Article

Multi-Scale CNN-Transformer Dual Network for Hyperspectral Compressive Snapshot Reconstruction

Kaixuan Huang *, Yubao Sun and Quan Gu

Engineering Research Center of Digital Forensics, Ministry of Education, Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; sunyb@nuist.edu.cn (Y.S.); 20211249321@nuist.edu.cn (Q.G.)
* Correspondence: 20211249413@nuist.edu.cn

Abstract: Coded aperture snapshot spectral imaging (CASSI) is a new imaging mode that captures the spectral characteristics of materials in real scenes. It encodes three-dimensional spatial–spectral data into two-dimensional snapshot measurements, and then recovers the original hyperspectral image (HSI) through a reconstruction algorithm. Hyperspectral data have multi-scale coupling correlations in both spatial and spectral dimensions. Designing a network architecture that effectively represents this coupling correlation is crucial for enhancing reconstruction quality. Although the convolutional neural network (CNN) can effectively represent local details, it cannot capture long-range correlation well. The Transformer excels at representing long-range correlation within the local window, but there are also issues of over-smoothing and loss of details. In order to cope with these problems, this paper proposes a dual-branch CNN-Transformer complementary module (DualCT). Its CNN branch mainly focuses on learning the spatial details of hyperspectral images, and the Transformer branch captures the global correlation between spectral bands. These two branches are linked through bidirectional interactions to promote the effective fusion of spatial–spectral features of the two branches. By utilizing characteristics of CASSI imaging, the residual mask attention is also designed and encapsulated in the DualCT module to refine the fused features. Furthermore, by using the DualCT module as a basic component, a multi-scale encoding and decoding model is designed to capture multi-scale spatial–spectral features of hyperspectral images and achieve end-to-end reconstruction. Experiments show that the proposed network can effectively improve reconstruction quality, and ablation experiments also verify the effectiveness of our network design.

Keywords: snapshot compressive imaging; spatial–spectral correlation; Transformer; complementary fusion

1. Introduction

Hyperspectral imaging technology can capture the spectral information of real-world scenes within a specific wavelength range. Compared with three-channel visible RGB images, hyperspectral images usually have a wider spectral response range and a finer spectral sampling resolution. Each pixel represents spectral signatures of objects in the scene and can be used to better identify their categories. Therefore, hyperspectral images are widely used in various tasks, including ground object classification [1–4], target detection and tracking, and change detection [2,5]. In order to obtain complete spatial and spectral information, conventional hyperspectral imaging technology needs to scan in spectral or spatial dimensions. However, the scanning operation requires more imaging time and is not suitable for capturing dynamic scenes. Recently, a new hyperspectral imaging technology, called snapshot compressive imaging (SCI), has emerged; it is based on compressive sensing theory, which can significantly improve imaging efficiency. In these SCI systems, coded aperture snapshot spectral imaging (CASSI) is a potential solution. The CASSI system modulates incoming light through a coded aperture and accumulates the modulated light on a two-dimensional sensor array into a two-dimensional snapshot measurement during...
a single exposure time. Then the 3D hyperspectral image (HSI) cube can be recovered from the 2D snapshot measurement during the reconstruction stage. The design of the reconstruction algorithm is the core issue of the CASSI system and a key factor affecting imaging quality.

Traditional reconstruction methods [6–9] recover the original hyperspectral image by solving optimization problems with prior regularization constraints. Some commonly used prior constraints include sparsity [10–12], total variation [8,10,13] and non-local similarity [9,14,15], etc. However, these handcrafted regularization constraints cannot adequately represent the spatial–spectral structure in hyperspectral images, which degrades the reconstruction quality. At the same time, the reconstruction procedure requires an iterative solution and the time complexity is high. In recent years, CNNs have been widely used in computer vision-related tasks due to their excellent capability of feature learning. They can adaptively learn the hierarchical feature representation of images in a data-driven way, and the learned features can be regarded as the deep prior of the underlying images. To this end, the researchers introduce CNN into the task of compressive snapshot imaging and guide the network to learn the end-to-end mapping from 2D snapshot measurements to 3D HSI. The CNN-based method has brought an effective improvement in reconstruction quality. However, due to the small receptive field of convolution kernels, CNN has certain limitations in capturing non-local self-similarity and long-range correlations in hyperspectral images. The Transformer is a new network structure that uses multi-head self-attention to model non-local correlations. It has received much attention in the field of computer vision and is used in image classification [16–20], target detection [21–24], semantic segmentation, etc. Researchers have also proposed a Transformer-based network for compressive snapshot reconstruction. A representative Transformer-based reconstruction model (dubbed as MST) calculates self-similarity in the spectral dimension, effectively utilizing global similarity between spectral bands for reconstruction. MST [25] can achieve better reconstruction performance than the competing CNN. However, it has shortcomings in the representation of small-scale spatial structures, as the reconstructed images have the drawback of being over-smooth. A single CNN or Transformer structure cannot effectively represent the multi-scale spatial–spectral correlation of hyperspectral images. Therefore, how to effectively represent the coupling correlation between the spatial and spectral dimensions of hyperspectral images is a key issue in improving reconstruction quality.

