

HyperTransformer: A Textural and Spectral Feature Fusion Transformer for Pansharpening

Wele Gedara Chaminda Bandara, Vishal M. Patel
Johns Hopkins University

Department of Electrical and Computer Engineering, Baltimore, MD 21218, USA

{wbandar1, vpatel136}@jhu.edu

Abstract

Pansharpening aims to fuse a registered high-resolution panchromatic image (PAN) with a low-resolution hyper-spectral image (LR-HSI) to generate an enhanced HSI with high spectral and spatial resolution. Existing pansharpening approaches neglect using an attention mechanism to transfer HR texture features from PAN to LR-HSI features, resulting in spatial and spectral distortions. In this paper, we present a novel attention mechanism for pansharpening called *HyperTransformer*, in which features of LR-HSI and PAN are formulated as queries and keys in a transformer, respectively. *HyperTransformer* consists of three main modules, namely two separate feature extractors for PAN and HSI, a multi-head feature soft-attention module, and a spatial-spectral feature fusion module. Such a network improves both spatial and spectral quality measures of the pansharpened HSI by learning cross-feature space dependencies and long-range details of PAN and LR-HSI. Furthermore, *HyperTransformer* can be utilized across multiple spatial scales at the backbone for obtaining improved performance. Extensive experiments conducted on three widely used datasets demonstrate that *HyperTransformer* achieves significant improvement over the state-of-the-art methods on both spatial and spectral quality measures. Implementation code and pre-trained weights can be accessed at <https://github.com/wgcban/HyperTransformer>.

1. Introduction

Hyperspectral (HS) pansharpening aims to spatially enhance Low-Resolution Hyperspectral Images (LR-HSIs) by transferring textural (spatial) details from better spatial resolution panchromatic (PAN) images, while preserving the spectral characteristics of LR-HSIs [28, 48]. The recent advancements in HS pansharpening greatly improve the amount of spectral and textural details in HSIs, which is indeed a crucial pre-processing for many remote sensing ap-

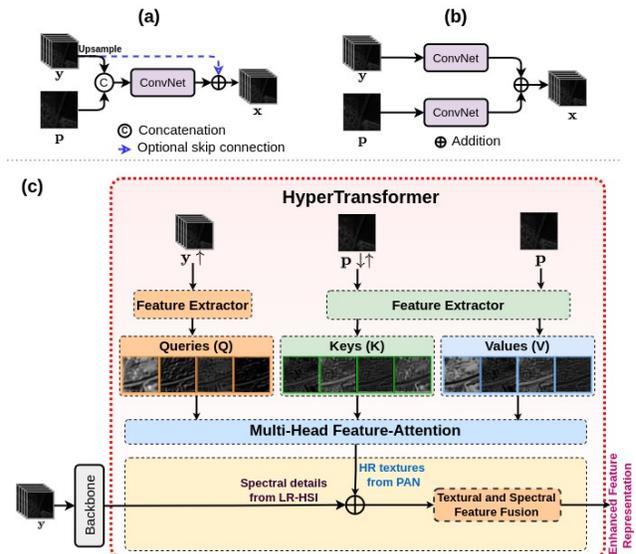


Figure 1: How our *HyperTransformer* differs from existing pansharpening architectures. Traditional pansharpening methods simply concatenate PAN (p) and LR-HSI (y) in (a) image domain [58, 22] or (b) feature domain [14, 41, 45] to learn the mapping function from LR-HSI to pansharpened HSI (x). In contrast, (c) our *HyperTransformer* utilizes feature representations of LR-HSI, PAN \downarrow , and PAN as Queries (Q), Keys (K), and Values (V) in an attention mechanism to transfer most relevant HR textural features to spectral features of LR-HSI from a backbone network. The output of *HyperTransformer* is an enhanced version of the feature representation of y . \uparrow and \downarrow denote bicubic upsampling and down-sampling, respectively.

plications to accurately and rapidly identify the underlying phenomena that would otherwise be difficult to see from LR-HSIs. HS pansharpening can be beneficial in a broad range of remote sensing tasks such as unmixing [8], change detection [37, 5], object recognition [31], scene interpretation [21], and classification [26, 6].

The early research on HS pansharpening employed component substitution (CS) [10, 17, 19], multi-resolution anal-

ysis (MRA) [29, 9], Bayesian [4, 15], and variational [32, 12, 50] methods to transform spatial details from PAN image to LR-HSI. However, these traditional pansharpening approaches often result in spatial and spectral distortions due to improper modeling of prior knowledge, inaccessibility of sensor characteristics, the mismatch between prior assumptions with the problem [57] (such as linear spectral mixture assumption [54] and the sparsity assumption [60]), and reliance on hand-crafted features such as dictionary [60, 61] with limited representation ability.

Recently, deep convolutional neural networks (ConvNets) have also been introduced for HS pansharpening due to their excellent ability to learn proper image features. However, state-of-the-art (SOTA) approaches often adopt straightforward ways to transfer textural and spectral details from PAN image to LR-HSI. For example, Lee et al. [23], Zheng et al. [58] and Bandara et al. [7] adopted a network shown in Figure 1-(a) as the backbone to learn the mapping function from the concatenation of up-sampled LR-HSI and PAN to the pansharpened HSI. However, we argue that the concatenation of PAN image along with hundreds of LR spectral bands makes textural and spectral feature fusion difficult, and inefficient. In addition, it could result in high spectral and spatial distortions in pansharpened HSI due to the inappropriate mixing of textural-spectral details. In contrast to the image-domain concatenation, researchers have also investigated feature-domain concatenation of PAN and LR-HSI as shown in Figure 1-(b). In this approach, two separate ConvNets are utilized to extract HR textural patterns from PAN, and spectral properties from LR-HSI [41, 14]. However, still the mixing process of textural and spectral details is just the addition without any appropriate guidance/attention over features. We argue that the above approaches do not effectively utilize the cross-feature space dependency between LR-HSI and PAN, and the long-range details of PAN during the textural-spectral mixing process. Instead, they completely rely upon the succeeding convolutional operations to propagate relevant textural-spectral features through the network. Although the convolution operation with sufficient depth is able to fuse the textural-spectral features appropriately to some extent, it is not intended to adjust each pixel value based on global (long-range) spectral-spatial details of the feature maps, but to adjust values of the small spatial regions together by employing the convolution kernel, which is not accurate and appropriate specially in HS pansharpening.

Motivated by a recent work on image super-resolution [47], we propose a novel textural-spectral feature fusion transformer called *HyperTransformer* for HS pansharpening that addresses the aforementioned issues of conventional pansharpening approaches as depicted in Figure 1-(c). In contrast to conventional pansharpening approaches, our *HyperTransformer* utilizes an attention

mechanism to extract cross-feature space dependency between PAN and LR-HSI features, and finds texturally advanced and more spectrally similar features for LR-HSI before fusion, which greatly helps to obtain pansharpened HSI with simultaneously high spectral and spatial qualities. Formally, our *HyperTransformer* consists of four interconnected modules, namely two feature extraction modules for PAN and LR-HSI called FE-PAN and FE-HSI, the attention mechanism, and textural-spectral feature fusion module (TSFF). Our *HyperTransformer* begins by transforming PAN and LR-HSI to their respective feature space by employing FE-PAN and FE-HSI, respectively. We then utilize LR-HSI, PAN $\downarrow\uparrow$, and PAN features as queries (Q), keys (K), and values (V) in an attention mechanism to compute texturally advanced and spectrally similar feature representations for LR-HSI features. The computed texturally advanced feature maps are then mixed with LR-HSI features from a backbone network which constitutes the pansharpened HSI. Furthermore, to obtain visually appealing pansharpened HSIs, we also introduce two new loss terms to the HS pansharpening, namely perceptual loss and transfer-perceptual loss in addition to the widely adopted L_1 loss. In summary, this paper makes the following contributions:

- We propose a novel transformer network called *HyperTransformer* for HS pansharpening which achieves significant improvements over SOTA approaches. To the best of our knowledge, we are one of the first to introduce fusion transformer architecture for HS pansharpening.
- We propose a novel *multi-scale* feature fusion strategy for HS pansharpening which enables our network to effectively capture multi-scale long-range details and cross-feature space dependencies of PAN and LR-HSI by employing *HyperTransformers* at different scales of the backbone network.
- We also introduce two novel loss functions for HS pansharpening, namely *synthesized perceptual loss* and *transfer perceptual loss* which enables our *HyperTransformer* to learn more powerful feature representations of PAN and LR-HSI.

