This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Flexible Window-based Self-attention Transformer in Thermal Image Super-Resolution

Hongcheng Jiang, ZhiQiang Chen
University of Missouri-Kansas City
hjq44@mail.umkc.edu, chenzhiq@umkc.edu

Abstract

The aim of this paper is to improve the resolution of low-quality thermal images obtained from downsampled images afflicted with noise and blur, alongside high-resolution visible images, to achieve high-resolution thermal imagery. Our proposed method, named Flexible Window-based Self-attention Transformer (FW-SAT), operates across global, regional, and local scales to effectively enhance the fine details in the thermal domain. FW-SAT integrates various attention mechanisms such as channel and spatial attention, window-based self-attention, and flexible window-based self-attention. Notably, flexible window-based self-attention aggregates regional window features based on window-based self-attention, while channel and spatial attention mechanisms capture global information. Additionally, window-based self-attention is employed to explore local features within the image. We assess the performance of FW-SAT in the PBVS-2024 Thermal Image Super-Resolution Challenge (GTISR) - Track2. Our extensive experiments demonstrate that our proposed approach surpasses state-of-the-art techniques in both qualitative and quantitative evaluations. Code will be available at https://github.com/jianghongcheng/FW-SAT.

1. Introduction

Optical cameras, which operate by capturing electromagnetic waves within the visible (VIS) and near-infrared (NIR) spectrum, encounter substantial challenges when faced with adverse lighting and weather conditions. Factors such as low illumination, precipitation, and atmospheric phenomena (e.g., fog) pose significant hurdles to their imaging capabilities due to the reliance on visible light for image formation. In such conditions, the captured images often lack detail and clarity, limiting their effectiveness in scenarios where precise visualization is crucial.

In contrast, thermal sensors function independently of visible light, detecting infrared radiation emitted by objects. This unique characteristic enables thermal cameras to maintain imaging consistency even in environments characterized by fluctuating lighting conditions and atmospheric obstructions. By detecting thermal energy emitted by objects, these cameras can effectively provide augmented and reliable imaging under challenging circumstances. Therefore, thermal cameras are widely utilized across various sectors, including military, agriculture, and medical fields, serving numerous surveillance and monitoring tasks [5, 14, 19]. The emergence of the COVID-19 pandemic has also introduced a new use for thermal cameras: body temperature measurement [11].

Despite their wide utilities, images from thermal cameras suffer from their low spatial resolution. To this end, high-resolution infrared focal plane arrays, necessary for detailed imaging, still present a significant cost barrier due to hardware process limitations. Physically for an object reflecting solar illumination and emitting thermal energy, thermal detectors typically require larger pixel sizes to gather enough emitted radiation to generate a measurable signal, since thermal radiation is less intense than the visible light reflected. For examples, in satellite sensors such as Landsat 8, the thermal bands have a solution of 100 m/pixel while visible and near-infrared bands have 30m/pixel. Consequently, thermal cameras typically fall short of matching the spatial resolution of VIS/NIR cameras; for example, a modern low-cost RGB camera can readily provide high-resolution images in the megapixel range. Additionally, increasing spatial resolution in thermal sensors is constrained by factors such as the Signal to Noise Ratio (SNR) of the sensor area. Attempts to improve resolution by increasing sensor size directly correlate with higher costs, making such technology less accessible and hindering efforts to enhance spatial resolution [17].

To address these challenges without resorting to costly hardware upgrades, there is growing interest in leveraging software algorithms to enhance the spatial resolution of thermal images. Super-resolution (SR) techniques, which aim to increase image resolution while preserving high-frequency details, have garnered attention in this regard.

Particularly, the Single Image Super-Resolution (SISR) methodology involves upscaling a low-resolution (LR) image to achieve high-resolution (HR) enhancement. Dong et al.[4] pioneered the adoption of deep learning for their SRCNN model, which has since propelled CNN-based solutions outperforming traditional SR methods. Noteworthy advancements in the arena have been made through various published works [10, 24, 29].

Transformers, renowned for their self-attention mechanisms enabling the capture of long-range dependencies, are believed to excel in understanding global context. In recent years, there has been a surge in the development of Transformers-based SR techniques, aiming to enhance the quality of SR images [7, 27, 28]. However, Chen et al. [2] presented a meticulous examination of the comparative effectiveness between Transformers and Convolutional Neural Networks (CNNs) in image super-resolution tasks. Additionally, they introduced an Overlapping Cross-attention (OCA) Module to facilitate more direct interaction among adjacent window features. This novel design enhances the collaboration between neighboring pixels, resulting in improved reconstruction performance. By activating a broader range of pixels for reconstruction, the model achieves notable performance gains. In addition, Li et al. [8] introduced the Anchored Stripe Attention (ASA) module, aimed at processing images of various resolutions while simultaneously reducing computational and space complexity.