In order to cope with the above-mentioned issues, we propose a dual-branch CNN-Transformer complementary module (DualCT). As shown in Figure 1, the CNN branch of DualCT focuses more on obtaining the spatial dimension information of hyperspectral images, while the Transformer branch of DualCT captures the global correlation between spectral bands. At the same time, we also introduce bidirectional cross-branch interactions to fuse the complementary features from the spectral-domain Transformer and spatial-domain convolution. Therefore, DualCT can effectively represent the spatial–spectral correlation in hyperspectral images. Meanwhile, considering the physical mechanism of the CASSI imaging mode, we introduce residual mask attention into the DualCT module. It uses the aperture mask to guide the network to focus on regions with high-fidelity spectral representation, thereby enhancing the spatial–spectral structure of the reconstructed hyperspectral images. Based on the DualCT module, we further build a multi-scale encoding and decoding reconstruction network (DualCT-Net) to learn the end-to-end reconstruction mapping from snapshot measurements to the original hyperspectral image. We conduct comparative experiments and ablation studies to verify the performance of the proposed network. The experimental results show that the proposed network can effectively improve reconstruction performance, and the ablation studies also verify the effectiveness of our network design. The main contributions of this paper include:

(1) The proposed DualCT module parallelizes the spectral Transformer branch and spatial CNN branch, and promotes complementary fusion of spatial and spectral features through bidirectional cross-branch interactions between the two branches. Therefore, the
DualCT module can effectively integrate the advantages of CNN and Transformer to capture the spatial–spectral features of hyperspectral images.

(2) Taking into account the physical mechanism of compressive snapshot imaging, we designed a residual mask attention mechanism to enhance the DualCT module. It can guide the DualCT module to focus on spatial regions with high-fidelity spectral representation, thereby enhancing the structural details of the reconstructed hyperspectral images.

(3) By using DualCT as the basic building block, we further construct a multi-scale encoding and decoding model to learn the inverse reconstruction mapping. The experimental results verify the effectiveness of the proposed reconstruction network.

Figure 1. The architecture of the proposed reconstruction network DualCT-Net. DualCT block acts as the basic building block of DualCT-Net.

2. Related Work

The compressive snapshot imaging system uses reconstruction algorithms to recover the original 3D hyperspectral images from the 2D snapshot measurements. According to the different models used, current reconstruction algorithms can mainly be divided into three types, namely prior regularization-based methods, deep convolutional network-based methods, and Transformer-based methods. Additionally, since we need to design mask attention based on the CASSI imaging process, we also introduce the CASSI imaging model.

2.1. Prior Regularized Reconstruction Method

Compressive snapshot reconstruction is an ill-posed inverse problem. The early work formulates it as a prior regularized optimization problem consisting of a prior regularization term and a data fidelity term. GAP-TV [8] uses total variation as a prior regularization constraint, and DeSCI [14] uses sparse representation and non-local self-similarity as a priori representation of hyperspectral images. However, these prior regularizations cannot effectively represent the complex spatial–spectral structure in hyperspectral images. At the same time, these reconstruction models require iterative optimization. Using the alternating...
direction multiplier method (ADMM) as an example, each iteration needs to solve the proximal operators of the prior regularization, and the computational complexity is high.

2.2. Deep Convolutional Network Reconstruction Method

In view of the excellent feature learning capabilities of deep networks, researchers have started using deep networks to directly learn the reconstruction mapping from snapshot measurements to original hyperspectral images [26–31]. Under this framework, Xiong et al. [32] used convolutional neural networks to learn hyperspectral compressed snapshot reconstruction. Miao et al. [30] proposed a two-stage hyperspectral image reconstruction model called λ-Net. The first stage of λ-Net employs a generative adversarial network (GAN) to obtain an initial reconstruction. The second stage is to refine each spectral band of the initial reconstruction. Meng et al. [33] expanded the iterative steps of the GAP optimization algorithm into different stages of the network model, and its denoising sub-steps were implemented using a pre-trained denoiser. TSA-Net [29] introduces a self-attention mechanism to capture the dependence between spatial and spectral dimensions of hyperspectral images. These methods based on deep networks can directly output the reconstruction results through a feed-forward network calculation without iteration. Therefore, they have high reconstruction efficiency. However, although the above-mentioned convolutional networks can effectively represent local details, they still have certain limitations in capturing long-term spatial–spectral correlation.

2.3. Transformer Reconstruction Method

The Transformer is widely used in natural language processing tasks [34]. Through long-range correlation matching and interaction between image feature tokens, the Transformer has also been introduced into computer vision tasks and brought effective performance improvement, including image classification [16,17,19,35], object detection [21–23] and segmentation [36–38]. The Transformer also shows good performance in low-level visual tasks [4,39–41]. Chen et al. [22] utilized a standard Transformer to construct the backbone model [34] for various image restoration tasks. Liu et al. [40] used the Swin Transformer to build a residual network, and achieved state-of-the-art results in image restoration tasks. Cai et al. [25] proposed a Transformer-based HSI reconstruction model named MST. It treats each spectral band as a token and computes self-attention in the spectral dimension, improving the reconstruction quality effectively. However, hyperspectral images exhibit coupled spatial–spectral correlations, and the MST model is insufficient at representing spatial correlations, which hinders the reconstruction of small-scale spatial structures.