2. Related Work

Classical approaches. Classical pansharpening approaches can be divided into four categories: component substitution (CS) [10, 17, 19], multi-resolution analysis (MRA) [29, 9], hybrid [25], and Bayesian methods [4, 15]. CS-based methods first decompose LR-HSI into spectral and spatial components. Subsequently, the spatial component is substituted with the PAN image and transformed back to the original space by employing the inverse transformation. The widely employed algorithms such as Gram-Schmidt (GS) [3], GS-adaptive (GSA)[20], and

principal component analysis (PCA) [18, 39] are examples of CS. The MRA-based pansharpening methods inject spatial features to LR-HSI by employing a spatial filter. The smoothing filter-based intensity modulation (SFIM) [27], MTF-GLP (MG) [1], MTF-GLP with high-pass modulation (MGH) [2] are examples of MRA. Considering the limitations of the CS and MRA, hybrid methods have been proposed, such as guided filter PCA (GFPCA) [25]. Bayesian methods formulate the fusion problem in a Bayesian inference framework. Examples of these include Bayesian Fusion (BF) [43], sparse BF (BSF) [42], and coupled non-negative matrix factorization (CNMF).

ConvNet-based approaches. ConvNet-based pansharpening approaches have recently shown significant progress in pansharpening due to their strong capability to learn high-level features from input data. Among those methods, Masi et al. [30] are the first to present a three-layer ConvNet architecture taking the up-sampled LR-HSI staked with PAN as input. Inspired by the wide adaptation of ResNet [13] in image super-resolution tasks, a deep residual pansharpening network (DRPNN) was proposed in [44] to learn the residual image between reference HSI and up-sampled HSI. Motivated by the 3-D characteristics of HSI data, Palsson et al. [33] proposed a 3d-ConvNet which has shown promising results when LR-HSI is corrupted by additive noise. Later, Dian et al. [11] proposed a deep pansharpening approach called DHSIS, which integrates priors learned by a ConvNet into the fusion of LR-HSI and PAN features. In order to improve the spectral prediction capability of HS pansharpening networks, two spectrally predictive ConvNet models called HyperPNN1 and HyperPNN2 were designed in [14]. More recently, attention mechanisms [58, 51, 59] (i.e., spectral and spatial attention) have been introduced to HS pansharpening to capture long-range details present in PAN and LR-HSI. In [58] (DHP-DARN), spectral and spatial attention residual blocks are utilized to map the residual image between the reference HR-HSI and the upsampled HSI, and has achieved better fusion performance compared to other SOTA methods. However, none of these attention-based methods mentioned above have explored attention for textural-spectral feature fusion process of pansharpening by utilizing feature representations of LR-HSI, PAN $\downarrow\uparrow$, and PAN as queries, keys, and values - which we will explore in this study.

3. Methodology

The overall structure of our HyperTransformer is shown in Figure 2, where \mathbf{p} , $\mathbf{p} \downarrow\uparrow$, and $\mathbf{y} \uparrow$ represent the PAN image, the sequentially $4\times$ down-sampled and $4\times$ up-sampled PAN image, and $4\times$ upsampled LR-HSI, respectively. We use bicubic interpolation for upsampling/down-sampling due to its experimentally proven less spatial and spectral distortions for HSIs [28]. The sequential down-sampling and up-sampling operations make $\mathbf{p} \downarrow\uparrow$ to

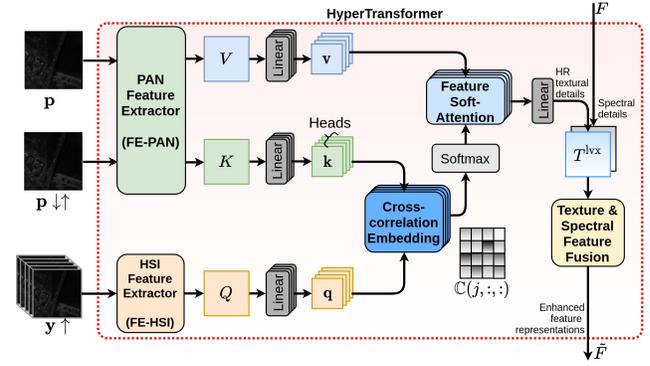


Figure 2: Overall structure of the proposed HyperTransformer for textural-spectral feature fusion.

domain-consistent with $\mathbf{y} \uparrow$: which is essential for smooth and reliable operation of the attention mechanism. HyperTransformer takes \mathbf{p} , $\mathbf{p} \downarrow\uparrow$, $\mathbf{y} \uparrow$, and the LR-HSI features (\mathbf{F}) produced by the backbone as the inputs, and outputs texturally advanced and spectrally similar feature representations of \mathbf{F} (denoted as $\hat{\mathbf{F}}$), which will be further used to generate the pansharpened HSI by the backbone network. The proposed transformer has four main modules: two separate feature extractors for PAN and HSIs (FE-PAN and FE-HSI), a feature attention module, and a textural-spectral feature fusion module. Next, we will discuss each of the modules in detail.

3.1. Feature Extractors for PAN and LR-HSI

We design two separate Feature Extractors (FE) to obtain HR textural and spectral features from PAN and LR-HSI, respectively. We employ VGG-like network architecture for the FEs (see Figure 3). The VGG-like design encourages the learning of precise mutual spectral and textural information in the LR-HSI and PAN image. The outputs from the FEs define the Query (Q), Key (K), and Value (V) features, which are the three basic elements of the attention mechanism inside the HyperTransformer. Formally, Q , K , and V are obtained as follows:

$$Q = f_{\text{FE-HSI}}(\mathbf{y} \uparrow), \quad (1)$$

$$K = f_{\text{FE-PAN}}(\mathbf{p} \downarrow\uparrow), \quad (2)$$

$$V = f_{\text{FE-PAN}}(\mathbf{p}), \quad (3)$$

where $f_{\text{FE-HSI}}(\cdot)$ and $f_{\text{FE-PAN}}(\cdot)$ are the parametric representations of FE-HSI and FE-PAN, respectively.

3.2. HR Texture Transfer through Multi-Head Feature Soft-Attention (MHFSA)

Feature attention aims to identify spectrally similar and texturally superior feature representations for LR-HSI features, which will be further used to produce pansharpened HSI by the backbone network. For this purpose, we utilize a multi-head feature soft-attention (MHFSA) mechanism instead of a single-head feature-attention due to experimentally recognized better spatial and spectral properties of pansharpened HSI. To facilitate MHFSA, we first derive

a set of N global descriptors for each feature in Q , K , and V by utilizing N fully-connected layers. Next, we compute feature soft-attention for each descriptor in parallel. Finally, we concatenate the output feature descriptors from feature soft-attention and employ linear layers to convert them back to the original feature space. The proposed MHFSA mechanism greatly assists the network in extracting cross-feature space dependencies between PAN and LR-HSI and long-range details of PAN.