Building upon this concept, we introduce the Flexible Window-based Self-attention (FWA) module, which realizes an adaptable attention mechanism. By aiming to examine the features of adjacent windows, this approach can dynamically adjust to neighboring regions, facilitating a more comprehensive understanding of regional features within the image. It is worth emphasizing that despite employing a similar overlapping window partition strategy, there exists a fundamental disparity between the ASA module and the FWA approach. ASA is primarily intended for the exploration of global features within images, whereas FWA is explicitly crafted for exploring regional features. This highlights a crucial distinction in the objectives and design principles of the two methods.

The main contributions of this paper are summarized as follows:

- We introduce the Flexible Window-based Self-Attention Transformer (FW-SAT) architecture, a comprehensive framework that combines advanced channel and spatial attention mechanisms with window-based self-attention and a flexible window-based self-attention mechanism. This integrated approach aims to significantly enhance the performance of thermal image super-resolution by effectively addressing global, regional, and local contextual factors with precision and efficacy.
- We propose a Flexible Window-based Self-Attention (FWA) module specifically designed to collect regional window features using window-based self-attention. This innovative approach enables the model to focus and analyze information from specific regions within the input data, resulting in enhanced performance and precision in thermal image super-resolution tasks.
- We introduce a Channel Spatial Attention Block (CSAB) that delves into global features by harnessing the strengths of both channel attention and spatial attention mechanisms through concatenation. Channel attention enhances the model's capability to emphasize critical features across diverse channels, allowing for a concentrated focus on the most pertinent information. Simultaneously, spatial attention facilitates the capture of spatial relationships and contextual details within features, thereby enhancing the model's comprehension and refining its performance.
- Extensive experimentation confirms the efficacy of our FW-SAT in comparison to contemporary thermal super-resolution techniques, demonstrating superior performance across both qualitative and quantitative evaluation metrics.

2. RelatedWork

2.1. Visible Image Super-Resolution

Visible image super-resolution (VISR) has been a popular research topic for over a decade, with early methods predominantly relying on model-based approaches such as neighbor embedding regression [21] and Random Forest [18]. However, in recent years, deep learning techniques have been extended for Single Image Super-Resolution (SISR). For example, Dong et al. first introduced SRCNN [4], paved the way by effectively extracting features from LR images and learning the mapping between LR and HR features to reconstruct HR images. EDSR [10] initially employed residual blocks without batch normalization as the fundamental building blocks, forming a deeper super-resolution network. RDN [30] combined residual blocks with dense connections, introducing residual dense blocks. RCAN [29] integrated channel attention into residual blocks, proposing residual attention modules and deepening the network.

Recently, Vision Transformers (ViTs) have emerged as powerful alternatives to CNNs, overcoming inherent biases and effectively modeling long-range dependencies, thereby achieving optimal performance in various high-level visual tasks. ViT-like structures have also been applied to low-level tasks, showcasing their versatility and effectiveness. For example, Niu et al. [12] introduced a Holistic Attention Network (HAN) for single image superresolution. It dynamically captures the global dependencies across various depths, channels, and positions through the self-attention

mechanism. SwinIR [9] adopted the Swin Transformer's architecture, leveraging a shifted window mechanism to capture long-range dependencies and achieving enhanced performance with reduced parameter complexity. Additionally, DAT [3] improved single-image super-resolution performance by aggregating spatial and channel features in both interblock and intra-block dual manners, enhancing representation competence. Furthermore, GRL [8] introduced a network structure with hierarchies in the Global, Regional, and Local range, leveraging anchored stripe self-attention, window-based self-attention, and channel attention enhanced convolution to yield impressive performance gains.

2.2. Thermal Image Super-Resolution

The advancements made by deep learning in Visible Image Super-Resolution (VISR) have sparked renewed interest and research in Thermal Image Super-Resolution (TISR). Alongside [16], which introduced a novel GAN-based architecture called CycleGAN and a comprehensive thermal image dataset, several other notable contributions have emerged in this domain. Thuan et al. [20] proposed a technique to enhance the resolution of thermal images by leveraging edge features from corresponding high-resolution visible images, providing an alternative approach to TISR. Prajapati et al. [13] introduced Channel Splitting-based Convolutional Neural Network (ChasNet) for thermal image SR, aiming to eliminate redundant features in the network and enhance performance. Additionally, Wang et al. [23] presented a Camera Internal Parameters Perception Network (CIPPSRNet) which could also be applied to other cross-camera super-resolution tasks and Compared with existing state-of-the-arts thermal and natural SR methods. Furthermore, Prajapati et al. [6] presented a CoReFusion architecture that is computationally inexpensive and lightweight with the ability to maintain performance despite missing one of the modalities. These contributions collectively signify the growing momentum and exploration within the realm of Thermal Image Super-Resolution research, highlighting the diverse methodologies and techniques being explored to address this challenging problem.