2.4. CASSI Imaging Model

Figure 2 shows the schematic diagram of the CASSI system. The hyperspectral image corresponding to the physical scene is denoted as $F \in \mathbb{R}^{H \times W \times N_{\lambda}}$, where $H, W, N_{\lambda}$ represent the height, width, and the number of spectral bands, respectively. The CASSI system first modulates $F$ by the aperture mask $M^\ast$, which is defined as follows:

$$F'(\cdot, \cdot, n_{\lambda}) = F(\cdot, \cdot, n_{\lambda}) \odot M^\ast, 1 \leq n_{\lambda} \leq N_{\lambda},$$  

where $F' \in \mathbb{R}^{H \times W \times N_{\lambda}}$ denotes the modulated HSI, $\odot$ represents the element-wise dot multiplication. After modulation, the disperser applies a shearing transformation to the modulated hyperspectral image. Taking the Y-axis shift as an example, the transformation process can be expressed as follows:

$$F''(u, v, n_{\lambda}) = F'(x, y + d(\lambda_{n} - \lambda_{c}), n_{\lambda}),$$  

where $\lambda_{c}$ represents the reference spectral band, and $d(\lambda_{n} - \lambda_{c})$ denotes the translation amount of the $n_{\lambda}$-th spectral band in the y direction, typically a linear function of the number of spectral index. All the spectral bands after shear transformation are projected
upon the two-dimensional sensor array, and the images of different spectral bands at the same position are accumulated, which can be calculated as follows:

\[ Y = \sum_{n=1}^{N_A} F''(\cdot; n_A) + N, \]

where \( Y \in \mathbb{R}^{H \times (W+d(N_A-1))} \) represents the captured two-dimensional snapshot measurement, and \( N \in \mathbb{R}^{H \times (W+d(N_A-1))} \) represents the measurement noise of the sensor during the imaging process. Equation (3) represents the forward observation equation of CASSI imaging. Based on this observation equation, the reconstruction algorithm recovers the original hyperspectral image from the two-dimensional snapshot measurements.

Figure 2. Schematic diagram of the CASSI system principle.

3. Materials and Methods
3.1. Multi-Scale CNN-Transformer Complementary Reconstruction Network

In order to better represent the spatial–spectral coupling correlations in hyperspectral images, we integrate the advantages of CNN and the Transformer and design a dual-branch CNN-Transformer complementary module (DualCT). Using DualCT as the basic module, we further construct a reconstruction network called DualCT-Net to learn the reconstruction mapping from snapshot measurements to the original hyperspectral image. Similar to the multi-scale structure in [25], DualCT-Net is configured as an encoder, a bottleneck module, and a decoder. As shown in Figure 1, the encoder in DualCT-Net takes the snapshot measurement \( Y \) as input. Through the reverse dispersion process, the snapshot measurement \( Y \) is backward-shifted to obtain the initial reconstruction \( H_0 \in \mathbb{R}^{H \times W \times N_A} \), and \( H_0 \) is mapped into features \( X_0 \in \mathbb{R}^{H \times W \times C} \) through a 3 × 3 convolutional process. \( X_0 \) then flows into the multi-scale feature extraction stage. Each scale of the encoder contains two DualCT modules and one downsampling operation. The downsampling operation employs a convolutional layer with a stride of 4 × 4 stride to reduce the spatial dimensions while doubling the number of channels. The output features of the third scale in the encoder enter the bottleneck module and then flow into the decoder. The bottleneck contains two DualCT modules.

The decoder of DualCT-Net is designed as a symmetrical structure of the encoder. Each scale of the decoder has two DualCT modules and one upsampling module, and the upsampling is realized as a deconvolution operation. The same spatial resolution scale of the encoder and decoder are connected through skip connections, which convey features from encoder to decoder. The skip connections are beneficial for reducing the information loss caused by downsampling operations. We denote the output of the highest resolution scale of the decoder as \( X'_0 \in \mathbb{R}^{H \times W \times C} \). After a 3 × 3 convolution processing upon \( X'_0 \), we can obtain the reconstruction \( R_0 \in \mathbb{R}^{H \times W \times N_A} \). Due to the presence of global residual connections in DualCT-Net, \( R_0 \) represents the incremental relative to the initial reconstruction \( H_0 \). Finally, the reconstructed hyperspectral image \( H_\in \mathbb{R}^{H \times W \times N_A} \) can be obtained by adding \( R_0 \) and \( H_0 \). The specific structure of the DualCT module is introduced in detail below.
3.2. DualCT Module