Obtaining a set of N global feature descriptors. We first reshape queries $Q \in \mathbb{R}^{f_q \times w \times h}$, keys $K \in \mathbb{R}^{f_k \times w \times h}$, and values $V \in \mathbb{R}^{f_v \times w \times h}$ into 2D tensors $q \in \mathbb{R}^{f_q \times wh}$, $k \in \mathbb{R}^{f_k \times wh}$, and $v \in \mathbb{R}^{f_v \times wh}$. Note that we discard the batch dimension from our notations for simplicity. The f_q , f_k , and f_v represent the number of feature maps in Q , K , and V , and w and h represent the width and height of a feature map, respectively. Next, we utilize a set of N linear (fully-connected) layers to transform each feature map in q , k , and v into a set of N global descriptors (heads) to facilitate multi-head feature soft-attention. The resulting N global descriptors for each feature map of q , k , and v represent 3-D tensors $\mathbf{q} \in \mathbb{R}^{f_q, N, \beta wh}$, $\mathbf{k} \in \mathbb{R}^{f_k, N, \beta wh}$, and $\mathbf{v} \in \mathbb{R}^{f_v, N, \beta wh}$, where β denotes the dimensionality reduction ratio. Formally, we can define the above process as:

$$\mathbf{q}(j, i, :) = f_{\text{linear-q}}^i(q(j, :)), \quad (4)$$

$$\mathbf{k}(j, i, :) = f_{\text{linear-k}}^i(k(j, :)), \quad (5)$$

$$\mathbf{v}(j, i, :) = f_{\text{linear-v}}^i(v(j, :)), \quad (6)$$

where $f_{\text{linear-q}}^i(\cdot)$, $f_{\text{linear-k}}^i(\cdot)$, and $f_{\text{linear-v}}^i(\cdot)$ are the parametric representation of the i -th linear-layer associated with query, key, and value, $q(j, :)$, $k(j, :)$, and $v(j, :)$ are the 1-D representation of the j -th feature map of q , k , and v , and $\mathbf{q}(j, i, :)$, $\mathbf{k}(j, i, :)$, and $\mathbf{v}(j, i, :)$ are the i -th global descriptor of the j -th feature map of q , k , and v , respectively.

Feature Cross-Correlation Embedding (FCCE). We compute the feature cross-correlation matrices between query (\mathbf{q}) and key (\mathbf{k}) for all N descriptors separately, and represent them in a 3-D matrix $\mathbb{C} \in \mathbb{R}^{N \times f_q \times f_k}$. In order to efficiently compute feature cross-correlation for all the N descriptors in parallel using matrix multiplication (i.e., without any “for” loops), we first permute the first two dimensions (dim 0 and 1) of \mathbf{q} , \mathbf{k} , and \mathbf{v} . The resulting permuted matrices are denoted as $\mathbf{q}' \in \mathbb{R}^{N \times f_q \times \beta wh}$, $\mathbf{k}' \in \mathbb{R}^{N \times f_k \times \beta wh}$, and $\mathbf{v}' \in \mathbb{R}^{N \times f_v \times \beta wh}$. We then compute the feature cross-correlation for N descriptors at once as follows:

$$\mathbb{C} = \text{MatMul}((\mathbf{q}' - \text{mean}(\mathbf{q}')), (\mathbf{k}' - \text{mean}(\mathbf{k}'))^T), \quad (7)$$

where T denotes the matrix transpose operation, MatMul denotes the batch matrix multiplication on dim-1 and 2, and $\text{mean}(\cdot)$ denotes the mean value. The rows of cross-correlation matrix for the j -th descriptor (i.e., $\mathbb{C}(j, :, :)$) tells us how a given query descriptor (i.e., an LR-HSI feature)

correlates with all the key descriptors. In other words, it extracts the cross-feature space dependencies between LR-HSI features and PAN features (note that the queries and values are the feature representations of LR-HSI and PAN $\downarrow\uparrow$, respectively). We then utilize a Softmax layer along the rows of each correlation matrix in \mathbb{C} to get the row-normalized cross-correlation matrices. Formally, we can define this process as follows:

$$\tilde{\mathbb{C}} = \text{Softmax}(\mathbb{C}, \text{dim} = 1), \quad (8)$$

where $\tilde{\mathbb{C}}$ contained row normalized (sums to 1) feature cross-correlation matrices of N descriptors.

Multi-Head Feature Soft-Attention (MHFSA). We then compute feature soft-attention on the N descriptors at once using matrix multiplication as follows:

$$t = \text{MatMul}(\tilde{\mathbb{C}}, \mathbf{v}') \quad (9)$$

where $t \in \mathbb{R}^{N, f_q, \beta hw}$ is the output of the MHFSA. Next, we permute the dimensions of t to its original format ($\mathbb{R}^{N, f_q, \beta hw} \rightarrow \mathbb{R}^{f_q, N, \beta hw}$) and apply a linear layer followed by reshaping to obtain texturally advanced feature representation $T \in \mathbb{R}^{f_q \times h \times w}$ from HyperTransformer as follows:

$$T \in \mathbb{R}^{f_t \times h \times w} = \text{Linear}(t), \quad (10)$$

$$T \in \mathbb{R}^{f_t \times h \times w} \leftarrow \text{Reshape}(T), \quad (11)$$

where f_t is the number of features in T . Note that to comply with matrix multiplications and additions we set $f_q = f_k = f_v = f_t$.

3.3. Textural-Spectral Feature Fusion (TSFF)

The texturally advanced and spectrally similar feature representation T of LR-HSI features Q that we obtain through MHFSA are further concatenated with spectral features F from the backbone network as shown in Figure 2. Next, we employ a 3×3 convolution followed by a Batch Normalization (BN) layer to fuse textural-spectral details together and generate the residual component required for the backbone network which will be further used to generate the pansharpened HSI. Formally, the process inside TSFF can be define as follows:

$$\tilde{F} = \text{BatchNorm}(\text{Conv}(\text{Cat}(T, F))), \quad (12)$$

where \tilde{F} is the output of HyperTransformer which will be further used by the backbone network to generate pansharpened HSI, and Cat denotes concatenation operation.

3.4. Multi-Scale Feature Fusion (MSFF)

Unlike the conventional pansharpening methods that fuse textural features from PAN only at the HR scale, we inject textural details from our HyperTransformer to the backbone network at multiple spatial scales, as depicted in Figure 3. In particular, we inject HR-textural details at three spatial scales: (1) LR-HSI spatial scale (denoted by $\times 1 \uparrow$), (2) two times upsample LR-HSI spatial scale (denoted by $\times 2 \uparrow$), and (3) desired HR spatial scale (denoted by $\times 4 \uparrow$). Accordingly, we denote the inputs and outputs of HyperTransformer as $X^{\times s \uparrow}$ in general where X could be

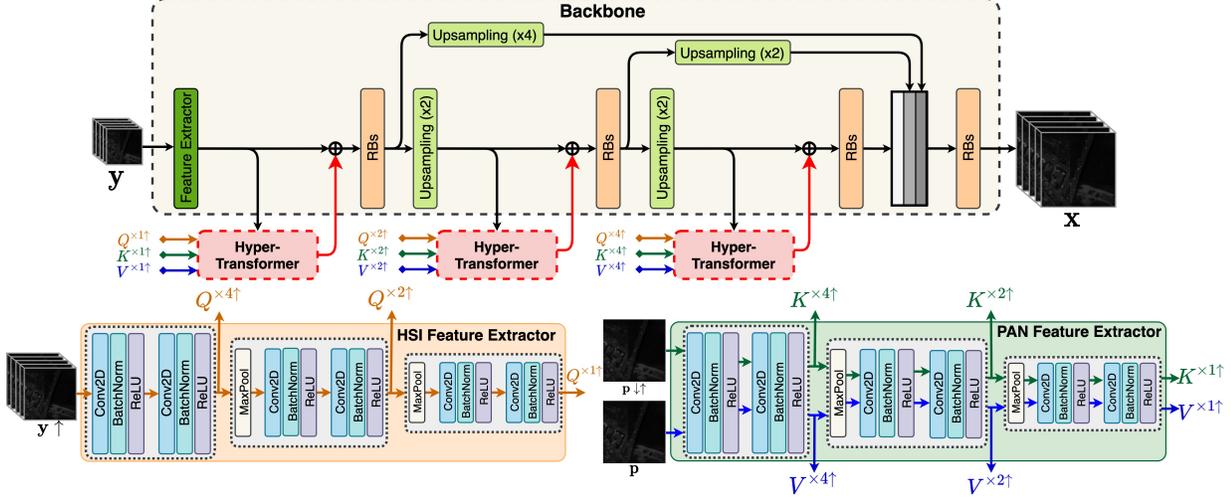


Figure 3: The complete pansharpening network. Note that we apply HyperTransformer at three scales: $\times 1 \uparrow$, $\times 2 \uparrow$, and $\times 4 \uparrow$. RBs denotes the residual blocks.

