentially address the aforementioned issues, it cannot be directly applied to single-HSI-SR. Firstly, Vision Transformers typically require a large amount of

*Corresponding author

data to learn inductive biases and produce reliable results. However, the difficulty in obtaining HSIs limits the collection of large-scale datasets, which poses a particular challenge compared to the millions of images available in RGB image datasets, thus hindering the training of Vision Transformers. Secondly, while Transformers can handle long-range dependencies, the self-attention process has a quadratic computation complexity of $\mathcal{O}(N^2)$ with respect to the token sequence N. This results in a massive computation burden for the network, particularly for ultra-high resolution HSI.

To address the above issues, we propose a novel Transformer model called ESSAformer. The ESSAformer is designed with several adaptations. First, it utilizes an iterative downsampling and upsampling strategy to capture global and local information at different scales and encode the detailed content of the hyperspectral images. Second, we propose to replace the conventional dot product (cosine similarity) with the robust and spectral-friendly spectral correlation coefficient of the spectrum, called SCC. Compared to traditional cosine similarity, the SCC has desirable properties such as spectral-wise shifting and scaling equivalence. This makes the model insensitive to amplitude-level changes in spectral curves caused by occlusions or shadows. As a result, SCC brings inductive biases into models, facilitates training efficiency, and even enables from-scratch training of Transformer models on small datasets. Third, we propose formulating the attention as kernelized ones to decrease the computation burden. Technically, we integrate SCC into a nonlinear square exponential kernel, *i.e*., Mercer's kernel, and then express SCC as a dot product of two individual terms according to the Mercer theorem. Subsequently, we change the multiplication order of self-attention, *i.e*., multiplying keys and values first and then queries, and thus lower the attention complexity from quadratic $\mathcal{O}(N^2)$ to linear $\mathcal{O}(N)$. Such a pipeline significantly relieves the computation burden since the token number N for high-resolution HSIs is usually significantly long. Consequently, we propose the novel SCC-kernel-based self-attention, called ESSA, and a new ESSAformer Vision Transformer architecture for the single-HSI-SR task.

Thanks to the proposed ESSA, our model efficiently enlarges the receptive field without imposing a significant computation burden, thus allowing the features to attend to the entire feature map at each layer and gather sufficient information. Consequently, ESSA effectively addresses the artifact issues caused by limited and inconsistent receptive fields between any two pixels in hyperspectral images, resulting in more natural high-resolution HSI generation. Unlike other attention variants [58, 45], our ESSA does not bring extra parameters and effectively introduces inductive biases. Consequently, the ESSAformer Transformer model obtains state-of-the-art performance on three public datasets without the need for pretraining.

In summary, this paper makes three main contributions. First, we introduce the use of Vision Transformer for the single-HSI-SR task and propose the ESSAformer model with strong learning ability. Second, we present a novel and efficient SCC-kernel-based self-attention method called ESSA. The approach significantly reduces the computation and data-hungry issues in the original Vision Transformer and helps the model better fit the single-HSI-SR task. Third, extensive experiments have been conducted to analyze the proposed model thoroughly, and the state-of-the-art performance on three popular datasets demonstrates its superiority regarding both visual quality and objective metrics.

2. Related Work

2.1. HSI super-resolution

For HSI-SR, the existing prevailing approaches are mostly based on CNNs. Since the researchers first apply CNN [25] into this task by proposing a deep spectral difference network, the architecture design for HSI-SR has attracted much attention from the community. For example, a 3D full convolutional network (3D-FCNN) [32] is proposed to recover high-quality HSI without any auxiliary information. Besides, a mixed 2D/3D convolutional network (MCNET) [23] and a bidirectional 3D convolutional network (Bi-3DQRNN) are proposed to take into account the forward and backward spectral dependence of HSIs [14]. Besides the 3D CNNs, the recursive strategy also demonstrates its effectiveness [27], which utilizes grouped recursive modules in the global residual structure to depict the complicated non-linear mapping function. Such group strategy has inspired various works to avoid over-processing the redundancy in HSIs and save computational costs. For example, SSPSR [19] employs grouped convolutional layers and channel attention to exploit spectral correlation. Besides, Zhang *et al*. [54] develop a difference curvature network (DCM-NET) in light of the grouping strategy. Nevertheless, prevailing methods are not beyond the limitation of CNN. They are not good at capturing long-range dependencies and modeling spatial-spectral correlation, resulting in insufficient utilization of spectral information and thus unwished artifacts after super-resolution. To this end, we propose a novel Transformer model ESSAformer to solve super-resolution in HSIs.

2.2. Efficient vision transformer

Transformer is originally proposed in the natural language process field [37, 10, 4, 44]. After Dosovitskiy *et al*. [12] fed images into a pure Transformer and gained great success on various image recognition tasks, Vision Transformer has received much attention from researchers in various computer vision tasks [7, 66, 63, 9, 65]. Since

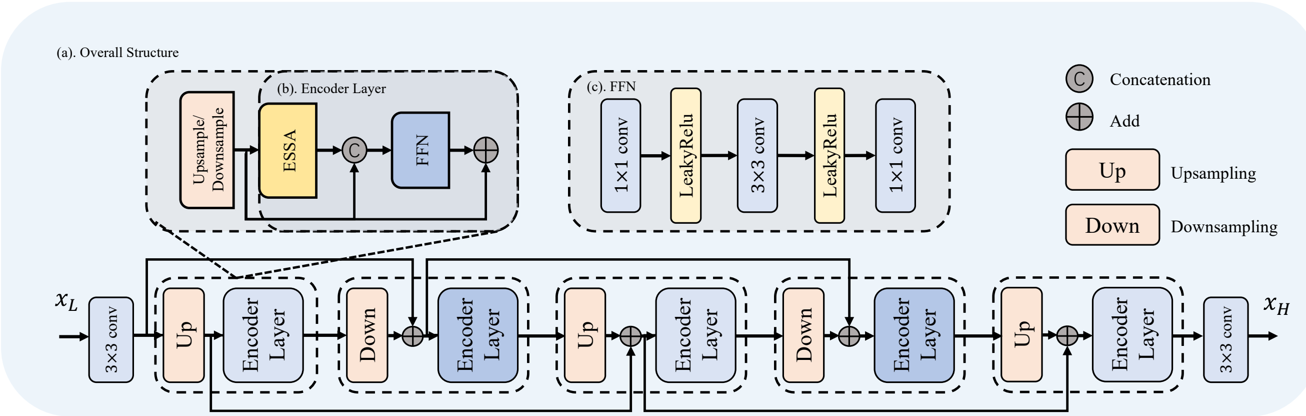


Figure 1. Overview of the proposed ESSAformer. The model is constructed by stacking upsampling/downsampling and encoder layers. The pipeline follows an iterative refinement process to handle the feature representations at different scales, thereby better encoding details and contextual information. All encoder layers having the same input resolution, *i.e*., the same color, share weight for a lightweight design.

the quadratic computation complexity of attention hinders the Transformer application, especially for high-resolution images, several works are proposed to relieve such issues. For example, Geng *et al*. [16] leverages matrix decomposition to substitute the original self-attention while modeling the dependence between different tokens. Similarly, self-attention is approximated as a linear dot-product of kernel feature maps [20] to avoid massive computation in attention. In contrast, kernelized attention [51, 31] aims to find kernels to approach the attention matrix and relieve the computation by changing the multiplication order. Our ESSA falls in this track to improve computation efficiency. However, previous works target the original attention with RGB images, while ESSA proposes an efficient attention method particularly designed for HSIs. Specifically, ESSA fully considers the characteristics of the hyperspectral field and brings channel-wise inductive bias into the models for better restoration performance.