CNN and Transformer have their own advantages in feature learning. In order to effectively represent the coupling of spatial and spectral features of HSIs, our DualCT module deploys the self-attention mechanism of the Transformer and localized convolution as two parallel branches, and uses bidirectional interactions to provide the fusion of complementary features in both spectral and spatial dimensions. At the same time, motivated by the mask guidance mechanism proposed in [25], we designed mask attention and added it to the DualCT module. According to the imaging characteristics of the CASSI system, mask attention can guide the network to focus on regions with high-fidelity HSI representation, enhancing the representation capability of spatial–spectral structural details. Specifically, as shown in Figure 1b, the DualCT module mainly consists of layer normalization (LN) [42], a parallel complementary fusion structure, and mask attention. The parallel complementary fusion structure contains a spatial domain CNN branch and a spectral domain Transformer branch. The two branches achieve effective feature fusion through bidirectional cross-linking interactions. The calculation process of DualCT can be expressed as follows:

$$
\hat{X}^l = CT(LN(X^l)) + X^l,
$$

$$
\hat{X}^{l+1} = CONV(LN(\hat{X}^l)) \odot M' + \hat{X}^l,
$$

where $X^l$ represents the feature input to the DualCT module, LN stands the layer normalization, and $CT(\cdot)$ represents the complementary fusion of dual branches. $M' \in R^{H \times W \times C}$ is the mask attention mapping, $\odot$ represents element-wise multiplication, and $\hat{X}^{l+1}$ represents the output feature of the DualCT module. We denote $CONV(LN(\hat{X}^l))$ as $X^{l+1}$. The specific design of the complementary fusion structure will be introduced in detail below.

3.2.1. Parallel Complementary Fusion

As shown in Figure 3, the parallel complementary fusion structure consists of a spatial domain CNN branch (abbreviated as Spa-CNN), a spectral domain Transformer branch (abbreviated as Spe-Trans), and bidirectional cross-linking between these two branches. In the Spa-CNN branch, it cascades three consecutive convolution operations with kernel sizes of $1 \times 1$, $5 \times 5$, and $1 \times 1$. The GELU layer is used after each convolution. The Spe-Trans branch exploits the Transformer to capture the global correlation between spectral bands, and its detailed design will be described in the next sub-section. Bidirectional cross-linking facilitates feature fusion between two branches, effectively capturing spatial–spectral correlations in the hyperspectral image. Specifically, the cross-linking is composed of spatial interaction and channel interaction. The channel interaction directs feature information from the Spa-CNN branch to the Spe-Trans branch, enhancing the feature modeling capability in the channel dimension. At the same time, the spatial interaction directs features from the Spe-Trans branch to the Spa-CNN branch, prompting the model to capture the spatial dependencies between different locations. The calculation flow of bidirectional interactions can be expressed as follows:

$$
F_c = Spa - CNN(LN(X^l)),
$$

$$
A = CI(F_c),
$$

$$
\hat{F}'_t = Spe - Trans(X^l, A),
$$

$$
\hat{F}'_c = SI(\hat{F}'_t) \odot F_c,
$$

where $X^l$ is the input of the parallel complementary fusion structure, $F_c$ is the feature output of the Spa-CNN branch, $CI(\cdot)$ is the channel interaction function, $SI$ is the spatial interaction function, and $A$ is the channel attention calculated from $F_c$ by $CI(\cdot)$. The spectral-Trans branch takes the $X^l$ and attention weight $A$ as input and uses $A$ to weight the value matrix. The detailed formulation of spectral-Trans branch will be presented in Section 3.2.2. Using
the output $F'_t$ of the spectral-Trans branch as input, we calculate the spatial attention map and use it to weigh $F_c$. Therefore, this parallel complementary structure can promote the interaction and fusion of spatial and spectral features.

![Diagram](image)

**Figure 3.** Parallel complementary fusion structure inside DualCT module.

**With regard to channel interaction**, we first use a global average pooling layer to obtain the global average feature vector of $F_c$. Then, the global average feature is processed through two consecutive $1 \times 1$ convolutional layers, and the GELU activation function [43] is used for non-linear transformation between the two convolutional layers. Finally, the channel attention map is generated along the channel dimension by applying the sigmoid function. The specific operations for channel interaction are defined as follows:

$$A = CI(F_c) = \sigma(Conv(GL(Conv(GP(F_c))))),$$  \hspace{1cm} (6)

where $\sigma$ represents the Sigmoid function, $Conv(\cdot)$ represents the $1 \times 1$ convolution, $GL$ represents the GELU activation function, and $GP$ represents the global average pooling layer. The channel attention, $A$, is used to weight each spectral band of the value matrix in the spectral Transformer.

**With regard to spatial interaction**, it calculates the spatial attention map from $F'_t$ to weight the output $F_c$ of the spatial-domain CNN branch. Specifically, it consists of two $1 \times 1$ convolutional layers, and the GELU activation function [43] is used between these two convolutional layers. Finally, the spatial attention map is generated by applying the sigmoid function. The specific operations for spatial interaction are defined as follows:
\[ SI(F'_t) = \sigma(Conv(GL(Conv(F'_t)))) , \]  

where \( SI(\cdot) \) represents the spatial interaction process, \( \sigma \) represents the Sigmoid function, \( Conv(\cdot) \) represents \( 1 \times 1 \) convolution, and \( GL \) represents the GELU activation function.