Q, K, V, F or \tilde{F} , and s represents the spatial-scale: 1, 2, or 4 (see Figure 3). Injecting HR-textural knowledge at multiple spatial scales helps the network to capture multi-scale long-range details and multi-scale cross-feature space dependencies of PAN and LR-HSI, resulting in better spatial and spectral quality of pansharpened HSI.

3.5. Loss Functions

We utilize a combination of three loss functions to train our network.

Reconstruction loss. We use L_1 loss as the reconstruction loss:

$$\mathcal{L}_{\text{rec}} = \frac{1}{CWH} \|\mathbf{x}_{\text{ref}} - \mathbf{x}\|_1, \quad (13)$$

where \mathbf{x}_{ref} is the target HR-HSI, \mathbf{x} is the predicted HR-HSI, and (C, H, W) is the size of the HR-HSI. We utilize L_1 loss which has been demonstrated to perform better compared to L_2 loss for HS pansharpening [28].

VGG perceptual loss. The VGG perceptual loss has been originally demonstrated useful for RGB super-resolution tasks to enhance the visual quality of images [16]. The underlying idea of the perceptual loss is to enhance the similarity in the feature space between the predicted image and the target image. The feature maps of predicted and target image are obtained from a pre-trained VGG network which is trained on RGB images. In order to evaluate VGG loss for HSI images, we first synthesize RGB image for a given HSI by defining Gaussian approximated spectral response curve for R, G, and B bands. Next, we evaluate the perceptual loss as:

$$\mathcal{L}_{\text{vgg-per}} = \frac{1}{C_i W_i H_i} \left\| f_i^{\text{vgg}}(\mathbf{x}_{\text{ref}}^{\text{rgb}}) - f_i^{\text{vgg}}(\mathbf{x}^{\text{rgb}}) \right\|_2, \quad (14)$$

where $f_i^{\text{vgg}}(\cdot)$ denotes the i -th layer's feature map of VGG-19 [36], and (C_i, H_i, W_i) represents the shape of the feature map at that layer. $\mathbf{x}_{\text{ref}}^{\text{rgb}}$ and \mathbf{x}^{rgb} are the synthesized RGB images of the target HSI \mathbf{x}_{ref} and predicted HSI \mathbf{x} , respectively.

Transfer perceptual loss. The transfer perceptual loss constrains the predicted HSI image \mathbf{x} to have similar texture features to the transferred texture features T from HyperTransformer, which makes our HR texture transfer process more effective. The transfer perceptual loss is calculated as follows:

$$\mathcal{L}_{\text{t-per}} = \frac{1}{C_s W_s H_s} \|f_{\text{FE-HSI}}(\mathbf{x})^s - T^s\|_2, \quad (15)$$

where T^s denotes the transferred feature map at the s -th spatial scale (i.e., 1, 2, or 4), $f_{\text{FE-HSI}}(\mathbf{x})^s$ is the feature map at the s -th spatial scale from the HSI feature extractor, and (C_s, W_s, H_s) represents the size of the feature map at that scale, respectively.

The overall loss function we use to train our network is defined as follows:

$$\mathcal{L}_{\text{overall}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{vgg-per}} \mathcal{L}_{\text{vgg-per}} + \lambda_{\text{t-per}} \mathcal{L}_{\text{t-per}}, \quad (16)$$

where $\lambda_{\text{rec}}, \lambda_{\text{vgg-per}}$ and $\lambda_{\text{t-per}}$ are regularization constants. We set $\lambda_{\text{rec}} = 1.0, \lambda_{\text{vgg-per}} = 0.1$ and $\lambda_{\text{t-per}} = 0.05$.

4. Experiments

4.1. Datasets and Performance Metrics

We use three publicly available and widely used HSI datasets for our experiments, namely Pavia Center [34], Botswana [38], and Chikusei [52]. Following the experimental and data preparation procedure outlined in [7] and [58], we create cubic patches of size $102 \times 160 \times 160$, $145 \times 120 \times 120$, and $128 \times 256 \times 256$ as the reference HSIs (\mathbf{x}_{ref}) for the Pavia Center, Botswana, and Chikusei datasets, respectively. We then utilize Wald's protocol [40, 56] to generate PAN and LR-HSI from the reference HSIs. As part of the Wald's protocol, we use 8×8 Gaussian filter followed by down-sampling operator with scaling factor of 4 to generate LR-HSI images for all the three datasets. We randomly select $\sim 80\%$ of cubic patches to form the training set for each dataset and the rest of the

cubic patches are used to form the testing set. We use 10-th, 10-th, and 12-th spectral bands as the **blue**-band, 30-th, 35-th, and 20-th spectral bands as the **green**-band, and 60-th, 61-th, and 29-th spectral bands as the **red**-band when synthesizing the RGB image for Pavia Center, Botswana, and Chikusei datasets, respectively.

To evaluate the quality of the proposed pansharpening method, we use different spatial and spectral quality measures. Following [28, 58, 7], we use Cross-Correlation (CC), Spectral Angle Mapping (SAM), Root Mean Square Error (RMSE), Reconstruction Signal to Noise Ratio (RSNR), Error Relative Globale Adimensionnelle Desynthese (ERGAS), and Peak Signal to Noise Ratio (PSNR). These measures have been widely used in the HSI processing community and are appropriate for evaluating fusion in spectral and spatial resolutions.

4.2. Results and Discussion

To demonstrate the effectiveness of HyperTransformer, we compare our model with both classical and ConvNet-based SOTA methods. The classical methods include PCA[18], GFPCA[25], BF[43], BFS[42], SFIM[27], GS[3], GSA[20], MGH[2], CNMF[55], MG[1], and HySure[35], among which HySure has achieved the SOTA performance on CC, SAM, RMSE, ERGAS, and PSNR in recent years. As for the ConvNet-based methods, HyperPNN [14], PanNet [49], DHP-DARN (abbreviated as DARN) [58], DIP-HyperKite (abbreviated as HyperKite) [7], SIPSA [22], and GPPNN [46] are six recent SOTA methods which significantly outperform previous ConvNet-based methods.

Quantitative results. Table 1 shows the quantitative evaluation results. As shown in the table, our HyperTransformer significantly outperforms both classical and ConvNet-based SOTA methods on all three datasets. The percentage improvement in CC/SAM/RMSE/ERGAS/PSNR performance measures for Pavia Center, Botswana, and Chikusei datasets are $\sim 0.9/26.9/32.6/29.4/13.3\%$, $\sim 0.3/19.2/11.9/14.0/3.2\%$, and $\sim 0.6/12.7/13.6/13.8/4.1\%$, respectively. These quantitative comparison results demonstrate the superiority of HyperTransformer over the SOTA approaches.