Efficient Vision Transformer has also demonstrated its effectiveness in image restoration tasks. For example, SwinIR [29, 28] uses window attention to calculate attention within local windows instead of the whole feature to reduce the computation cost. Stoformer [46] studies the window partition mechanism and proposes a stochastic shifting method. Wang *et al*. [43] designs a locally-enhanced window-based attention mechanism and a U-Shape model architecture for the restoration task, while CAT [62] extends the window shape to rectangles. Different from them, Lee *et al*. [22] established local attention with non-local connectivity using local-sensitive hashing. Besides the spatial-wise attention, channel-wise attention also proves to work well for high-resolution image restoration tasks [53, 6]. They have linear complexity and are most related to our methods. However, ESSA still conducts spatial-wise attention. With strict theoretical support, it uses kernelizable techniques to enable the multiplication exchange, which is significantly different from channel-wise attention in motivation, mathematical formulation, and performance.

3. Method

In this section, we will first introduce the overall structure of our proposed ESSAformer model. Then ESSA's implementation details, theoretical support, and complexity analysis are presented in the following part.

3.1. Overall structure

The overall structure is presented in Fig. 1. Given the predefined scaling ratio s and low-resolution HSIs $\mathbf{x}_L \in R^{h \times w \times c}$, ESSAformer outputs the high-resolution HSIs $\mathbf{x}_H \in R^{sh \times sw \times c}$ through the learned mapping function $M(\cdot)$ of low to high resolution, where h, w, c denote the height, width, and channel of HSIs respectively and s usually set to 2, 4, 8, etc.

$$\mathbf{x}_H = M(\mathbf{x}_L) \tag{1}$$

The channel dimension c is usually larger than that in RGB images for HSIs. Each channel describes the real world at discrete bands from a wide range of continuous spectrums.

The model first projects the raw HSIs into features with a projection layer and then contains several stages to sequentially process the features. At the beginning of each stage is a rescaling module to upsample/downsample the feature maps. Then the encoder layer that has an ESSA and an FFN module follows to encode the features, as shown at the top of Fig. 1. For each upsampling/downsampling module, multiple projections and PixelShuffle/PixelUnshuffle [36] layers with a rescale ratio of 2 are used sequentially until they output the feature map to the expected resolution. For example, we use 3 layers in the rescaling module for the required scaling ratio $s = 8$. The features channel dimension keeps after both upsampling and downsampling modules.

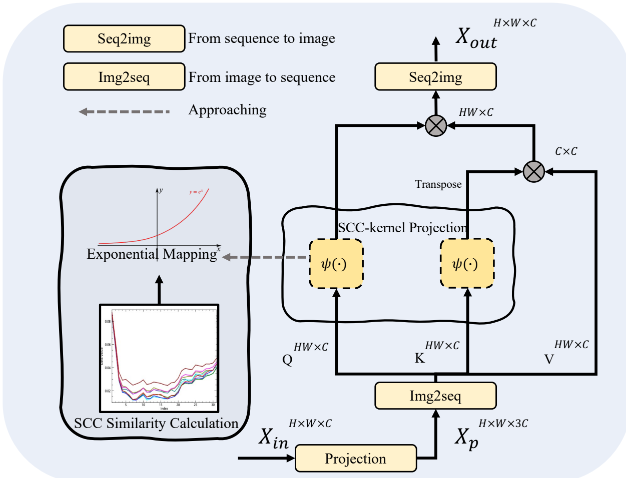


Figure 2. The structure of our ESSA.

In encoder layers, the feature map after ESSA is concatenated with ESSA's input and then fed into the feed-forward layer (FFN), which is consisted of several convolution and activation layers following the common practice [26].

These sequential stages follow an up-down strategy and thus construct an iterative refinement process by encoding feature representations at different scales, allowing the model to effectively capture content details and contextual information. Each stage includes one encoder layer, and all stages that have the same input resolution (the same color in Fig. 1) share weights to enable a lightweight model design. Besides, ESSAformer uses a residual connection between features right after the up/downsampling layers of the adjacent three stages as illustrated in Fig. 1. This approach enables the model to utilize multi-scale features and generate better feature representations. Finally, after the last stage with an upsampling module, a convolution layer with a 3×3 kernel projects the feature maps into the required channel dimension c, resulting in high-resolution HSIs $\mathbf{x}_H$.

3.2. SCC self-attention

As one of the cores of Transformers, self-attention enlarges the dependence distance by attending to the feature at each position. We will start from the original attention process and then introduce SCC self-attention that introduces channel-wise inductive biases to improve the data efficiency. We take the input resolution $H \times W$ for example.

Given the input features in $R^{H\times W\times C}$, a projection layer first embeds them into three token sequences, *i.e.*, $Q, K, V \in R^{N\times C}$, where N equals to $H \times W$, *i.e.*, the sequence length, and C is the channel dimension. The attention function computes the dot products of queries with all keys and applies a softmax function as the weights on the values. We take the single-head self-attention for simplicity and the function is thus given by:

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{C}})V \tag{2}$$

Then we will introduce the proposed SCC self-attention. It considers the specific characteristics in hyperspectral images and brings inductive biases to improve data efficiency and representation ability.

To make attention spectral-friendly, we utilize a robust spectral similarity measure named Spectral Correlation Coefficient of Spectrum, *i.e.*, SCC. SCC is the utilization of Pearson's correlation coefficient r in the hyperspectral field and represents the cosine of the generalized angle of tokens after the spectral curves minus their averages, which can be obtained by:

$$r(q,k) = \frac{(q-\overline{q})(k-\overline{k})^T}{||(q-\overline{q})|| \cdot ||(k-\overline{k})||}, \tag{3}$$

where $\overline{q}, \overline{k}$ denote the mean value of the any two token vectors q, k in Q, K. $r \in [-1, 1]$ represents whether the correlation degree of q and k are positive or negative. Following the practice in [11], we propose to use r^2 to ensure the non-negativity value and regard it as the attention matrix to represent the relationship between any two tokens. Then we have the following theorem for the inductive biases.