3. Methodology

3.1. Motivation

The HAT model, as demonstrated by Chen et al. [2], has exhibited remarkable performance in image super-resolution. The authors assert that leveraging more information leads to improved performance. Additionally, Li et al. [8] introduced the Anchored Stripe Attention (ASA) module, designed to process images of varying resolutions while concurrently reducing computational complexity and memory usage. Notably, they introduce the concept of anchors in addition to the traditional triplets of queries, keys, and values. Anchors serve as a condensed representation of the information within the image feature map, possessing lower dimensionality. Based on the ideas presented in the two papers, it can be inferred that the efficacy of image super-resolution techniques can be significantly improved by incorporating more information and introducing novel attention mechanisms. Additionally, there is a clear need to develop a network architecture that effectively integrates global, regional, and local features.

To address the modeling of global, regional, and local features for thermal image super-resolution, we introduce a novel Flexible Window-based Self-Attention Transformer (FW-SAT). Our FW-SAT architecture incorporates channel and spatial attention, window-based self-attention, and flexible window-based self-attention mechanisms. Specifically, the flexible window-based self-attention is devised to aggregate regional window features based on window-based self-attention.

This proposed architecture aims to leverage diverse attention mechanisms to effectively capture global, regional, and local information within thermal images. The channel and spatial attention mechanisms enable the model to focus on relevant channels and spatial locations, respectively, allowing for the selective enhancement of features at different scales. The window-based self-attention mechanism facilitates the extraction of regional features by attending to specific windows within the image, thereby capturing contextually relevant information. Additionally, the flexible window-based self-attention mechanism further refines feature representations by adaptively attending to informative regions based on the learned attention weights.

By combining these attention mechanisms within the FW-SAT architecture, we aim to create a versatile and adaptive framework for thermal image super-resolution. This approach enables the model to effectively exploit global context, capture regional details, and preserve local information, ultimately leading to enhanced performance in super-resolving thermal images.

3.2. Network Architecture

3.2.1 The Overall Structure

Fig. 1 illustrates the overall architecture of our network, which comprises three key components: shallow feature extraction, deep feature extraction, and image reconstruction. This architectural approach is commonly utilized in the HAT model [2]. The network architecture of the Flexible Window-based Self-attention Transformer (FW-SAT) for thermal image super-resolution commences with the input of the downsampled thermal image and visible image, thus reaching the upsampled thermal image resolution. Initially, bicubic interpolation is applied to enhance the thermal image's resolution to match that of the visible image.

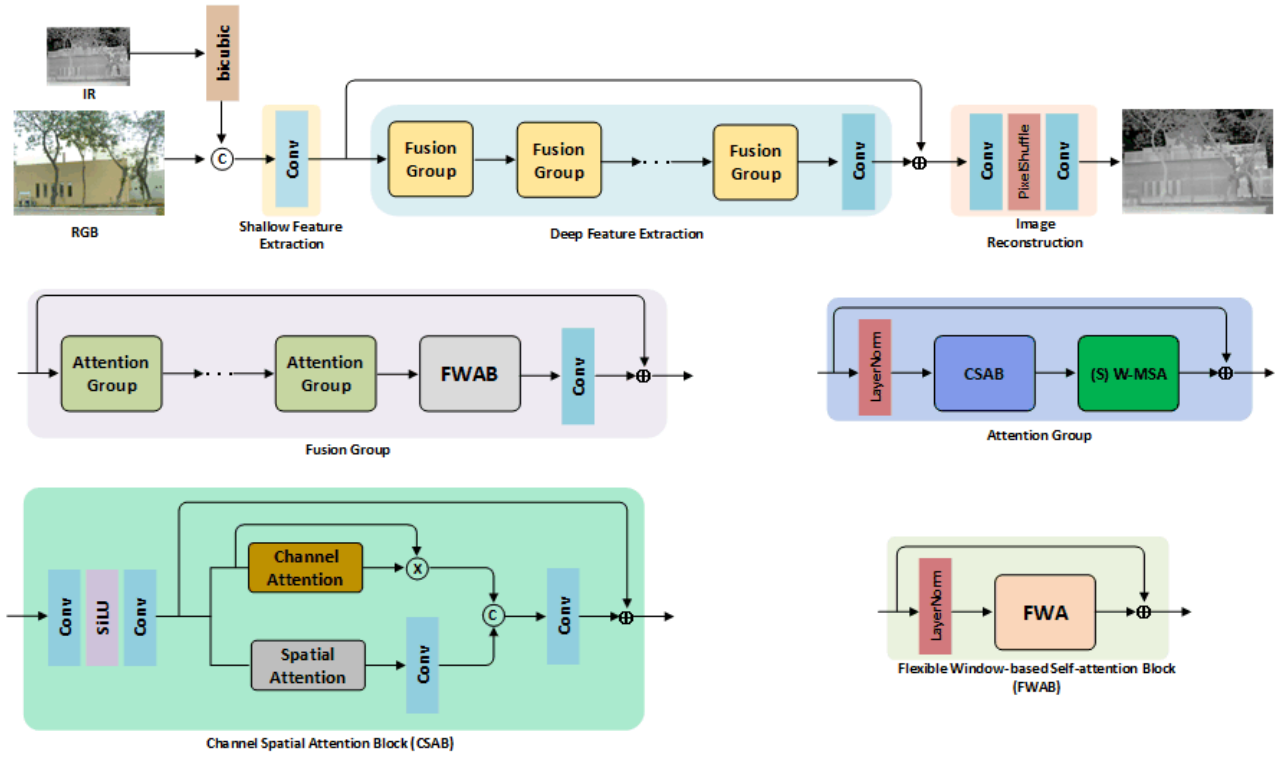

Figure 1. The overall architecture of FW-SAT and the structure of FG, AG, CSAB and FWAB