Qualitative results. Figure 2 shows the visual evaluation results where we randomly select one image from the testing set of each dataset and present the synthesized RGB images of HSIs along with the corresponding mean absolute error (MAE) images between the reconstructed HSIs and the reference HSI. Though the difference in the synthesized RGB images is minute, we can observe the difference between each method from the MAE images. As can be seen from the MAE images, our HyperTransformer achieves significantly lower MAE than all the other methods. These visual results further demonstrate the excellent ability of HyperTransformer to extract fine details more effectively.

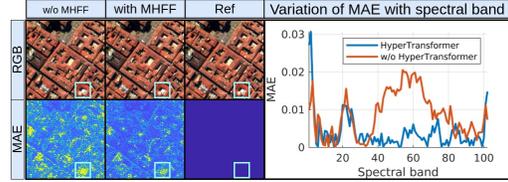


Figure 4: The visual results for the ablation study to demonstrate the effect of HyperTransformer for HS pansharpening on the Pavia Center dataset.

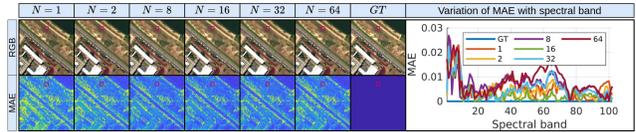


Figure 5: The visual results for the ablation study on the number of heads (N) in our HyperTransformer.

4.3. Ablation Studies

HyperTransformer. To further demonstrate the effectiveness of HyperTransformer on HS pansharpening, we conduct an ablation study, and results are presented in Table 3 and Figure 4. In this study, we consider the baseline results (B/L) as the results from our proposed pansharpening network without feature attention mechanism (i.e., we bypass the MHFSA and consider transferred texture features (T) as PAN features (V)). The proposed HyperTransformer significantly improves the baseline results in CC/SAM/RMSE/ERGAS/PSNR by $\sim 1.5/21/35/29/13\%$ when $N = 16$, respectively. In addition, Figure 4 depicts the synthesized RGB images, MAE plots, and variation of average MAE with spectral bands for a randomly selected region (marked by blue color) for with and without MHFSA. From these plots also we can clearly observe the reduction in MAE over the spectral bands (specially in infrared region) when we utilize HyperTransformer. All of these qualitative and quantitative comparisons empirically show the effectiveness of our HyperTransformer which captures long-range and cross-feature space dependencies of PAN and LR-HSI for HS pansharpening.

Number of global descriptors N . Table 3 and Figure 5 present the results when we increase the number of global descriptors in our HyperTransformer from 1 to 64. We can see that increasing the number of global descriptors results in significant improvement over spatial and spectral performance measures. However, after $N = 16$, the performance metrics start getting saturated or start degrading. Therefore, we set $N = 16$ as the optimal value. We can further verify these quantitative results by observing the qualitative results shown in Figure 5 in which we observe overall low MAE across all spectral bands when $N = 16$. Therefore, this study clarifies the reason for selecting $N = 16$.

Multi-scale feature fusion. As we discussed previously, the conventional pansharpening algorithms usually fuse PAN and LR-HSI features only at a single spatial-scale

Table 1: The average quantitative pansharpening results on the Pavia Center [34], Botswana [38], and Chikusei dataset [53].*

Method	Pavia Center Dataset [34]					Botswana Dataset [38]					Chikusei Dataset [53]				
	CC	SAM	RMSE	ERGAS	PSNR	CC	SAM	RMSE	ERGAS	PSNR	CC	SAM	RMSE	ERGAS	PSNR
			$\times 10^{-2}$					$\times 10^{-2}$					$\times 10^{-2}$		
PCA [18] ^{PERS-2014}	0.845	8.92	3.45	6.64	31.26	0.943	2.38	1.98	2.22	40.03	0.887	6.99	2.47	7.71	30.98
GFPCA [25] ^{DataFusion-2014}	0.902	8.31	3.98	7.44	29.09	0.901	2.66	2.45	2.75	37.83	0.883	4.76	1.98	7.00	30.96
BF [43] ^{JSTSP-2015}	0.918	9.60	3.44	6.63	30.22	0.931	2.47	1.88	2.34	40.01	0.903	5.15	1.94	6.62	37.89
BFS [42] ^{TGRS-2015}	0.925	8.10	3.05	6.00	31.09	0.932	2.39	1.85	2.32	40.15	0.917	4.69	1.72	6.39	37.99
SFIM [27] ^{IIRS-2000}	0.946	6.76	2.55	5.43	32.61	0.932	3.44	2.81	2.25	39.58	0.928	3.79	1.43	6.43	39.55
GS [3] ^{TGRS-2007}	0.961	6.62	2.55	4.95	32.93	0.946	2.34	1.93	2.17	40.14	0.733	5.64	2.96	8.17	35.13
GSA [20] ^{US Pat.-2000}	0.950	7.15	2.34	4.70	33.52	0.955	2.04	1.59	1.85	41.89	0.943	3.52	1.42	4.30	41.38
MGH [2] ^{PERS-2006}	0.955	6.81	2.25	4.77	33.97	0.960	2.07	1.54	1.69	42.43	0.929	3.82	1.45	6.40	39.85
CNMF [55] ^{TGRS-2011}	0.960	6.64	2.20	4.39	34.14	0.942	2.61	1.73	2.10	40.98	0.900	4.72	1.91	5.75	39.65
MG [1] ^{TGRS-2002}	0.956	6.55	2.20	4.45	34.12	0.960	2.02	1.51	1.68	42.47	0.938	3.81	1.52	4.41	41.05
HySure [35] ^{TGRS-2014}	0.966	6.13	1.80	3.77	35.91	0.956	2.15	1.46	1.77	42.30	0.960	2.98	1.13	3.69	43.14
HyperPNN [14] ^{JST-RS-2019}	0.967	6.09	1.67	3.82	36.70	0.970	1.67	1.15	1.44	44.45	0.946	3.97	1.11	4.77	41.57
PanNet [49] ^{ICCV-2017}	0.968	6.36	1.83	3.89	35.61	0.926	2.17	1.53	2.82	40.41	0.956	3.79	0.88	5.32	41.90
DARN [58] ^{TGRS-2020}	0.969	6.43	1.56	3.95	37.30	0.973	1.58	1.09	1.35	44.42	0.953	3.60	1.05	4.44	42.24
HyperKite [7] ^{TGRS-2021}	0.980	5.61	1.29	2.85	38.65	0.979	1.46	1.01	1.21	45.53	0.974	2.85	1.03	3.62	43.53
SIPSA [24] ^{CVPR-2021}	0.948	5.27	2.38	4.52	33.65	0.901	2.34	2.20	2.54	38.55	0.947	2.87	1.06	5.09	41.02
GPPNN [46] ^{CVPR-2021}	0.963	6.52	1.91	4.05	35.36	0.962	1.90	1.36	1.65	43.01	0.970	2.75	0.66	4.24	44.07
Ours	0.989	3.85	0.87	2.01	43.80	0.982	1.18	0.89	1.04	46.97	0.980	2.40	0.57	3.12	45.87

*Higher values of CC and PSNR, and lower values of SAM, RMSE, and ERGAS indicate good pansharpening performance. The ideal values of CC, SAM, RMSE, ERGAS, and PSNR are 1, 0, 0, 0, and ∞ , respectively. Color convention: **best**, **2nd-best**, and **3rd-best**.