Theorem 3.1 *Let q, k be vectors $\in R^{1\times C}$ and r^2 be the correlation degree that describes the relationship between any two token vectors, we have the channel-wise translational inductive biases,* i.e.*, scales and shifts, in r^2:*

$$\begin{aligned} r^2(q, s\cdot k+t) &= (\frac{(q-\overline{q})(s\cdot k+t-(s\cdot\overline{k}+t))^T}{||q-\overline{q}|| \cdot ||s\cdot k+t-(s\cdot\overline{k}+t)||})^2 \\ &= r^2(q,k) \end{aligned} \tag{4}$$

for any $s \in R, t \in R^{1\times C}$.

Such characteristics indicate that SCC self-attention is affected by neither shadows nor occlusions that usually lead to scales and offsets transformation in HSIs, where

$$\mathrm{SCC}(Q,K,V) = r^2(Q,K)V. \tag{5}$$

Such inductive biases improve the model convergence and make it easy to train on small datasets from scratch. Although the operations on local features, such as pixel unshuffle, limit the receptive field and tend to produce blocking artifacts around the 'block' boundaries, as demonstrated in [42], the attention mechanism can effectively enlarge the receptive field by building the long-range dependencies among feature maps, which thus helps to produce natural and smooth images.

3.3. Efficient SCC-kernel-based self-attention

Based on SCC self-attention, we introduce the proposed efficient SCC-kernel-based self-attention (ESSA) to relieve the computation burden in attention, and the calculation process is shown in Fig. 2. In a kernel machine, in order to get the results of $\psi(Q) \cdot \psi(K)$, it is common to find the kernel function $\mathcal{K}(Q, K) = \psi(Q) \cdot \psi(K)$ so that we can directly calculate $\mathcal{K}(Q, K)$ instead of calculating $\psi(Q)$ and $\psi(K)$ separately, which is known as the kernel trick. In contrast, to lower the computation cost of SCC self-attention, a solution is finding the mapping function $\psi(Q)$ and $\psi(K)$ so that $\psi(Q)\psi(K) = r^2(Q, K)$. Then Equation 5 can be reformulated as $\psi(Q)(\psi(K)^T V)$ and decreases from quadratic complexity to linear complexity regarding the token sequence N. Before finding the mapping function, we first develop SCC into a radial basis function (RBF)-like kernel [35] for non-linearity and good derivatives:

$$\mathcal{K}_{SCC} = \exp(r^2). \tag{6}$$

In the following, we demonstrate that the SCC-kernel $\mathcal{K}_{SCC}$ is a Mercer's kernel to confirm the existence of the mapping function $\psi()$.

Theorem 3.2 *(Mercer's theorem [33]) Let $\mathcal{X}, \mathcal{Y}$ be the input space, and $\mathcal{H}$ be the Hilbert space. If the mapping $\psi(x) : \mathcal{X} \longrightarrow \mathcal{H}$ exists, then there is a kernel function $\mathcal{K}(x, y)$ satisfied:*

$$\mathcal{K}(x, y) = \langle \psi(x), \psi(y) \rangle \tag{7}$$

Theorem 3.3 *(Mercer's kernel closure properties [5]) Let $\mathcal{X}, \mathcal{Y}$ be the input space and $\mathcal{K}_1$,$\mathcal{K}_2$ be the Mercer's kernels, then:*

- *if $\mathcal{K}(x, y) = \mathcal{K}_1(x, y)\mathcal{K}_2(x, y)$, then $\mathcal{K}$ is Mercer's kernel.*
- *if $a, b > 0$ and $\mathcal{K}(x, y) = a\mathcal{K}_1(x, y) + b\mathcal{K}_2(x, y)$, then $\mathcal{K}$ is Mercer's kernel.*

According to Theorem 3.2, we can conclude r is a normalized linear kernel and also a Mercer's kernel. Mathematically, the Taylor expansion of $\exp(r^2)$ can be expressed as $\exp(r^2) = 1 + r^2 + \frac{(r^2)^2}{2!} + \frac{(r^2)^3}{3!} + \cdots$. According to Theorem 3.3, we can easily find r^2 and $\exp(r^2)$ Mercer's kernels. It guarantees the existence of the mapping function $\psi()$, which can be acquired via Taylor expansion as follows:

$$\begin{aligned}
\mathcal{K}_{SCC}(q, k) &= \exp\Big(\frac{q_{norm}^2 k_{norm}^2}{\sigma}\Big) \\
&= \sum_{i=0}^{\infty} \frac{(q_{norm}^2 k_{norm}^2)^i}{\sigma^i i!} \\
&= \sum_{i=0}^{\infty} \Big(\frac{q_{norm}^{2i}}{\sigma^{\frac{1}{2}i}\sqrt{i!}} \frac{k_{norm}^{2i}}{\sigma^{\frac{1}{2}i}\sqrt{i!}}\Big) \\
&= \langle \psi(q), \psi(k) \rangle
\end{aligned} \tag{8}$$

where $q_{norm} = q - \overline{q}$, $k_{norm} = k - \overline{k}$ and $\psi(q) = (1, \frac{q_{norm}^2}{\sigma^{\frac{1}{2}}}, \frac{q_{norm}^4}{2^{\frac{1}{2}}\sigma}, \cdots,)$. We choose the order of the polynomial, *i.e.*, the number of terms, to balance the performance and computation cost during experiments. Finally, the calculation of ESSA is given by

$$ESSA(Q, K, V) = (\psi(Q)\psi(K)^T)V = \psi(Q)(\psi(K)^T V) \tag{9}$$

By exchanging the multiplication order, the total computation complexity of ESSA is $\mathcal{O}(NC^2)$, which is significantly smaller than conventional attention $\mathcal{O}(N^2C)$ because the channel dimension is usually much smaller than the sequence length, especially for high-resolution images in HSI-SR. It is noted that the mathematical formulation of ESSA has a significant difference from channel-wise attention [53], *i.e.*, $Q \times \mathrm{softmax}(K^T V)$, demonstrating the different motivations and behaviors between the two methods.

4. Experiments

4.1. Datasets and settings

Chikusei dataset: The Chikusei dataset [50] is acquired using the Headwall Hyperspec-VNIR-C imaging sensor over agricultural and urban areas in Chikusei, Ibaraki, Japan. The dataset comprises images of 19 different classes, which are collected via a field survey and visual inspection, along with high-resolution images that are captured concurrently with the hyperspectral data.
Cave dataset: The Cave dataset [49] is a multispectral dataset that comprises 32 images of everyday objects. The dataset contains full spectral resolution reflectance data from 400 nm to 700 nm at a resolution of 10 nm, resulting in a total of 31 bands. The images have a resolution of 512 × 512 pixels and are stored as 16-bit grayscale PNG images per band.
Pavia dataset: The Pavia Dataset [15] is a hyperspectral dataset acquired by the ROSIS sensor during a flight campaign over Pavia, northern Italy. The images contain 102 spectral bands with a spatial resolution of 1096 × 1096, with a geometric resolution of 1.3 m. The image feature area is divided into 9 categories, each containing 9 samples.