r	0	0.25	0.5	0.75
PSNR/SSIM (×8)	27.87/0.8521	28.05/0.8536	**28.56/0.8698**	27.95/0.8525
PSNR/SSIM (×16)	23.68/0.7574	23.77/0.7581	**24.05/0.7773**	23.63/0.7552

Table 1. Ablation study on the different flexible window-based self-attention (FWA) ratio with window size of 8x8 on test dataset.

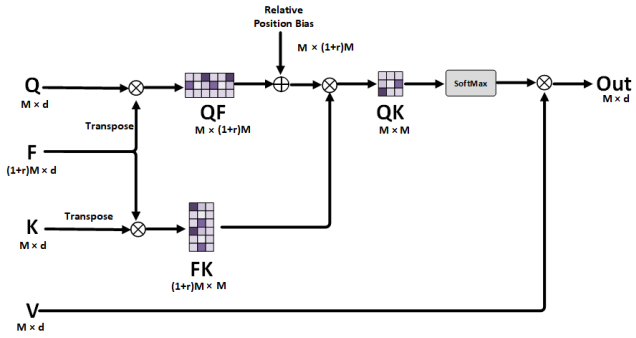

Figure 2. The structure of the FWA

Subsequently, the upsampled thermal image is concatenated with the visible image, and convolutional layers (Conv) are employed to extract shallow features. To enable deep feature extraction, a deep network is constructed by stacking several Fusion Groups (FG) with a Conv layer at the end. This architecture facilitates the bypassing of abundant low-frequency information through multiple skip connections, guiding the main network's attention towards learning high-frequency information. Each Fusion Group (FG) comprises several Attention Groups (AG), followed by a combination of Flexible Window-based Self-attention (FWA) and a Conv layer. Finally, the image is reconstructed using a Conv layer, followed by pixel-shuffle layers [15], and another Conv layer. We utilize a sophisticated fusion of L1 loss, SIMM loss, and Perceptual loss to intricately optimize the network parameters throughout the training process. It is important to note that, unlike the HAT model [2], which performs upsampling in the image reconstruction module, in our approach, we reach the upsampled image size before the shallow feature extraction module.

	EDSR	SwinIR	HAN	GRL	FW-SAT
PSNR/SSIM (×8)	25.66/0.8394	24.98/0.8170	25.86/0.8430	25.59/0.8405	**27.80/0.8815**
PSNR/SSIM (×16)	22.59/0.7562	21.22/0.7277	22.69/0.7591	22.38/0.75	**24.61/0.8116**

Table 2. Quantitative comparison with state-of-the-art methods on validation dataset.

Size	(8 × 8)	(16 × 16)
PSNR/SSIM (×8)	**28.56/0.8698**	27.93/0.8518
PSNR/SSIM (×16)	**24.05/0.7773**	23.97/0.7706

Table 3. Ablation study on the different window sizes with a flexible window-based self-attention (FWA) ratio of 0.5 on the test dataset

3.2.2 Attention Group (AG)

Previous studies have emphasized the benefits of convolution in enhancing the visual representation and optimization capabilities of Transformers [2, 25, 26, 31]. To capture global information effectively, we introduce a Channel Spatial Attention Block (CSAB) into the standard Transformer block. As illustrated in Fig. 1, the CSAB is seamlessly integrated into the standard Swin Transformer block after the first LayerNorm (LN) layer, positioned before the Shifted Window-based Self-attention ((S)W-MSA) module. The process of AG is computed as follows:

$$Y = \text{(S)W-MSA}(\text{CSAB}(\text{LN}(X))) + X; \quad (1)$$

Where X denotes a given input feature and Y represents the output of HAB.

For the (S)W-MSA, given an input feature of size $H \times W \times C$, it is initially divided into $\frac{HW}{M^2}$ local windows of size $M \times M$. Within each local window, self-attention is computed independently. Let $X_W \in \mathbb{R}^{M^2 \times C}$ denote the feature matrix within a local window. The query, key, and value matrices, denoted as Q, K, and V respectively, are obtained through linear mappings. The window-based self-attention is then expressed as follows:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V, \quad (2)$$

where d denotes the dimension of the query and key matrices. The term B represents the relative position encoding, computed according to the method described in reference [22].