According to Equation (5), we can obtain the attention-enhanced feature \( F'_t \) and \( F'_c \). The bidirectional cross-branch interactions can effectively enhance the modeling capabilities of the Transformer and deep convolution in both spectral and spatial dimensions, providing complementary features for the two branches. \( F'_t, F'_c \) are combined through a concatenation operation and then flow into the continuous feed-forward network. FFN consists of two linear layers with a GELU layer in the middle. The final output feature \( \hat{X}_l \) of parallel complementary fusion is calculated as follows:

\[ \hat{X}_l = FFN(\text{Concat}(F'_c, F'_t)) + X_l. \]  

3.2.2. Spectral Domain Transformer Branch

Due to the effectiveness of the Transformer in capturing long-range dependencies, we utilize the Transformer branch to capture the global correlation information between spectral bands, and name it the spectral Transformer. As in [25], it treats each spectral band as a token and performs self-attention calculations between each token. Figure 4 shows the diagram of the spectral Transformer. The input \( X^l \in \mathbb{R}^{H \times W \times C} \) of the spectral Transformer is reshaped into \( X \in \mathbb{R}^{HW \times C} \), and each column of \( X \) is taken as a token. \( X \) is then linearly projected to obtain query \( Q \in \mathbb{R}^{HW \times C} \), the key \( K \in \mathbb{R}^{HW \times C} \), and the value \( V \in \mathbb{R}^{HW \times C} \). They are defined as

\[ Q = XW^Q, K = XW^K, V = XW^V, \]  

where \( W^Q, W^K, W^V \in \mathbb{R}^{C \times C} \) are learnable projection matrices. We split \( Q, K, \) and \( V \) into \( N \) heads along the channel dimension, and each head independently calculates self-attention within its channel group which is defined as follows:

\[ S_{a_j} = \text{softmax}(\mu_j K^T_j Q_j), \]
\[ V'_j = A \odot V_j, \]  

\[ \text{head}_j = V'_j S_{a_j}, \]  

where \( j \) is the index of head number, \( \mu_j \in \mathbb{R} \) is the learnable parameter. In order to introduce the complementary feature from the spatial-domain CNN branch, we use the channel interaction to weight the value matrix \( V \) with channel attention \( A \). Channel attention \( A \) is calculated according to Equation (6). Then the outputs of \( N \) heads are concatenated along the spectral dimension to form a larger feature matrix. Considering the spatial location relationship within each token, we also add positional embeddings to integrate positional information into the feature representation. The output of spectral Transformer can be calculated as

\[ \text{Spe} - \text{Trans}(X^l) = (\sqcup_{j=1}^{N} \text{head}_j)W + f_p(V'), \]  

where \( W \in \mathbb{R}^{C \times C} \) is the linear projection matrix, and \( \sqcup \) is the concatenation of \( N \) heads. The positional embeddings function \( f_p(\cdot) \) consists of two \( 3 \times 3 \) convolutional layers, a GELU activation layer, and a reshaping operation. Finally, after a reshaping operation, we can obtain the output feature map \( X_{out} \in \mathbb{R}^{H \times W \times C} \).
3.2.3. Residual Mask Attention

In the imaging process of the CASSI system, the original scene is first modulated through an aperture mask, and the aperture mask is usually set to a random matrix with elements 0 and 1. The positions of element 1 in the mask allow the light reflected from the target in the scene to pass through, so these positions have high information fidelity. Based on this imaging mode, the literature [25] designed a mask guidance mechanism to enhance the quality of reconstructed images. Inspired by [25], we designed a residual mask attention mechanism and encapsulated it into the DualCT module. This mask attention can guide the network to focus on spatial regions with high-fidelity spectral representations. As shown in Figure 5, in accordance with the CASSI imaging mechanism, we first perform the reverse dispersion on the aperture mask $M^*$:

$$M_s(x, y, n_\lambda) = M^*(x, y + d(\lambda_n - \lambda_c)),$$  \hspace{0.5cm} (12)

where $M_s \in R^{H \times (W + d(N_\lambda - 1)) \times N_\lambda}$ represents the shifted version of $M^*$. Then $M_s$ undergoes a sequence of operations to obtain the two-dimensional weight map. These operations include two consecutive $3 \times 3$ convolutional layers followed by GELU activation, a $5 \times 5$...
convolutional layer and a sigmoid activation function. With the residual link, we can obtain the spatially enhanced mask \( M'_s \in \mathbb{R}^{H \times (W+d(N_\lambda-1)) \times C} \), and it is calculated as follows:

\[
M'_s = M_s \odot (1 + \sigma(\beta^2 \phi \beta^2 M_s)), \quad (13)
\]

where \( \sigma(\cdot) \) represents the sigmoid activation function, \( \beta(\cdot) \) represents the conv 3 \times 3 convolution operation and GELU operation, and \( \phi(\cdot) \) represents the mapping function of the 5 \times 5 deep convolution layer. We perform the reverse dispersion process and shift \( M'_s \) backward to obtain the mask attention map \( M' \in \mathbb{R}^{H \times W \times C} \):

\[
M'(x, y, n_\lambda) = M'_s(x, y - d(n_\lambda - \lambda_c), n_\lambda), \quad (14)
\]

where the spectral bands are indexed as \( n_\lambda \in [1, \ldots, C] \). The obtained mask attention map \( M' \) adequately indicates the spatial positions with high-fidelity spectral information. Therefore, we integrate the mask attention \( M' \) into the DualCT module and enhance feature maps in a residual formulation:

\[
\hat{X}^{l+1} = X^{l+1} \odot M' + \hat{X}^l, \quad (15)
\]

where \( \hat{X}^{l+1} = \text{CONV} (\text{LN} (\hat{X}^l)) \) represents the deep feature map in the DualCT module, \( \hat{X}^l \) is the output feature of the parallel complementary fusion structure, and \( \hat{X}^{l+1} \) represents the output feature of the DualCT module. The mask attention is helpful for recovering spatial–spectral structures and improving reconstruction performance. Ablation experiments in the next section verify the effectiveness of mask attention.