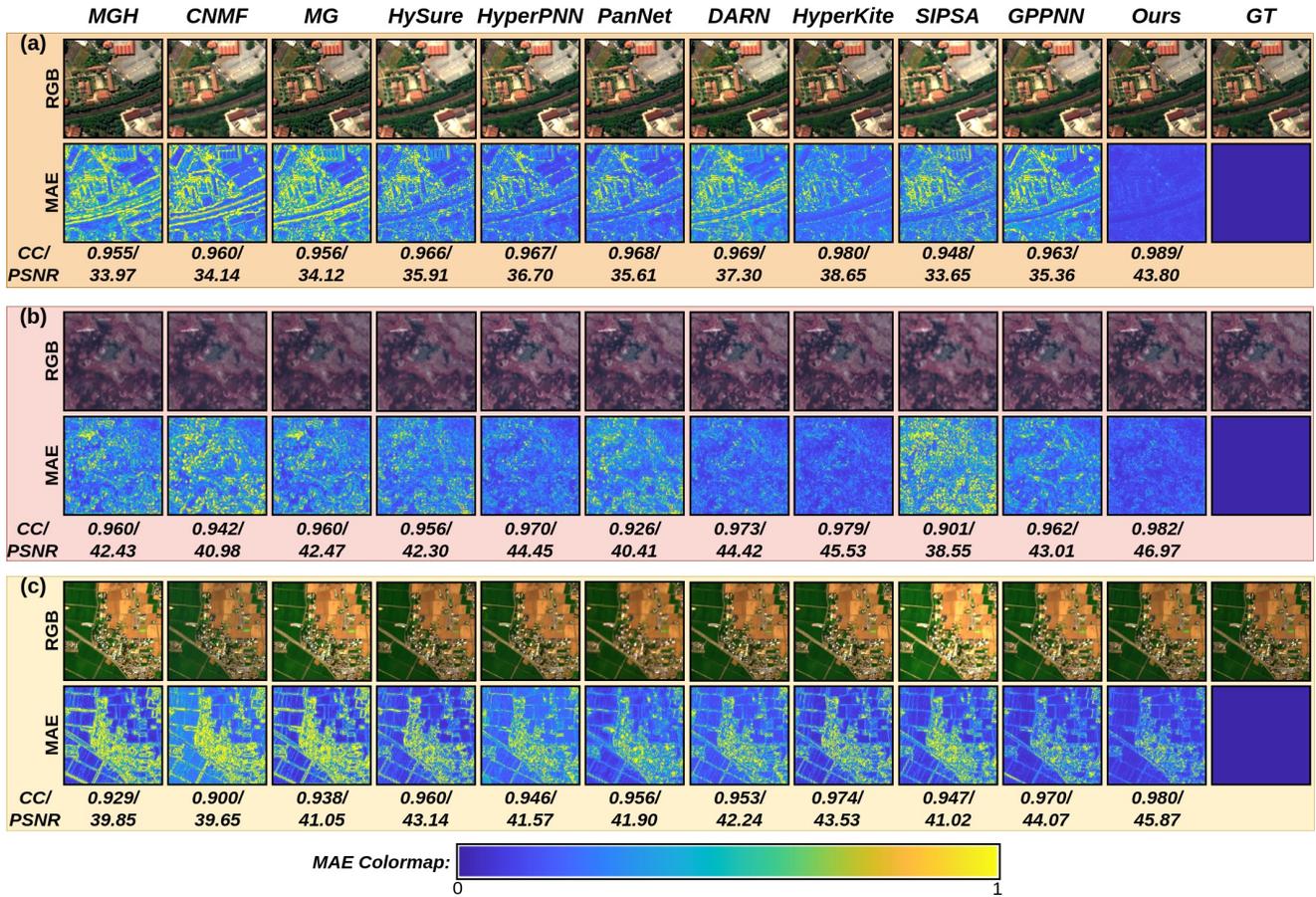


Table 2: Visual results generated by different pansharpening algorithms (left to right: MGH [2], CNMF [55], MG [1], HySure [35], HyperPNN [14], PanNet [49], DARN [58], HyperKite [7], SIPSA [22], GPPNN [46], HyperTransformer (**ours**), and Ground Truth (GT)) for (a) Pavia Center [34] (20-th patch), (b) Botswana [38] (12-th patch), (c) Chikusei datasets [53] (32-nd patch). MAE denotes the (normalized) Mean Absolute Error across all spectral bands.

Table 3: Ablation study on the number of heads (N) in HyperTransformer for the Pavia Center dataset.

N	CC	SAM	RMSE	ERGAS	PSNR
B/L	0.981	4.88	0.0133	2.84	38.71
1	0.975	4.32	0.0103	2.31	40.59
2	0.976	4.19	0.0095	2.18	42.52
8	0.987	4.06	0.0092	2.13	43.20
16	0.989	3.85	0.0087	2.01	43.80
32	0.988	4.02	0.0090	2.10	43.47
64	0.987	4.04	0.0091	2.12	43.19

Table 4: Ablation study on utilizing HyperTransformer at multiple scales for the Pavia Center dataset.

$\times 1$	$\times 2$	$\times 4$	CC	SAM	RMSE	ERGAS	PSNR
B/L			0.956	4.86	0.0204	3.90	35.81
✓	✗	✗	0.975	4.76	0.0149	3.00	38.42
✗	✓	✗	0.985	4.29	0.0108	2.40	40.80
✓	✓	✗	0.985	4.41	0.0109	2.38	41.02
✗	✗	✓	0.986	4.01	0.0098	2.21	42.88
✓	✗	✓	0.988	3.96	0.0092	2.20	43.58
✗	✓	✓	0.988	3.85	0.0089	2.09	43.60
✓	✓	✓	0.989	3.85	0.0087	2.01	43.80

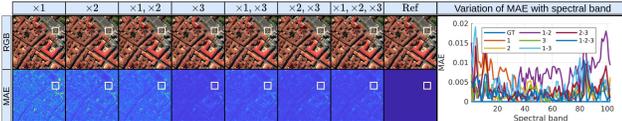


Figure 6: The visual results for the ablation study on the effect of HyperTransformer at multiple scales on the Pavia Center dataset.

(i.e., $\times 4$). Instead, our proposed pansharpening network injects textural features from PAN at three different scales, namely $\times 1$, $\times 2$, and $\times 4$ of spatial resolution of LR-HSI. To demonstrate the reason for utilizing HyperTransformers at multiple spatial-scales, we conduct an ablation study, and the results are presented in Table 4 and Figure 6. In Table 4, B/L corresponds to the case where no PAN features are injected to the backbone at any spatial-scale (i.e., sharpening without PAN image). As can be seen from Table 4, when we inject PAN features to the backbone, the quality of pansharpened HSI improves, and a significant improvement can be noticed when we inject PAN features to the backbone at HR scale (i.e., $\times 4$). Furthermore, as can be seen from the final row of Table 4, the best pansharpening performance is observed when we utilize HyperTransformers across all the three spatial-scales. Concretely, we observe an improvement of $\sim 0.3/3.7/11.2/9.0/2.1\%$ in CC/SAM/RMSE/ERGAS/PSNR when we utilize HyperTransformers at all scales compared to HyperTransformer only at the $\times 4$ spatial-scale. In addition to the quantitative results, we also present synthesized RGB images, MAE

Table 5: Ablation study on different loss functions.

L_1	$\mathcal{L}_{\text{vgg-per}}$	$\mathcal{L}_{\text{t-per}}$	CC	SAM	RMSE	ERGAS	PSNR
✓	✗	✗	0.987	4.07	0.0091	2.12	43.00
✓	✓	✗	0.988	4.01	0.0090	2.09	43.60
✓	✓	✓	0.989	3.85	0.0087	2.01	43.80

plots, and the variation of MAE with spectral bands for a randomly selected region in Figure 6. All the above observations verify that the multi-scale feature fusion is better than the conventional single-scale feature fusion for HS pansharpening.

VGG perceptual loss and Transfer perceptual loss. Table 5 shows how each loss function improves the quality of the pansharpened HSI. Combining the synthesized perceptual loss $\mathcal{L}_{\text{vgg-per}}$ with L_1 loss improves CC/SAM/RMSE/ERGAS/PSNR metrics by $\sim 0.1/1.5/1.1/1.4/1.4\%$, respectively. It is further improved by the transferred perceptual loss $\mathcal{L}_{\text{t-per}}$ in CC/SAM/RMSE/ERGAS/PSNR by $\sim 0.1/4.0/3.3/3.8/0.5\%$, respectively. Note that the significant improvement of PSNR by $\mathcal{L}_{\text{vgg-per}}$, and improvement of SAM/RMSE/ERGAS metrics by $\mathcal{L}_{\text{t-per}}$.