To effectively explore global information, a Channel Spatial Attention Block (CSAB) is devised, comprising two standard convolution layers with SiLU activation, along with a Channel Attention (CA) module [29] and Spatial Attention (SA) module [1], as depicted in Fig. 1. Channel attention enhances the model's ability to emphasize crucial features across different channels, enabling a focused attention on the most relevant information. Concurrently, spatial attention aids in capturing spatial relationships and contextual details within features, thereby enhancing the model's understanding and refining its performance. To leverage the benefits of both modules, we combine them through concatenation, harnessing the strengths of both channel attention and spatial attention mechanisms.

3.2.3 Flexible Window-based Self-attention Block (FWAB)

To explore regional features, we introduce an additional concept called a "flexible pointer" alongside the triplets of queries, keys, and values. The flexible pointer set serves as a condensed summary of information from the image feature map and possesses a high dimensionality. Rather than directly comparing similarities between queries and keys, the flexible pointer acts as an intermediary for this comparison. The window size for the flexible pointer is calculated as follows:

$$M_o = (1 + \gamma) \times M, \quad (3)$$

where γ represents a positive constant used to control the window size of the flexible pointer, ensuring that the dimensionality of M_o is much bigger compared to the original dimensionality M. To clarify this operation further, consider the standard window partition as a sliding partition with both the kernel size and the stride set to the window size M. Conversely, the flexible pointer window partition can be conceptualized as a sliding partition with the kernel size equal to M_o, while the stride remains equal to M. Zero-padding with a size of $\frac{\gamma M}{2}$ is employed to maintain the consistency of flexible pointer windows.

As illustrated in Fig. 2, the Flexible Window-based Self-attention (FWA) mechanism is proposed, as outlined in the following equation:

$$\text{M}_f = \text{SoftMax}\left[\left(QF^T/\sqrt{d} + B\right)\left(FK^T/\sqrt{d}\right)\right]V \quad (4)$$

Where $F \in \mathbb{R}^{M_o \times d}$ represents the flexible pointer, while $Q, K, V \in \mathbb{R}^{M \times d}$ denote the query, key, and value matrices respectively, where d represents the embedding dimension. Additionally, the relative position bias $B \in \mathbb{R}^{M \times M_o}$ is employed to incorporate positional information into the attention mechanism.

	Baseline			
CSAB	×	×	✓	✓
FWAB	×	✓	×	✓
PSNR/SSIM (×8)	27.45/0.8418	27.65/0.8448	27.86/0.8490	**28.56/0.8698**
PSNR/SSIM (×16)	23.43/0.7495	23.46/0.7529	23.53/0.7561	**24.05/0.7773**

Table 4. Ablation study on the proposed channel spatial attention block (CSAB) and flexible window-based self-attention block (FWAB) on test dataset.

Structure	w/o CA	w/o SA	w/ CA & SA
PSNR/SSIM (×8)	27.82/0.8509	27.74/0.8510	**28.56/0.8698**
PSNR/SSIM (×16)	23.59/0.7555	23.72/0.7578	**24.05/0.7773**

Table 5. Ablation study on the channel attention(CA) and spatial attention(SA) modules in channel spatial attention block (CSAB) on test dataset.

4. Experiments

4.1. Dataset

The PBVS-2024 Thermal Image Super-Resolution Challenge (GTISR) - Track2 dataset utilized for this Thermal Image Super-Resolution (TISR) challenge consists of registered pairs of images captured in both the visible and thermal spectra, ensuring accurate alignment of scenes across modalities. The dataset comprises a total of 1000 registered images, simultaneously captured using both the Balster and TAU2 cameras. Participants have access to 900 images, with 700 designated for training and 200 for validation. The remaining 100 images serve as the test dataset for evaluating outcomes, with ground truth data withheld for assessment.

4.2. Experimental Setup

We utilize the PBVS-2024 Thermal Image Super-Resolution Challenge (GTISR) - Track2 dataset as our training dataset. During training, we utilize the validation dataset and incorporate data augmentation techniques by randomly rotating or flipping the images. Specifically, we configure the number of Fusion Groups (FG) and Attention Groups (AG) to 6 each. The channel number is set to 96 to facilitate effective feature processing. Additionally, both the attention head number and window size are defined as 6 and 8, respectively, for both the (S)W-MSA and Flexible Window-based Self-attention (FWA) modules. Regarding the hyperparameters of our proposed modules, we set the squeeze factor in channel attention to 3 and the FWA ratio to 0.5.

4.3. Ablation Study

4.3.1 Comparison of HAT and FW-SAT

The HAT model, as introduced by Chen et al. [2], integrates self-attention, channel attention, and a unique overlapping cross-attention mechanism. This integration is achieved through the use of residual Hybrid Attention Groups (RHAG), which comprise Hybrid Attention Blocks (HAB) and Overlapping Cross-attention Blocks (OCAB). By activating more pixels, this method aids in reconstruction and differs from prior approaches. Notably, HAT utilizes same-task pre-training on large-scale datasets, demonstrating the effectiveness of this strategy. Furthermore, HAT scales up the model, setting new state-of-the-art benchmarks for single-image super-resolution tasks.