**Figure 5.** Residual mask attention.

### 3.3. Network Parameters Learning

We train the constructed DualCT-Net in a supervised way. We denote \( \Omega \) as the training dataset. \( \Omega \) is composed of \( N \) data pairs \((x_i, y_i)\), where \( y_i \) denotes the compressive snapshot measurements corresponding to the original hyperspectral images, \( x_i \). In order to better guide the parameter learning, we design a composite loss function composed of root mean square error (RMSE) loss \( L_{\text{RMSE}} \) and spectrum constancy loss \([44] L_{\text{SCL}} \). This composite loss function is defined as follows:

\[
L_{\text{mix}}(\theta) = L_{\text{RMSE}}(\theta, \Omega) + \gamma_1 L_{\text{SCL}}(\theta, \Omega), \quad (16)
\]

where \( \theta \) represents the network parameters, and \( \gamma_1 \) denotes the weight parameters that coordinate the importance of the root mean square error and the spectral constant loss.
Specifically, the $L_{\text{RMSE}}$ loss calculates the root mean square error (RMSE) between the reconstructed images and the original images; the formula is defined as follows:

$$L_{\text{RMSE}}(\theta, \Omega) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - F_{\text{Net}}(y_i, \theta))^2}, \quad (17)$$

where $F_{\text{Net}}(y_i, \theta)$ is the hyperspectral image reconstructed from $y_i$ by DualCT-Net. Considering the spatial–spectral correlation of hyperspectral images, the spectrum constancy loss $L_{\text{SCL}}$ is introduced and is defined as follows:

$$L_{\text{SCL}}(\theta, \Omega) = \frac{1}{N} \sum_{i=1}^{N} \| \nabla_{\lambda} \Omega_i - \nabla_{\lambda} x_i \|^2, \quad (18)$$

where $\nabla_{\lambda}$ represents the gradient along the spectrum across different spectral bands. This composite loss function can constrain the reconstructed hyperspectral image to approximate in the original domain and gradient domain, resulting in better reconstruction quality.

4. Results

This section verifies the effectiveness of the proposed DualCT-Net through comparative experiments. The proposed network model is quantitatively and qualitatively compared with multiple state-of-the-art methods.

4.1. Experimental Setting

We use CAVE [5] as the training hyperspectral image set. CAVE has 205 hyperspectral images with a spatial size of $1024 \times 1024$ and 28 spectral bands. The wavelengths of these 28 spectral bands are within the wavelength range of 450–650 nm, and they are 453.3, 457.6, 462.1, 466.8, 471.6, 476.5, 481.6, 486.9, 492.4, 498.0, 503.9, 509.9, 516.2, 522.7, 529.5, 536.5, 543.8, 551.4, 558.6, 567.5, 575.3, 584.3, 594.4, 604.2, 614.4, 625.1, 636.3, and 648.1 nm, respectively. We randomly crop massive sub-images with a spatial size of $256 \times 256$ from the original large-sized image, thereby enhancing the training image set. According to the forward model of CASSI, we can obtain the snapshot measurement with a spatial size of $256 \times 310$, corresponding to each training sample. By performing a reverse dispersion process and backward shift upon the snapshot measurement, we can obtain the reconstruction $H_0$ and take it as the network input. As in [25], 10 scenes are selected from KAIST [26] for testing.

Furthermore, this paper initialized weights using the method described in [29]. All models were trained for 300 iterations using the Adam [45] optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). We used Adam as the optimizer to train the proposed DualCT-Net, the training was implemented on the PyTorch platform, and CuDNN was used for acceleration. In the network training process, the number of iterations is set to 300, the batch size is set to 5, and the initial learning rate is set to 0.0004, which is halved every 50 epochs.

4.2. Comparative Experiments

In order to evaluate the reconstruction quality, we introduce two quantitative metrics: peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [46]. Higher values for both metrics indicate better reconstruction quality. We compare the proposed DualCT-Net with four state-of-the-art deep network-based reconstruction methods: TSA-Net [29], SRN [47], GST [48], and MST [25]. A small variant of the MST model is used in our experiments and we denote it as MST-S. Table 1 shows the PSNR and SSIM values of the five reconstruction methods across ten scenes.
Table 1. PSNR and SSIM values of reconstruction results of 5 methods across 10 scenes.