Method		MPSNR↑	SAM↓	ERGAS↓	MSSIM↑	RMSE↓	CC↑	MACs(G)
Bicubic	2×	43.2125	1.7880	3.5981	0.9721	0.0082	0.9781	N/A
GDRRN [25]	2×	46.5412	1.3779	2.5896	0.9872	0.0055	0.9884	1.66
SSPSR [19]	2×	47.4403	1.2072	2.2805	0.9897	0.0050	0.9910	23.47
MCNet [23]	2×	46.7882	1.3311	2.4382	0.9872	0.0055	0.9893	82.77
Bi-3DQRNN [14]	2×	45.7107	1.4306	2.7407	0.9843	0.0061	0.9867	30.24
DCM-NET [54]	2×	48.0238	1.1160	2.1766	0.9906	0.0047	0.9916	101.3
Ours	2×	**48.2886**	**1.1004**	**2.1268**	**0.9912**	**0.0045**	**0.9920**	12.82
Bicubic	4×	37.6377	3.4040	6.7564	0.8954	0.0156	0.9212	N/A
GDRRN [25]	4×	39.6456	2.6306	5.3946	0.9353	0.0122	0.9490	6.65
SSPSR [19]	4×	40.3612	2.3527	4.9894	0.9413	0.0114	0.9565	42.44
MCNet [23]	4×	39.5599	2.7831	5.3687	0.9317	0.0126	0.9481	289.63
Bi-3DQRNN [14]	4×	39.8938	2.5221	5.1923	0.9377	0.0120	0.9518	120.97
DCM-NET [54]	4×	40.5139	2.3012	4.8584	0.9464	0.0112	0.9581	130.9
Ours	4×	**40.7648**	**2.2126**	**4.7231**	**0.9487**	**0.0109**	**0.9601**	48.65
Bicubic	8×	34.5049	5.0436	9.6975	0.8069	0.0224	0.8314	N/A
GDRRN [25]	8×	35.2210	4.6363	9.0720	0.8354	0.0202	0.7977	26.62
SSPSR [19]	8×	35.8279	4.0282	8.3177	0.8538	0.0192	0.8773	118.33
MCNet [23]	8×	35.2643	4.6107	8.7438	0.8321	0.0208	0.8588	2637.51
Bi-3DQRNN [14]	8×	35.6284	4.2259	8.4955	0.8456	0.0196	0.8701	483.89
DCM-NET [54]	8×	35.9809	3.9310	8.1459	0.8580	0.0189	0.8811	249.95
Ours	8×	**36.1405**	**3.8979**	**8.1181**	**0.8599**	**0.0187**	**0.8823**	192.64

Table 1. Quantitative comparison of different methods on the Chikusei dataset.

Harvard dataset: The Harvard dataset [8] collects fifty real-world images from indoor and outdoor scenes. Images are taken from a hyperspectral camera (Nuance FX, CRI Inc.) with wavelengths ranging from 420nm to 730nm at steps of 10nm. Each image has a spatial resolution of 1392 × 1040 with thirty-one spectral measurements at each pixel.

Implementation details: We trained our ESSAformer model from scratch using PyTorch and the Adam optimizer. The loss function is L1 loss. We set the initial learning rate to 1×10^{-4} and gradually decreased it to a minimum of 1×10^{-5}. We used the same training settings for all four datasets (Chikusei, Pavia, Cave, Harvard), without any special tuning. Our ESSAformer model consists of five stages, each with one encoder layer, and an upsampling module in the last stage. The first 3×3 convolution layer projects the channel dimension to $C = 256$, which is maintained in all stages except the last 3×3 convolution layer that recovers the original channel size. The temperature σ is set to 1. We used NVIDIA RTX 3090 GPUs for all experiments.

Evaluation metrics: We use six popular metrics for all experiments to thoroughly evaluate the model's performance in both spatial and spectral perspectives: the peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM) [52], erreur relative globale adimensionnelle de synthese (ERGAS) [38], structure similarity (SSIM) [41], root mean square error (RMSE), and cross correlation (CC) [30].

4.2. Qualitative results

We compare our methods with five representative deep learning methods includes GDRRN [25], SSPSR [19], MCNet [23], Bi-3DQRNN [14], and DCM-NET [54], besides traditional bicubic interpolation. All the models are trained from scratch. The qualitative and visual results will be presented below by datasets:

Experiments on the Chikusei dataset: Images in Chikusei have 2517×2335 pixels with 128 bands. We follow SSPSR [19] and crop non-overlapped patches of 512×512 resolution. For each image in Chikusei, four cropped patches are used for testing and the rest are for training. We have three scale factors for experiments. Specifically, the input resolutions for scale factors of 2, 4, and 8 are set to 32×32, 16×16, and 16×16, respectively. The output resolutions are 64×64, 64×64, and 128×128, respectively.

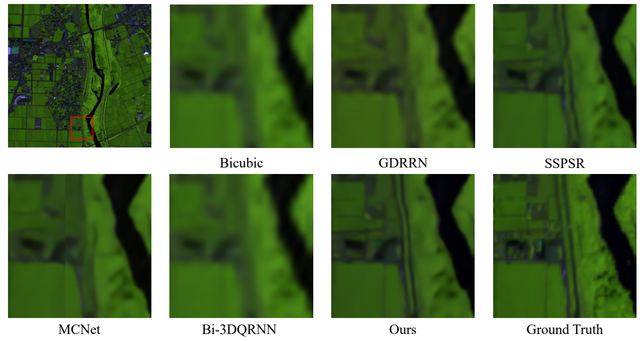


Figure 3. Chikusei's visual results, *i.e.*, the patch in red rectangle, of different methods are provided for comparison. We set the bands of 70/100/36 as the R/G/B channels for better visualization.

The quantitative results for the different methods are demonstrated in Tab. 1 with the best performance highlighted in bold. Our ESSA outperforms other approaches in almost all metrics at all three scale factors, demonstrating the effectiveness of our model. For example, ESSAformer outperforms the second, *i.e.*, DCM-Net [54], by 0.26 dB in PSNR and 0.13 in ERGAS for the 4× scale factor. To further measure the performance of our model, the visual results are presented in Fig. 3. We zoomed in the area in the red rectangle and provide the details produced by each method. It can be seen that only SSPSR, DCM-NET, and our method recovered the two vertical lines of the original image clearly, and among these three methods, SSPSR yields the worst result whose reconstructed lines are broken in many places. DCM-NET recovers relatively better than SSPSR, but discontinuation also appears, and our ESSA produces the best results.

Experiments on the Cave dataset: The Cave dataset [49] has 32 images with 512×512 resolution of different scenes with full spectral resolution reflectance data from 400 to 700 nm at a resolution of 10 nm (31 bands in total). We randomly chose 8 scenes for testing and the remained images are for training. The same cropping settings of Chikusei are used for experiments with Cave.