In the case of our Flexible Window-based Attention Transformer (FW-SAT), we made modifications by removing the multi-layer perceptron (MLP) and CAB module. Instead, we incorporated the Channel and Spatial Attention Blocks (CSAB) before the (S)W-MSA module. This design decision stems from our aim to enable the network to learn global information through CSAB, which contains both Channel Attention and Spatial Attention, and then transfer these features to the (S)W-MSA module to explore local features. Finally, the Flexible Window-based Self-attention Block (FWAB) is employed to capture regional features. This hierarchical learning approach allows our network to optimize performance by learning local, regional, and global features.

Additionally, we combined the visible image and downsampled thermal images before the shallow feature extraction stage. This enables our network to learn RGB information, in contrast to the HAT network, which only takes downsampled images as input. To maintain a consistent comparison standard, we kept the input of the shallow feature extraction module of FW-SAT the same as HAT. Therefore, we set the upsampling ratio of the image reconstruction to 1 and maintained the same network configuration for both HAT and FW-SAT. Our experimental results in Tab. 6 demonstrate that FW-SAT achieves better performance compared to HAT.

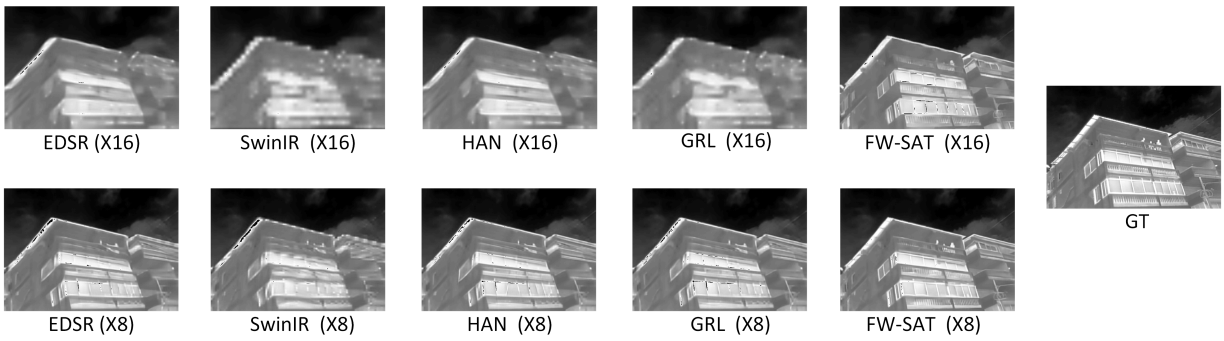

Figure 3. Visual comparison on validation dataset

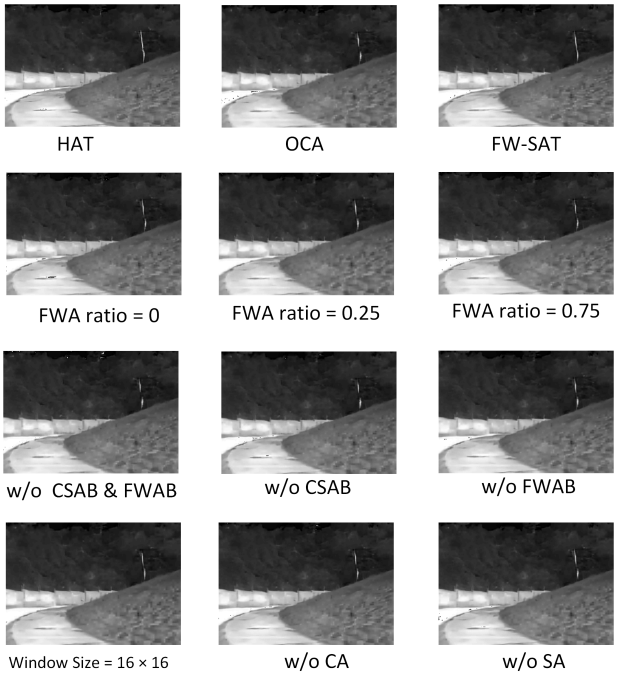

Figure 4. The visual comparison x8 upscaled images on the test dataset

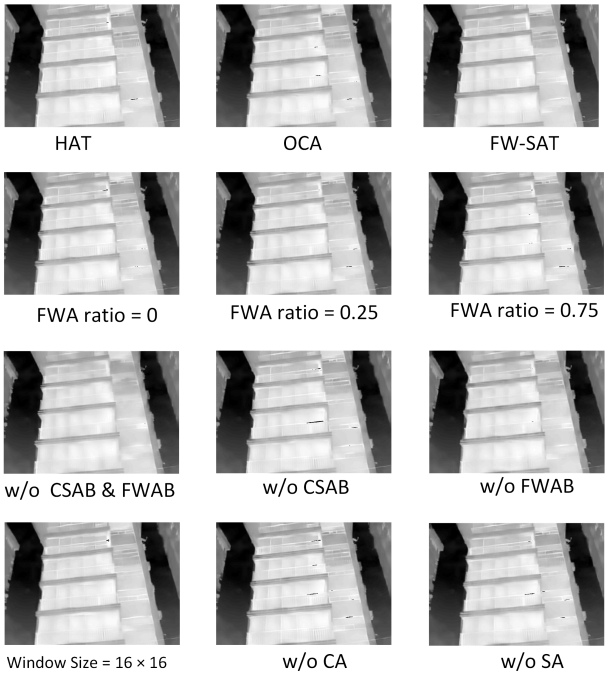

Figure 5. The visual comparison x16 upscaled images on the test dataset

Structure	FW-SAT	HAT
PSNR/SSIM ($\times$8)	**28.56/0.8698**	28.48/0.8688
PSNR/SSIM ($\times$16)	**24.05/0.7773**	24.01/0.7705

Table 6. Ablation study on the flexible window-based self-attention transformer (FW-SAT) and hybrid attention transformer (HAT) on test dataset.

4.3.2 Comparison of FWA and OCA

To compare the two modules, we directly replaced the Overlapping Cross-Attention (OCA) with Flexible Window-based Self-attention (FWA) in FW-SAT. The primary difference between FWA and OCA lies in their approach to achieving overlapping windows. While OCA adjusts the size of keys and values to achieve this purpose, we designed a flexible pointer to dynamically adjust the key and value sizes based on the idea proposed in [8].

Our experimental results in Tab. 7 demonstrate that FWA is significantly more effective than OCA. This indicates that our approach to handling overlapping windows through flexible adjustment of key and value sizes yields superior performance compared to the conventional approach employed by OCA.

Structure	FWA	OCA
PSNR/SSIM (×8)	**28.56/0.8698**	27.67/0.8467
PSNR/SSIM (×16)	**24.05/0.7773**	23.60/0.7547

Table 7. Ablation study on the flexible window-based self-attention (FWA) and overlapping cross-attention (OCA) of hybrid attention transformer (HAT) on test dataset.

4.3.3 Effectiveness of CA and SA in CSAB