<table>
<thead>
<tr>
<th>Scene</th>
<th>TSA-Net</th>
<th>SRN</th>
<th>GST</th>
<th>MST-S</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene 1</td>
<td>32.03/0.892</td>
<td>33.26/0.910</td>
<td>33.99/0.926</td>
<td>34.71/0.930</td>
<td>36.01/0.948</td>
</tr>
<tr>
<td>Scene 2</td>
<td>31.00/0.858</td>
<td>29.86/0.881</td>
<td>30.49/0.900</td>
<td>34.45/0.925</td>
<td>37.21/0.953</td>
</tr>
<tr>
<td>Scene 3</td>
<td>32.25/0.915</td>
<td>31.69/0.909</td>
<td>32.63/0.921</td>
<td>35.32/0.943</td>
<td>37.66/0.959</td>
</tr>
<tr>
<td>Scene 4</td>
<td>39.19/0.953</td>
<td>39.90/0.947</td>
<td>41.04/0.967</td>
<td>41.50/0.967</td>
<td>43.71/0.985</td>
</tr>
<tr>
<td>Scene 5</td>
<td>29.39/0.884</td>
<td>30.86/0.923</td>
<td>31.49/0.938</td>
<td>31.90/0.933</td>
<td>33.21/0.958</td>
</tr>
<tr>
<td>Scene 6</td>
<td>31.44/0.908</td>
<td>34.20/0.941</td>
<td>34.89/0.955</td>
<td>33.85/0.943</td>
<td>34.95/0.959</td>
</tr>
<tr>
<td>Scene 7</td>
<td>30.32/0.878</td>
<td>27.27/0.852</td>
<td>27.63/0.866</td>
<td>32.69/0.911</td>
<td>34.23/0.943</td>
</tr>
<tr>
<td>Scene 8</td>
<td>29.35/0.888</td>
<td>32.35/0.932</td>
<td>33.02/0.947</td>
<td>31.69/0.933</td>
<td>32.94/0.954</td>
</tr>
<tr>
<td>Scene 9</td>
<td>30.01/0.890</td>
<td>32.83/0.921</td>
<td>33.45/0.932</td>
<td>34.67/0.939</td>
<td>35.70/0.952</td>
</tr>
<tr>
<td>Scene 10</td>
<td>29.59/0.874</td>
<td>30.25/0.905</td>
<td>31.49/0.935</td>
<td>31.82/0.926</td>
<td>32.40/0.942</td>
</tr>
<tr>
<td>Average</td>
<td>31.46/0.894</td>
<td>32.25/0.912</td>
<td>33.01/0.929</td>
<td>34.26/0.935</td>
<td>35.80/0.955</td>
</tr>
</tbody>
</table>

As can be seen from Table 1, our DualCT-Net has the best reconstruction results upon all 10 test scenes. TSA-Net [29], SRN [47], and GST [48] are CNN-based reconstruction methods, and MST-S [25] is a Transformer-based reconstruction method. Our network surpasses comparative methods by a large margin. More specifically, the average PSNR value of the DualCT-Net model is 4.34 dB, 3.55 dB, 2.79 dB, and 1.54 dB higher than the state-of-the-art deep network-based methods TSA-Net, SRN, GST, and MST-S, respectively. These results verify the effectiveness of configuring the CNN + Transformer in the DualCT module. The DualCT-Net can effectively integrate the advantages of CNN and the Transformer, capturing spatial–spectral features of hyperspectral images. Figures 6 and 7 visualize spectral bands reconstructed by five methods and the ground-truth spectral bands. We also zoom in on some areas for comparison.

Figure 6. Visualization of three spectral bands reconstructed by five methods in scene 2.
It can be clearly seen that the reconstructed images of TSA-Net can recover the outline of the object, but produce blurry, distorted, and incomplete feature maps. Reconstructed images of SRN and GST are prone to having artifacts along the object contours. Compared with the MST-S method, the DualCT-Net network can reconstruct more clear details.

In addition, Figures 8 and 9 show reconstructed spectral signatures corresponding to annotated rectangular regions in the RGB images. Correlation coefficients are also calculated to quantitatively evaluate the fidelity of reconstructed spectral signatures. Spectral signatures reconstructed by our method can adequately match the reference signatures and have spectral fidelity.
5. Discussion

In this section, we conduct ablation experiments to verify the impact of the parallel complementary fusion structure and mask attention on the reconstruction performance. Additionally, the computational complexity and parameter count of these methods are also analyzed.

5.1. Ablation Experiments

In this section, we conduct ablation experiments to examine the impact of the complementary fusion structure and mask attention on the reconstruction results. The main verification method is to remove specific modules from the DualCT-Net network and determine their impact on reconstruction performance. The ablation experiments are performed on an HSI dataset [26]. The model obtained by removing our mask guidance mechanism and parallel complementary fusion structure from DualCT-Net is regarded as the baseline model. The ablation experiment results are shown in Table 2. The reconstruction quality of the baseline model is 33.88 dB, and when we use the parallel complementary fusion structure and masked attention, the reconstruction performance is continuously improved by 0.87 dB and 1.05 dB. These experimental results fully demonstrate the effectiveness of the parallel complementary fusion structure and mask attention. Next, we further analyze the impact of different settings within the parallel complementary fusion structure and mask attention on the reconstruction results.

Table 2. Ablation experiments of DualCT-Net. (√ indicates using this module).