The results are represented in Tab. 2. As can be seen, our network still obtains the best performance in most metrics and scale factors, which confirms ESSAformer's superiority. Besides, we compare the spectral profile of the images from different methods in Fig. 4. As shown in the right image of Fig. 4, our method, *i.e.*, the green line, is the clos-

Method		Pavia						Cave					
		MPSNR↑	SAM↓	ERGAS↓	MSSIM↑	RMSE↓	CC↑	MPSNR↑	SAM↓	ERGAS↓	MSSIM↑	RMSE↓	CC↑
Bicubic	2×	34.4107	5.2881	4.4877	0.9387	0.0197	0.9760	38.0603	3.237	4.9579	0.9662	0.0147	0.9907
GDRRN [25]		37.4868	4.7773	3.2812	0.9651	0.0138	0.9867	40.9785	3.7454	4.2106	0.9738	0.0126	0.9948
SSPSR [19]		37.7264	4.6532	3.1757	0.9668	0.0135	0.9875	41.3895	3.1472	3.3333	0.9752	0.0101	0.9953
MCNet [23]		37.2858	4.7874	3.3173	0.9638	0.0143	0.9864	41.9772	2.6656	3.1483	0.9765	0.0095	0.9956
Bi-3DQRNN [14]		36.9049	4.7343	3.4460	0.9623	0.0148	0.9855	40.7212	2.8859	3.6087	0.9744	0.0108	0.9945
DCM-NET [54]		37.1815	5.2427	3.3738	0.9600	0.0144	0.9861	41.9867	2.7051	3.1217	0.9771	0.0095	0.9957
Ours		**38.7896**	**4.5576**	**2.8806**	**0.9712**	**0.0120**	**0.9896**	**42.2174**	**2.6623**	**3.0443**	**0.9778**	**0.0092**	**0.9958**
Bicubic	4×	29.6732	6.9353	7.5858	0.8154	0.0347	0.9321	33.0421	4.7962	7.846	0.9202	0.0258	0.9767
GDRRN [25]		30.8474	6.5915	6.7641	0.8677	0.0302	0.9477	34.897	4.3822	6.7552	0.938	0.0206	0.9830
SSPSR [19]		30.6447	6.4081	6.776	0.8619	0.0312	0.9461	35.3433	4.1654	6.5045	0.9434	0.0200	0.9838
MCNet [23]		30.9330	6.6822	6.5725	0.8628	0.0299	0.9488	35.5813	**3.7189**	6.3928	**0.9470**	0.0194	0.9845
Bi-3DQRNN [14]		30.4242	7.2323	7.0308	0.8525	0.0317	0.9417	35.3269	3.9294	6.5136	0.9438	0.0199	0.9839
DCM-NET [54]		30.6048	7.2623	6.8190	0.8507	0.0312	.9454	35.5055	3.9460	6.4092	0.9445	0.0197	0.9957
Ours		**31.6126**	**6.3032**	**6.1063**	**0.8847**	**0.0275**	**0.9558**	**35.8947**	3.8390	**6.1621**	0.9467	**0.0190**	**0.9869**
Bicubic	8×	26.6043	8.4814	10.7741	0.6509	0.0500	0.8597	29.2466	6.6079	12.2687	0.832	0.039	0.9439
GDRRN [25]		26.7832	9.2154	10.5724	0.6718	0.0488	0.8698	30.3026	7.151	11.0352	0.8513	0.0353	0.9533
SSPSR [19]		26.8435	10.1753	10.4754	0.6692	0.0487	0.9595	31.129	5.5101	10.1804	0.8749	0.0325	0.9595
MCNet [23]		27.0388	8.7111	10.2412	0.6808	0.0475	0.8734	31.323	5.7398	10.0206	0.8804	0.0320	0.9607
Bi-3DQRNN [14]		26.9813	**8.4730**	10.312	0.6802	0.0479	0.8715	31.1791	5.3401	10.1370	0.8792	0.0322	0.9601
DCM-NET [54]		25.6571	13.4698	12.014	0.5487	0.0559	0.8331	31.3766	5.3067	9.9363	0.8822	0.0316	0.9618
Ours		**27.1114**	8.9197	**10.1681**	**0.6918**	**0.0471**	**0.8757**	**31.4387**	**5.3041**	**9.9058**	**0.8845**	**0.0316**	**0.9629**

Table 2. Quantitative comparison of different methods on the Cave and Pavia datasets. The results of different scales are given respectively, where the best performance of each scale is highlighted in bold.

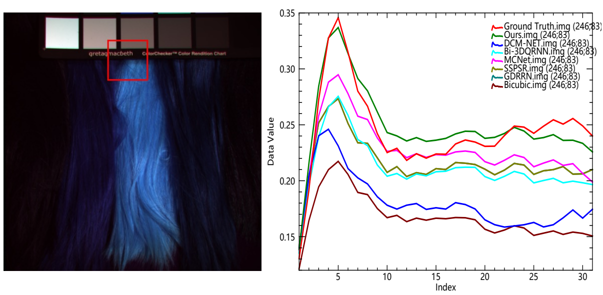


Figure 4. The test image in Cave and the spectral profile of the area in the red rectangle are provided respectively. The red line refers to the ground truth and the closest green line refers to the ESSAformer's result, which shows the strong ability of the ESSAformer to recover the texture detail.

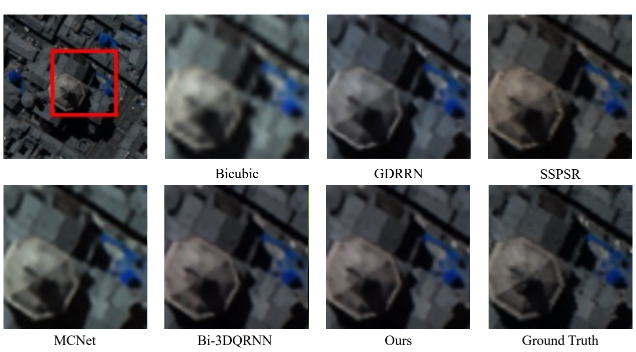


Figure 5. The test image from the Pavia dataset and details of the area in red rectangle offered by various methods. Bands 10/70/50 are visualized as the R/G/B channels.

est to ground truth (the red line) and recovers better details, while the other methods cannot well restore the peak value.

Experiments on the Pavia dataset: The size of Pavia is significantly small compared with Chikusei and Cave. It has only one 1096×1096 image with the available part 1096×714. Similarly, we crop non-overlapped patches with a spatial resolution of 120×714. For each image in Pavia, three patches are used for testing and the rest for training.

As shown in Tab. 2, ESSAformer reaches the best performance and significantly outperforms other methods in most metrics. For example, it obtains 38.79 dB and 31.61 dB in PSNR and has more than 1 dB gain when comparing the second. ESSAformer obtains the best 0.0275 and 0.955 for RMSE and CC, respectively. This confirms the superiority of the proposed techniques for small datasets, *i.e.*, the channel-wise inductive bias in ESSA. It is noted the performance decreases when the scale factor rises from 2× to 8× because of the increased task difficulty. The visual comparisons shown in Fig. 5 also demonstrate ESSA's superiority.