Channel attention (CA) and spatial attention (SA) are two complementary techniques utilized to enhance model performance. CA enables models to prioritize critical features across different channels, allowing for focused attention on the most relevant information. Meanwhile, SA captures spatial relationships and contextual details within features, improving the model's understanding and performance. As shown in Tab. 5, comparative analyses against baseline results reveal that integrating both CA and SA led to notable enhancements in performance metrics such as PSNR and Structural SSIM, with improvements exceeding 0.3 dB and 0.01,

4.3.4 Effectiveness of CSAB and FWAB

The Channel Spatial Attention Block (CSAB) and Flexible Window-based Self-attention Block (FWAB) are implemented to investigate global and residual features, respectively. Experimental assessments were conducted to highlight the efficacy of these proposed blocks. Comparative analyses with baseline results in Tab. 4 demonstrate that both CSAB and FWAB contributed to performance improvements of over 0.5 dB on PSNR and 0.02 on SSIM.

4.3.5 Comparison of Different Window Size

The design of the (S)W-MSA and FWA modules serves the purpose of exploring both local and regional features within the input features. Our experimentation reveals that overly large window sizes introduce computational complexities that hinder the network's capacity to meticulously examine window features. Based on our observations in Tab. 3, a window size of 8x8 emerges as the optimal choice, striking the right balance between computational efficiency and performance.

4.3.6 Comparison of Different Ratio of FWA

In the FWA, a constant parameter γ is introduced to regulate the window size for the flexible pointer. To investigate the impact of different ratios, a range of γ values from 0 to 0.75 were examined, with $\gamma = 0$ representing a standard Transformer block. The analysis, detailed in Tab. 1 , indicates that the model achieves optimal performance when γ is set to 0.5. Conversely, setting γ to 0.75 results in either minimal performance gains or even a decrease in performance, highlighting the importance of selecting an appropriate window size of the flexible pointer to facilitate effective interaction among neighboring windows.

4.4. Comparison with State-of-the-Art Methods

In a meticulous evaluation of FW-SAT's performance, we rigorously compared it with state-of-the-art methods by training them on the train datasets. Utilizing the validation data for evaluation was crucial, as the dataset of 200 images offered ample opportunity to assess the performance of each network comprehensively. Notably, since these networks are optimized for processing super-resolution of downsampled visible images, we maintained consistency by utilizing downsampled thermal images as inputs. Fig. 3 and Tab. 2 present the qualitative and quantitative comparisons, respectively. The notable consistency in performance across these methods highlights the effectiveness of FW-SAT in skillfully addressing the super-resolution task within the domain of thermal imaging.