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Parallel Complementary Fusion</th>
<th>Mask Attention</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>√</td>
<td>☐</td>
<td>☐</td>
<td>33.88</td>
<td>0.931</td>
</tr>
<tr>
<td>√</td>
<td>☑</td>
<td>☐</td>
<td>34.75</td>
<td>0.943</td>
</tr>
<tr>
<td>√</td>
<td>☑</td>
<td>☑</td>
<td>35.80</td>
<td>0.955</td>
</tr>
</tbody>
</table>

5.1.1. The Ablation of Complementary Fusion Structure

This section continues to verify the effectiveness of parallel branches and bidirectional cross-linking interactions in the DualCT module. The bidirectional cross-linking attention is composed of the channel iteration attention and spatial interaction attention. The experimental results are shown in Table 3.
Table 3. Ablation experiments of parallel branches and bidirectional cross-linking interactions in the DualCT module. (✓ indicates using this module).

<table>
<thead>
<tr>
<th>Parallel Branches</th>
<th>Interaction</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Channel</td>
<td>Spatial</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td>✓</td>
<td>33.88</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td>✓</td>
<td>34.38</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>34.61</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>34.61</td>
</tr>
</tbody>
</table>

According to the PSNR and SSIM values in Table 3, we can see that the design of parallel branches and bidirectional cross-linking interactions are both positive factors that can improve reconstruction performance.

5.1.2. Mask Attention Ablation

In this section, we further verify the impact of mask attention on reconstruction performance. Experimental results are shown in Table 4.

Table 4. Ablation experiments of mask attention in DualCT module. (✓ indicates using this module).

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>Mask Attention</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>H₀</td>
<td></td>
<td>34.75</td>
<td>0.943</td>
</tr>
<tr>
<td>B</td>
<td>H₀ ⊗ M⁺</td>
<td>✓</td>
<td>35.08</td>
<td>0.946</td>
</tr>
<tr>
<td>C</td>
<td>H₀</td>
<td>✓</td>
<td>35.80</td>
<td>0.955</td>
</tr>
<tr>
<td>D</td>
<td>H₀ ⊗ M⁺</td>
<td>✓</td>
<td>35.62</td>
<td>0.951</td>
</tr>
</tbody>
</table>

Methods A and C utilize the initial reconstruction H₀, while methods B and D adopt H₀ ⊗ M⁺ as input. Methods C and D use the Mask Attention in the DualCT module. Method B achieves limited improvements due to the HSI representation damage and insufficient mask utilization. In both cases of input, mask attention can improve the reconstruction performance. These results demonstrate the effectiveness of our mask attention design.

5.2. Computational Complexity and Parameter Amount Analysis

This section further analyzes the computational complexity and parameter amount of the five reconstruction methods. These five methods are all run on NVIDIA GTX 3090 GPU. Table 5 displays the FLOPS and the number of parameters for each reconstruction method. Due to the introduction of bidirectional interactions between the CNN branch and the Transformer branch in the DualCT module, the FLOPS value of our network becomes higher. On the other hand, the DualCT module integrates the advantages of CNN and the Transformer, bringing effective reconstruction performance improvements compared to only using the Transformer or CNN networks.

Table 5. Analysis of computational complexity and parameter quantity.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Parameter Amount (M)</th>
<th>FLOPS (G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSA-Net</td>
<td>44.25</td>
<td>110.06</td>
</tr>
<tr>
<td>SRN</td>
<td>1.25</td>
<td>81.84</td>
</tr>
<tr>
<td>GST</td>
<td>1.27</td>
<td>82.87</td>
</tr>
<tr>
<td>MST-S</td>
<td>0.93</td>
<td>12.96</td>
</tr>
<tr>
<td>Ours</td>
<td>0.92</td>
<td>122.27</td>
</tr>
</tbody>
</table>
6. Conclusions

In this paper, we aim to integrate the advantages of CNN and Transformer for representing multi-scale coupling correlations within hyperspectral images. Specifically, we propose a dual-branch CNN-Transformer complementary module. It combines the Transformer self-attention mechanism with deep convolution through dual branches and bidirectional cross-branch interactions. Therefore, it can learn complementary features in both spectral and spatial dimensions. In addition, according to the compressive snapshot imaging mode, this paper also introduces a mask guidance mechanism to refine the spatial–spectral structure of the reconstructed image. By using the DualCT module as a basic component, this paper further designs a multi-scale encoding and decoding model to learn the end-to-end reconstruction mapping from snapshot measurements to the original hyperspectral images. Quantitative experimental results have verified the effectiveness of our network design. It is hoped that the network design and potential insights proposed in this paper will contribute to future work in emerging SCI research.

Author Contributions: Conceptualization: Y.S.; data curation: K.H. and Q.G.; methodology: K.H.; software: K.H.; supervision: Y.S.; writing—original draft: K.H.; writing—review and editing: Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under grants 62276139 and U2001211.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study can be available on request from the corresponding author. The data are not publicly available due to privacy.

Acknowledgments: The authors would like to thank the people who shared data and basic models from their research with the community. Their selflessness significantly drives the research process for the entire community. In addition, we sincerely express our gratitude to Huang Junru for her assistance with the experiments and valuable discussions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References


Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.