Table 3. Comparison of different methods on the Harvard dataset.

Methods	Harvard 2×		Harvard 4×	
	PSNR ↑	SAM ↓	PSNR ↑	SAM ↓
DCM-NET [54]	50.2559	2.7389	45.4087	3.3102
SSPSR [19]	50.2929	2.7017	45.2164	3.4292
Ours	**50.6928**	**2.6489**	**45.5091**	**3.3031**

Experiments on the Harvard dataset: In addition to the aforementioned three datasets, we conduct experiments on the Harvard dataset [8] and compare ESSAFormer with other methods. We follow the training settings in [54] for our method and present the results in Tab. 3. It can be seen that ESSAFormer outperforms the others by a significant margin regarding both PSNR and SAM metrics, demonstrating the effectiveness of our proposed model.

Method	CNN	MHSA[37]	Swin[29]	ESSA(0)	ESSA(1)	ESSA(2)	ESSA(3)
PSNR	40.5672	40.4310	40.1122	40.5474	40.7648	40.7891	40.8012
SAM	2.2455	2.3835	2.5107	2.3000	2.2126	2.2100	2.2106
SSIM	0.9473	0.9456	0.9426	0.9466	0.9487	0.9491	0.9495
MACs(G)	62.23	77.14	50.51	47.12	48.65	50.47	52.75
Params(M)	13.47	11.64	11.64	11.1	11.1	11.1	11.1

Table 4. The ablation study results of ESSAformer with different attention mechanisms.

Blocks	PSNR	SAM	SSIM	MACs(G)
3	40.4200	2.3703	0.9451	31.58
5	40.7648	2.2126	0.9487	48.39
7	40.7314	2.2359	0.9488	63.39
9	40.7191	2.2279	0.9489	78.30

Table 5. The ablation study results regarding the impact of the block number.

	computation cost (FLOPs)			
image size	10×10	20×20	30×30	40×40
MHSA [37]	23.67 G	142.09 G	494.70 G	1326.80 G
ESSA	19.06 G	76.22 G	173.66 G	304.97 G

Table 6. Computation cost comparison between MHSA and ESSA. The scale factor and input channel dimension are $4\times$ and 128.

	GDRRN [25]	SSPSR [19]	MCNet [23]	Bi-3DQRNN [14]	DCM-NET [54]	Ours
MPSNR (dB)	39.65	40.36	39.56	39.89	40.51	40.76
MACs(G)	6.65	42.44	289.63	120.97	130.9	48.65
Params(M)	0.442	14.88	2.17	1.29	12.61	11.1
Time (per epoch)	2m23s	6m00s	46m52s	53m26s	53m04s	4m26s

Table 7. Model efficiency of different approaches on the Chikusei dataset $4\times$.

Methods	DCM-NET [54]	MCNet [23]	Bi-3DQRNN [14]	SSPSR [19]	Performer [51]
FPS (n/s)	5.58	20.00	14.12	36.75	55.74
Methods	NPRF [31]	ELAN [61]	RLFN [21]	SwinIR [28]	**Ours**
FPS (n/s)	71.43	51.47	**336.70**	63.29	151.06

Table 8. Inference speed of various methods on the Chikusei dataset $4\times$ with an NVIDIA 3090 GPU.

4.3. Ablation study

Several ablation studies are conducted to thoroughly understand the proposed network. All the experiments are based on the Chikusei dataset with a scale factor of 4 unless specifically indicated.

Ablation study on polynomial order in ESSA: To verify the effectiveness of ESSA, we compare several attention types with ESSA, and the results are demonstrated in Tab. 4. 'CNN' refers to using two 3×3 convolution layers with interval LeakyRelu to replace ESSA. For 'MHSA', we use the original Multi-head Self-Attention. For 'Swin', we substitute ESSA with the shifted window self-attention from Swin Transformer. 'ESSA(0)', 'ESSA(1)', 'ESSA(2)', and 'ESSA(3)' all denote using ESSA, while the difference is the order of Taylor expansion. '3' means that the term below or equal to the third order of the Taylor expansion is kept to approach the function. For 'ESSA(0)', the constant value of 1 is used for the mapping function $\psi()$ and the attention matrix becomes constant.

As can be seen in Tab. 4, using two approaching terms for the mapping function $\psi()$, *i.e.*, ESSA(1), is significantly better than ESSA(0) and the performance increases from 40.5474dB/2.3/0.9466 to 40.7340dB/2.2283/0.9487 on MPSNR/SAM/MSSIM. When increasing the order of kept terms, the PSNR performance improves at the cost of increased computation. Thus we choose to expand the Taylor polynomial to an order of 1 to balance between performance and computation cost. Thanks to the inductive biases in ESSA, ESSAformer fits HSIs well and has better data efficiency. Therefore, it obtains significantly better performance and less computation than CNN, MHSA, and Swin, which demonstrates the effectiveness of the proposed ESSA method in the super-resolution task.

Ablation study on the number of stages: ESSAformer uses shared parameters to save the model size and thus each stage can be regarded as a refinement process. We ablate how the number of stages affects the performance as shown in Tab. 5, where the results of PSNR, SAM, SSIM, and MACs are given for comparison. When the stage increases from 3 to 5, the performance significantly improves from 40.42 to 40.76 dB in PSNR and the computation cost also rises by 28 G. When the number of stages further increases to 7 and 9, the performance saturates. Thus we set 5 stages for the proposed ESSAformer.

Efficiency analysis: Due to the $\mathcal{O}(N^2)$ complexity, MHSA handles images with large resolutions in an expensive manner. In Tab. 6, we compare the overall computational cost with different image resolutions regarding different attention, *i.e.*, conventional MHSA and our proposed ESSA. We adopt the ESSAformer architecture for a fair comparison. It can be seen that the computational cost of ESSA and MHSA is relatively similar for infrequent small images with the resolution 10×10. With the resolution going up, the computational cost of MHSA rises sharply and the difference between MHSA and ESSA increases rapidly. The cost of MHSA is 4 times over ESSA when the input image has 40×40 resolution, *i.e.*, 1326 GFlops, a huge computation burden for the computing resources.

To further highlight the efficiency of our ESSA, we have conducted comprehensive experiments comparing the efficiency of various HSI-SR models in Tab. 7. We also evaluate the inference speed of HSI-SR SOTAs, RGB-SR SOTAs, and linear attention variants in Tab. 8. The results clearly demonstrate that our model excels in striking a balance between computational cost, training speed, and performance compared to other state-of-the-art methods. This further underscores the effectiveness of ESSAformer.