More experimental results and analysis of the proposed method can be found in the supplementary document.

5. Limitations and Future Work

From the MAE figures, a relatively high MAE can be observed around UV (\sim bands 1-10) and IR (\sim bands 90-104) regions. This could be due to the lack of UV and IR features in the PAN image. Additional research must be conducted to improve the performance in these regions.

6. Conclusion

In this paper, we proposed a novel textural-spectral feature fusion network for HS pansharpening called HyperTransformer, which transfers HR textural features from PAN image to spectral features from LR-HSI through a multi-head feature soft-attention mechanism. The proposed HyperTransformer consists of two separate feature extractors to extract PAN and LR-HSI features, a multi-head feature soft-attention network to capture the long-range and cross feature-space dependencies between PAN and LR-HSI, and a textural-spectral feature fusion module to fuse HR texture features and spectral features effectively. Furthermore, the proposed HyperTransformer can be utilized in multiple spatial scales to learn more powerful texture representations. Extensive experiments conducted on three widely used HSI datasets demonstrate the superiority of our HyperTransformer over the SOTA approaches on both quantitative and qualitative evaluations.

7. Acknowledgment

This work was supported by NSF CARRER award 2045489.

References

- [1] Bruno Aiazzi, Luciano Alparone, Stefano Baronti, and Andrea Garzelli. Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Transactions on geoscience and remote sensing*, 40(10):2300–2312, 2002.
- [2] B Aiazzi, L Alparone, S Baronti, A Garzelli, and M Selva. Mtf-tailored multiscale fusion of high-resolution ms and pan imagery. *Photogrammetric Engineering & Remote Sensing*, 72(5):591–596, 2006.
- [3] Bruno Aiazzi, Stefano Baronti, and Massimo Selva. Improving component substitution pansharpening through multivariate regression of ms + pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3230–3239, 2007.
- [4] Coloma Ballester, Vicent Caselles, Laura Igual, Joan Verdera, and Bernard Rougé. A variational model for p+xs image fusion. *International Journal of Computer Vision*, 69(1):43–58, 2006.
- [5] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. *arXiv preprint arXiv:2201.01293*, 2022.
- [6] Wele Gedara Chaminda Bandara, Jeya Maria Jose Valanarasu, and Vishal M Patel. Spin road mapper: Extracting roads from aerial images via spatial and interaction space graph reasoning for autonomous driving. *arXiv preprint arXiv:2109.07701*, 2021.
- [7] Wele Gedara Chaminda Bandara, Jeya Maria Jose Valanarasu, and Vishal M. Patel. Hyperspectral pansharpening based on improved deep image prior and residual reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.
- [8] José M Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE journal of selected topics in applied earth observations and remote sensing*, 5(2):354–379, 2012.
- [9] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pages 671–679. Elsevier, 1987.
- [10] Wjoseph Carper, Thomasm Lillesand, and Ralphw Kiefer. The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data. *Photogrammetric Engineering and remote sensing*, 56(4):459–467, 1990.
- [11] Renwei Dian, Shutao Li, Anjing Guo, and Leyuan Fang. Deep hyperspectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):5345–5355, 2018.
- [12] Xueyang Fu, Zihuang Lin, Yue Huang, and Xinghao Ding. A variational pan-sharpening with local gradient constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10265–10274, 2019.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [14] Lin He, Jiawei Zhu, Jun Li, Antonio Plaza, Jocelyn Chanussot, and Bo Li. Hyperpnn: Hyperspectral pansharpening via spectrally predictive convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(8):3092–3100, 2019.
- [15] Xiyan He, Laurent Condat, José M Bioucas-Dias, Jocelyn Chanussot, and Junshi Xia. A new pansharpening method based on spatial and spectral sparsity priors. *IEEE Transactions on Image Processing*, 23(9):4160–4174, 2014.
- [16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.
- [17] P Kwarteng and A Chavez. Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis. *Photogramm. Eng. Remote Sens.*, 55(1):339–348, 1989.
- [18] P Kwarteng and A Chavez. Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis. *Photogramm. Eng. Remote Sens.*, 55(1):339–348, 1989.
- [19] Craig A Laben and Bernard V Brower. Process for enhancing the spatial resolution of multispectral imagery using pansharpening, Jan. 4 2000. US Patent 6,011,875.
- [20] Craig A Laben and Bernard V Brower. Process for enhancing the spatial resolution of multispectral imagery using pansharpening, Jan. 4 2000. US Patent 6,011,875.
- [21] Florence Laporterie-Déjean, Hélène de Boissezon, Guy Flouzat, and Marie-José Lefèvre-Fonollosa. Thematic and statistical evaluations of five panchromatic/multispectral fusion methods on simulated pleiades-hr images. *Information Fusion*, 6(3):193–212, 2005.
- [22] Jaehyup Lee, Soomin Seo, and Munchurl Kim. Sipsa-net: Shift-invariant pan sharpening with moving object alignment for satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [23] Jaehyup Lee, Soomin Seo, and Munchurl Kim. Sipsa-net: Shift-invariant pan sharpening with moving object alignment for satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10166–10174, 2021.
- [24] Jaehyup Lee, Soomin Seo, and Munchurl Kim. Sipsa-net: Shift-invariant pan sharpening with moving object alignment for satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10166–10174, 2021.
- [25] Wenzhi Liao, Frieke Van Coillie, Sidharta Gautama, Aleksandra Piurica, and Wilfried Philips. Fusion of thermal infrared hyperspectral and vis rgb data using guided filter and supervised fusion graph. 2014.
- [26] GA Licciardi, Alberto Villa, Muhammad Murtaza Khan, and Jocelyn Chanussot. Image fusion and spectral unmixing of hyperspectral images for spatial improvement of classification maps. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, pages 7290–7293. IEEE, 2012.
- [27] JG Liu. Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving

- spatial details. *International Journal of Remote Sensing*, 21(18):3461–3472, 2000.
- [28] Laetitia Loncan, Luis B De Almeida, Jose 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.
- [29] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. In *Fundamental Papers in Wavelet Theory*, pages 494–513. Princeton University Press, 2009.
- [30] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016.
- [31] Ali Mohammadzadeh, Ahad Tavakoli, and Mohammad J Valadan Zoej. Road extraction based on fuzzy logic and mathematical morphology from pan-sharpened ikonos images. *The photogrammetric record*, 21(113):44–60, 2006.
- [32] Frosti Palsson, Johannes R Sveinsson, and Magnus O Ulfarsson. A new pansharpening algorithm based on total variation. *IEEE Geoscience and Remote Sensing Letters*, 11(1):318–322, 2013.
- [33] 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.
- [34] Antonio Plaza, Jon Atli Benediktsson, Joseph W. Boardman, Jason Brazile, Lorenzo Bruzzone, Gustavo Camps-Valls, Jocelyn Chanussot, Mathieu Fauvel, Paolo Gamba, Anthony Gualtieri, Mattia Marconcini, James C. Tilton, and Giovanna Trianni. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113:S110–S122, 2009. Imaging Spectroscopy Special Issue.
- [35] Miguel Simoes, José Bioucas-Dias, Luis B Almeida, and Jocelyn Chanussot. A convex formulation for hyperspectral image superresolution via subspace-based regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6):3373–3388, 2014.
- [36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [37] Carlos Souza Jr, Laurel Firestone, Luciano Moreira Silva, and Dar Roberts. Mapping forest degradation in the eastern amazon from spot 4 through spectral mixture models. *Remote sensing of environment*, 87(4):494–506, 2003.
- [38] S.G. Ungar, J.S. Pearlman, J.A. Mendenhall, and D. Reuter. Overview of the earth observing one (eo-1) mission. *IEEE Transactions on Geoscience and Remote Sensing*, 41(6):1149–1159, 2003.
- [39] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586, 2014.
- [40] Lucien Wald. Quality of high resolution synthesised images: Is there a simple criterion? In *Third conference " Fusion of Earth data: merging point measurements, raster maps and remotely sensed images"*, pages 99–103. SEE/URISCA, 2000.
- [41] Jiaming Wang, Zhenfeng Shao, Xiao Huang, Tao Lu, and Ruiqian Zhang. A dual-path fusion network for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [42] Qi Wei, José Bioucas-Dias, Nicolas Dobigeon, and Jean-Yves Tourneret. Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7):3658–3668, 2015.
- [43] Qi Wei, Nicolas Dobigeon, and Jean-Yves Tourneret. Bayesian fusion of multi-band images. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):1117–1127, 2015.
- [44] Yancong Wei, Qiangqiang Yuan, Huanfeng Shen, and Liangpei Zhang. Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. *IEEE Geoscience and Remote Sensing Letters*, 14(10):1795–1799, 2017.
- [45] Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, and Tian-Jing Zhang. Dynamic cross feature fusion for remote sensing pansharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14687–14696, 2021.
- [46] Shuang Xu, Jianshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pansharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1366–1375, 2021.
- [47] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Bain-ing Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020.
- [48] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pansharpening. In *Proceedings of the IEEE international conference on computer vision*, pages 5449–5457, 2017.
- [49] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pansharpening. In *Proceedings of the IEEE international conference on computer vision*, pages 5449–5457, 2017.
- [50] Jing Yao, Danfeng Hong, Jocelyn Chanussot, Deyu Meng, Xiaoxiang Zhu, and Zongben Xu. Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution. In *European Conference on Computer Vision*, pages 208–224. Springer, 2020.
- [51] Jing Yao, Danfeng Hong, Jocelyn Chanussot, Deyu Meng, Xiaoxiang Zhu, and Zongben Xu. Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution. In *European Conference on Computer Vision*, pages 208–224. Springer, 2020.
- [52] Naoto Yokoya and Akira Iwasaki. Airborne hyperspectral data over chikusei. *Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27*, 2016.
- [53] Naoto Yokoya and Akira Iwasaki. Airborne hyperspectral data over chikusei. *Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27*, 2016.

- [54] Naoto Yokoya, Takehisa Yairi, and Akira Iwasaki. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2):528–537, 2011.
- [55] Naoto Yokoya, Takehisa Yairi, and Akira Iwasaki. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2):528–537, 2011.
- [56] Yongnian Zeng, Wei Huang, Maoguo Liu, Honghui Zhang, and Bin Zou. Fusion of satellite images in urban area: Assessing the quality of resulting images. In *2010 18th International Conference on Geoinformatics*, pages 1–4. IEEE, 2010.
- [57] Lei Zhang, Jiangtao Nie, Wei Wei, Yanning Zhang, Shengcai Liao, and Ling Shao. Unsupervised adaptation learning for hyperspectral imagery super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3073–3082, 2020.
- [58] 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.
- [59] Man Zhou, Xueyang Fu, Jie Huang, Feng Zhao, Aiping Liu, and Rujing Wang. Effective pan-sharpening with transformer and invertible neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.
- [60] Xiao Xiang Zhu and Richard Bamler. A sparse image fusion algorithm with application to pan-sharpening. *IEEE transactions on geoscience and remote sensing*, 51(5):2827–2836, 2012.
- [61] Xiao Xiang Zhu, Claas Grohnfeldt, and Richard Bamler. Exploiting joint sparsity for pansharpening: The j-sparsefi algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, 54(5):2664–2681, 2015.

Supplementary Material for HyperTransformer: A Textural and Spectral Feature Fusion Transformer for Pansharpining

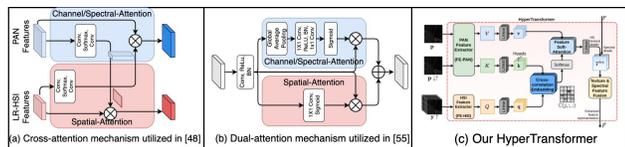


Figure 7: How our proposed *HyperTransformer* differs from the attention mechanisms utilized in previous pansharpining works.

How our proposed HyperTransformer differs from previous pansharpining methods that utilize attention? In previous pansharpining works [48, 55], attention (channel and spatial) mechanisms are used to *re-weight* the PAN and LR-HSI features from ConvNets along the channel and spatial dimensions as shown in Fig. 7 (a) and (b) without having an explicit consideration of special properties of PAN and LR-HSI features. Different from these previous methods, the proposed HyperTransformer is specifically designed to cater to the pansharpining problem by taking into consideration spatial and spectral properties of PAN and LR-HSI. Instead of simply re-weighting the feature maps, our HyperTransformer first computes the cross-correlation between the feature representations of PAN $\downarrow\uparrow$ and LR-HSI. Then multi-head feature soft-attention (MHFA) is utilized to identify texturally advanced and spectrally similar feature representations from PAN that will be further mixed with spectral features from the backbone network. Hence, the proposed HyperTransformer re-defines queries, keys, and values in standard attention mechanisms as LR-HSI, PAN $\downarrow\uparrow$, and PAN features, respectively that not only deliver better intuitive understanding to the pansharpining problem under the context of attention but also result in better pansharpining performance. HyperTransformer outperforms many previous classical [1, 2, 3, 16, 18, 23, 25, 33, 40, 41, 52], ConvNet-based [5, 12, 22, 44, 46], and attention-based [55] methods in terms of CC, SAM, RSNR, ERGAS, and PSNR on three datasets.

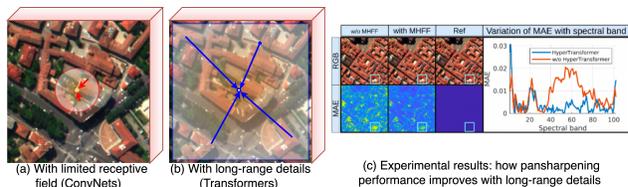


Figure 9: Necessity of long-range details for pansharpining.

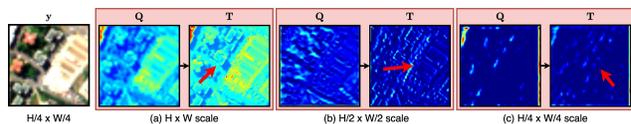


Figure 8: For a given query feature Q , we visualize corresponding spectrally similar and texturally advanced feature map T processed from our *HyperTransformer* at multiple spatial scales.

Visualization of *in* and *out* feature maps from our HyperTransformer. As shown in Fig. 8, at each spatial scale, HyperTransformer adds missing texture details to LR-HSI features (queries - Q) while maintaining their spectral characteristics (i.e., the cross-correlation).

Why are long-range details necessary for pansharpining? We explain our intuition of why pansharpining should benefit from long-range details in Fig. 9. As shown in Fig. 9, when the pansharpining network has a larger receptive field (i.e., it can capture long-range details), it can enhance the texture and spectral details of a given pixel not only by looking at adjacent pixels but also from the pixels far away. As shown in 9 - (c) (in paper Fig. 4), we can see a significant reduction in MAE across the spectral bands when we add our HyperTransformer to the main pansharpining network, which empirically shows that pansharpining indeed benefits from long-range details. Furthermore, it has been shown in the literature that not only segmentation, detection, and classification tasks benefit from long-range details but also restoration, fusion, and super-resolution [?]. Pansharpining consists of both super-resolution and fusion

tasks and as a result it should also benefit from long-range details.