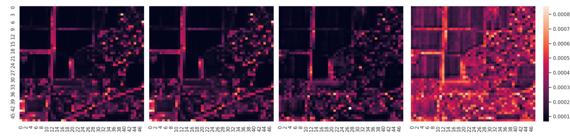
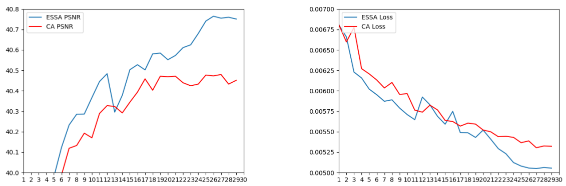


Figure 6. Visualization of the similarity map of ESSA and MHSA before and after noise for comparison.



(a) Performance (b) Train Loss

Figure 7. Comparison between the proposed ESSA and channel-wise attention.

Comparison to other linear attention: In contrast to traditional linear attention targeting RGB images, our ESSA considers the characteristics of the hyperspectral field and excels at handling HSIs. We compare ESSA and other linear attention mechanisms using the Chikusei and Pavia datasets (4× and show the experimental results in Tab. 9, where all methods use the same model architecture of ESSAFormer except attention for fair comparisons. From the table, we can see that the best performance of ESSA showcases its superiority among all attention choices, underscoring the method's effectiveness in processing HSIs.

Comparisons on large size datasets: We perform comparisons on the popular large ICVL [1] and NTIRE [2, 3] datasets, whose data volumes are larger than the previous Chikusei and Pavia. According to the results in Tab. 11, although the initial design targets improving data efficiency from small-scale datasets, ESSAFormer obtains the best performance on ICVL and NTIRE2022, demonstrating its efficacy in learning from large-size datasets. It is noted that ESSAFormer achieves comparable performance to SwinIR on NTIRE2020 while exhibiting more than 2× faster inference speed as shown in Tab. 8, showing a better trade-off between performance and efficiency.

Attention map analysis: We compare the similarity maps from ESSA and original attention in Fig. 6. We consider the first-order results for ESSA due to the infinite Fourier series decomposition. We simulate the occlusions or shadows by using scaling factors and shifts as noise and visualize the resulting heatmaps before and after the simulation. We can see ESSA recognizes a highly similar pattern to MHSA. Also, ESSA demonstrates insensitivity to the introduced noise and still focuses on the discriminative features, while MHSA collapses under the same conditions, highlighting the suitability of ESSA for denoising HSIs in the SR task. This merit originates from ESSA's superior designs and accompanies the mathematical principles, *i.e.*, channel-wise translational inductive biases, as proved in *Theorem* 3.1.

Comparison between ESSA and channel-wise attention: We plot the performance and training loss of ESSA and channel-wise attention [53, 6] in Fig. 7. The same architecture is adopted for two attentions for a fair comparison. It can be seen that the ESSA has much better training efficiency than the counterpart channel-wise attention, demonstrating ESSA's effectiveness.

Methods	Chikusei 4×		Pavia 4×	
	PSNR ↑	SAM ↓	PSNR ↑	SAM ↓
NPRF [31]	40.4344	2.3435	31.3306	7.372
Performer [51]	40.5833	2.2373	31.4690	6.2658
Linear [20]	40.3824	2.4021	31.3164	6.7403
Ours	**40.7648**	**2.2126**	**31.6126**	**6.1063**

Table 9. Comparison with different linear attention methods.

	ELAN [61]	RLFN [21]	SwinIR [28]	**Ours**
PSNR ↑	39.5227	39.6813	39.5165	**40.7648**
SAM ↓	2.6945	2.6583	2.6212	**2.2126**

Table 10. Comparisons with SOTA methods designed for RGB images on the Chikusei dataset 4×.

Methods	ICVL [1]		NTIRE2020 [2]		NTIRE2022 [3]	
	PSNR ↑	SAM ↓	PSNR ↑	SAM ↓	PSNR ↑	SAM ↓
RLFN [21]	38.2526	2.2179	34.2946	1.9554	37.0744	1.3873
SwinIR [28]	39.0727	1.9303	**34.9186**	1.8421	37.7432	1.4881
DCMNet [54]	38.7663	2.3337	34.5316	2.0283	37.4761	1.6257
Ours	**39.5158**	**1.9130**	34.8674	**1.7040**	**37.8432**	**1.2202**

Table 11. Comparisons on large-scale datasets.

5. Conclusion

This paper presents a novel Transformer network, *i.e.*, ESSAformer, for the single-hyperspectral image super-resolution task. The Transformer has an iterative refinement architecture to encode the feature representation and context information at different scales. Besides, we utilize the characteristics in HSI and propose a particular ESSA attention mechanism to effectively improve the data efficiency and model performance by involving channel-wise inductive biases. The model also significantly relieves the computation burden with strict theoretical support from kernel machines. Thanks to the particular design, the model builds long-range dependencies and produces better restoration results without involving much computing cost. Extensive experiments on different datasets at various super-resolution scales demonstrate the SOTA performance of ESSAformer regarding both visual quality and objective metrics.

6. Acknowledgement

This work is supported in part by the National Natural Science Foundation of China under Grants 62272363, 62036007, 62061047, 62176195, and U21A20514, the Young Elite Scientists Sponsorship Program by CAST under Grant 2021QNRC001, the Youth Talent Promotion Project of Shaanxi University Science and Technology Association under Grant 20200103, the Special Project on Technological Innovation and Application Development under Grant No.cstc2020jscx-dxwtB0032, and the Chongqing Excellent Scientist Project under Grant No. cstc2021ycjh-bgzxm0339.

References

[1] Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In *European Conference on Computer Vision*, pages 19–34. Springer, 2016. 9

[2] Boaz Arad, Radu Timofte, Ohad Ben-Shahar, Yi-Tun Lin, and Graham D Finlayson. Ntire 2020 challenge on spectral reconstruction from an rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 446–447, 2020. 9

[3] Boaz Arad, Radu Timofte, Rony Yahel, Nimrod Morag, Amir Bernat, Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, et al. Ntire 2022 spectral recovery challenge and data set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 863–881, 2022. 9

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 2

[5] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998. 5

[6] 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, 2022. 3, 9

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1, 2

[8] Ayan Chakrabarti and Todd Zickler. Statistics of real-world hyperspectral images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6, 7

[9] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 1, 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[11] JG Ding, XB Li, and LQ Huang. A novel method for spectral similarity measure by fusing shape and amplitude features. *Journal of Engineering Science & Technology Review*, 8(5), 2015. 4

[12] 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. 2

[13] Ying Fu, Antony Lam, Imari Sato, and Yoichi Sato. Adaptive spatial-spectral dictionary learning for hyperspectral image restoration. *International Journal of Computer Vision*, 122(2):228–245, 2017. 1

[14] Ying Fu, Zhiyuan Liang, and Shaodi You. Bidirectional 3d quasi-recurrent neural network for hyperspectral image super-resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2674–2688, 2021. 2, 6, 7, 8

[15] Paolo Gamba. A collection of data for urban area characterization. In *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium*, volume 1. IEEE, 2004. 5

[16] Zhengyang Geng, Meng-Hao Guo, Hongxu Chen, Xia Li, Ke Wei, and Zhouchen Lin. Is attention better than matrix decomposition? *arXiv preprint arXiv:2109.04553*, 2021. 3

[17] Huijuan Huang, Anthony G Christodoulou, and Weidong Sun. Super-resolution hyperspectral imaging with unknown blurring by low-rank and group-sparse modeling. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 2155–2159. IEEE, 2014. 1

[18] Hasan Irmak, Gozde Bozdagi Akar, and Seniha Esen Yuksel. A map-based approach for hyperspectral imagery super-resolution. *IEEE Transactions on Image Processing*, 27(6):2942–2951, 2018. 1

[19] Junjun Jiang, He Sun, Xianming Liu, and Jiayi Ma. Learning spatial-spectral prior for super-resolution of hyperspectral imagery. *IEEE Transactions on Computational Imaging*, 6:1082–1096, 2020. 1, 2, 6, 7, 8

[20] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020. 3, 9

[21] Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 766–776, 2022. 8, 9

[22] Hunsang Lee, Hyesong Choi, Kwanghoon Sohn, and Dongbo Min. Knn local attention for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2139–2149, 2022. 3

[23] Qiang Li, Qi Wang, and Xuelong Li. Mixed 2d/3d convolutional network for hyperspectral image super-resolution. *Remote Sensing*, 12(10):1660, 2020. 1, 2, 6, 7, 8

[24] Shutao Li, Renwei Dian, Leyuan Fang, and José M Bioucas-Dias. Fusing hyperspectral and multispectral images via coupled sparse tensor factorization. *IEEE Transactions on Image Processing*, 27(8):4118–4130, 2018. 1

[25] Yunsong Li, Jing Hu, Xi Zhao, Weiying Xie, and JiaoJiao Li. Hyperspectral image super-resolution using deep convolutional neural network. *Neurocomputing*, 266:29–41, 2017. 1, 2, 6, 7, 8

[26] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021. 4

[27] Yong Li, Lei Zhang, Chen Dingl, Wei Wei, and Yanning Zhang. Single hyperspectral image super-resolution with grouped deep recursive residual network. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pages 1–4. IEEE, 2018. 1, 2

[28] 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*, pages 1833–1844, 2021. 3, 8, 9

[29] 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*, pages 10012–10022, 2021. 3, 8

[30] Laetitia Loncan, Luis B De Almeida, José M Bioucas-Dias, Xavier Briottet, Jocelyn Chanussot, Nicolas Dobigeon, Sophie Fabre, Wenzhi Liao, Giorgio A Licciardi, Miguel Simoes, et al. Hyperspectral pansharpening: A review. *IEEE Geoscience and Remote Sensing Magazine*, 3(3):27–46, 2015. 1, 6

[31] Shengjie Luo, Shanda Li, Tianle Cai, Di He, Dinglan Peng, Shuxin Zheng, Guolin Ke, Liwei Wang, and Tie-Yan Liu. Stable, fast and accurate: Kernelized attention with relative positional encoding. *Advances in Neural Information Processing Systems*, 34:22795–22807, 2021. 3, 8, 9

[32] Shaohui Mei, Xin Yuan, Jingyu Ji, Yifan Zhang, Shuai Wan, and Qian Du. Hyperspectral image spatial super-resolution via 3d full convolutional neural network. *Remote Sensing*, 9(11):1139, 2017. 1, 2

[33] James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909. 5

[34] Frosti Palsson, Johannes R Sveinsson, and Magnus O Ulfarsson. Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, 14(5):639–643, 2017. 1

[35] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002. 5

[36] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016. 3

[37] 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*, 30, 2017. 2, 8

[38] Lucien Wald. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES, 2002. 6

[39] Qi Wang, Qiang Li, and Xuelong Li. Spatial-spectral residual network for hyperspectral image super-resolution. *arXiv preprint arXiv:2001.04609*, 2020. 1

[40] Yao Wang, Xi'ai Chen, Zhi Han, and Shiying He. Hyperspectral image super-resolution via nonlocal low-rank tensor approximation and total variation regularization. *Remote Sensing*, 9(12):1286, 2017. 1

[41] 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. 6

[42] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3365–3387, 2020. 4

[43] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 3

[44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020. 2

[45] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4794–4803, 2022. 2

[46] Jie Xiao, Xueyang Fu, Feng Wu, and Zheng-Jun Zha. Stochastic window transformer for image restoration. In *Advances in Neural Information Processing Systems*, 2022. 3

[47] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34:28522–28535, 2021. 1

[48] Jingxiang Yang, Yong-Qiang Zhao, Jonathan Cheung-Wai Chan, and Liang Xiao. A multi-scale wavelet 3d-cnn for hyperspectral image super-resolution. *Remote Sensing*, 11(13):1557, 2019. 1

[49] 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. 5, 6

[50] Naoto Yokoya and Akira Iwasaki. Airborne hyperspectral data over chikusei. *Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27*, 5, 2016. 5

[51] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021. 3, 8, 9

[52] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*, 1992. 6

[53] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 3, 5, 9

[54] Chi Zhang, Mingjin Zhang, Yunsong Li, Xinbo Gao, and Shi Qiu. Difference curvature multidimensional network for hyperspectral image super-resolution. *Remote Sensing*, 13(17):3455, 2021. 1, 2, 6, 7, 8, 9

[55] Mingjin Zhang, Qianqian Wu, Jie Guo, Yunsong Li, and Xinbo Gao. Heat transfer-inspired network for image super-resolution reconstruction. *IEEE Transactions on neural networks and learning systems*, 2022. 1

[56] Mingjin Zhang, Qianqian Wu, Jing Zhang, Xinbo Gao, Jie Guo, and Dacheng Tao. Fluid micelle network for image super-resolution reconstruction. *IEEE Transactions on Cybernetics*, 53(1):578–591, 2022. 1

[57] Mingjin Zhang, Jingwei Xin, Jing Zhang, Dacheng Tao, and Xinbo Gao. Curvature consistent network for microscope chip image super-resolution. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 1

[58] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vsa: Learning varied-size window attention in vision transformers. *arXiv preprint arXiv:2204.08446*, 2022. 2

[59] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *International Journal of Computer Vision*, pages 1–22, 2023. 1

[60] Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. Vision transformer with quadrangle attention. *arXiv preprint arXiv:2303.15105*, 2023. 1

[61] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *European Conference on Computer Vision*, pages 649–667. Springer, 2022. 8, 9

[62] Chen Zheng, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Cross aggregation transformer for image restoration. In *Advances in Neural Information Processing Systems*, 2022. 3

[63] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021. 1, 2

[64] Yuxuan Zheng, Jiaojiao Li, Yunsong Li, Jie Guo, Xianyun Wu, and Jocelyn Chanussot. Hyperspectral pansharpening using deep prior and dual attention residual network. *IEEE Transactions on Geoscience and Remote Sensing*, 58(11):8059–8076, 2020. 1

[65] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 1, 2

[66] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1, 2