5. Conclusion

The paper presents a novel Flexible Window-based Self-attention Transformer (FW-SAT) designed specifically for thermal image super-resolution. The FW-SAT architecture integrates channel and spatial attention, window-based self-attention, and flexible window-based self-attention mechanisms. Particularly, the Channel Spatial Attention Block (CSAB) is introduced to exploit global features by combining the strengths of channel attention and spatial attention through concatenation. Additionally, the Flexible Window-based Self-attention (FWA) is developed to effectively explore regional features by leveraging window-based self-attention. Extensive experimental evaluations confirm the effectiveness of the proposed network, showcasing its state-of-the-art performance both quantitatively and qualitatively.

While focusing the GTISR dataset in this work, the proposed technique can be extended to perform SR for multi-spectral remote sensing images (e.g., Landsat or Sentinel-2) which feature high- and low-resolution at different bandwidths. By performing SR over these low-resolution bands, they can effectively enable advanced image understanding (e.g., semantic mapping for global hazards and impact) using multiple bands with unified resolution. This capability is subject to our future effort.

6. Acknowledgement

The second author thanks to the support from the University of Alabama in Huntsville under a contract with the National Aeronautics and Space Administration (80NSSC22K0014).

References

[1] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017. 5

[2] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 2, 3, 4, 5, 6

[3] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12312–12321, 2023. 3

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 2

[5] Arnold C Goldberg, Theodore Fischer, and Zenon I Derzko. Application of dual-band infrared focal plane arrays to tactical and strategic military problems. In *Infrared Technology and Applications XXVIII*, pages 500–514. SPIE, 2003. 1

[6] Aditya Kasliwal, Pratinav Seth, Sriya Rallabandi, and Sanchit Singhal. Corefusion: Contrastive regularized fusion for guided thermal super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 507–514, 2023. 3

[7] Ao Li, Le Zhang, Yun Liu, and Ce Zhu. Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12514–12524, 2023. 2

[8] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289, 2023. 2, 3, 7

[9] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 3

[10] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2

[11] Jia-Wei Lin, Ming-Hung Lu, and Yuan-Hsiang Lin. A thermal camera based continuous body temperature measurement system. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[12] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 191–207. Springer, 2020. 2

[13] Kalpesh Prajapati, Vishal Chudasama, Heena Patel, Anjali Sarvaiya, Kishor P Upla, Kiran Raja, Raghavendra Ramachandra, and Christoph Busch. Channel split convolutional neural network (chasnet) for thermal image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4368–4377, 2021. 3

[14] Hairong Qi and Nicholas A Diakides. Thermal infrared imaging in early breast cancer detection-a survey of recent research. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)*, pages 1109–1112. IEEE, 2003. 1

[15] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in neural information processing systems*, 32, 2019. 4

[16] Rafael E Rivadeneira, Angel D Sappa, and Boris Xavier Vintimilla. Thermal image super-resolution: A novel architecture and dataset. In *VISIGRAPP (4: VISAPP)*, pages 111–119, 2020. 3

[17] Antoni Rogalski, Piotr Martyniuk, and Małgorzata Kopytko. Challenges of small-pixel infrared detectors: a review. *Reports on Progress in Physics*, 79(4):046501, 2016. 1

[18] Samuel Schulter, Christian Leistner, and Horst Bischof. Fast and accurate image upscaling with super-resolution forests. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3791–3799, 2015. 2

[19] Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Vegetation index estimation from monospectral images. In *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15*, pages 353–362. Springer, 2018. 1

[20] Nguyen Duc Thuan, Trinh Phuong Dong, Bui Quang Manh, Hoang Anh Thai, Tran Quang Trung, and Hoang Si Hong. Edge-focus thermal image super-resolution using generative adversarial network. In *2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6. IEEE, 2022. 3

[21] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 1920–1927, 2013. 2

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[23] Kai Wang, Qigong Sun, Yicheng Wang, Huiyuan Wei, Chonghua Lv, Xiaolin Tian, and Xu Liu. Cippsrnet: A camera internal parameters perception network based contrastive learning for thermal image super-resolution. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 342–349, 2022. 3

[24] Xuehui Wang, Qing Wang, Yuzhi Zhao, Junchi Yan, Lei Fan, and Long Chen. Lightweight single-image super-resolution network with attentive auxiliary feature learning. In *Proceedings of the Asian conference on computer vision*, 2020. 2

[25] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021. 5

[26] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021. 5

[27] Jinsu Yoo, Taehoon Kim, Sihaeng Lee, Seung Hwan Kim, Honglak Lee, and Tae Hyun Kim. Enriched cnn-transformer feature aggregation networks for super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4956–4965, 2023. 2

[28] Mingjin Zhang, Chi Zhang, Qiming Zhang, Jie Guo, Xinbo Gao, and Jing Zhang. Essaformer: Efficient transformer for hyperspectral image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23073–23084, 2023. 2

[29] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 2, 5

[30] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 2

[31] Yucheng Zhao, Guangting Wang, Chuanxin Tang, Chong Luo, Wenjun Zeng, and Zheng-Jun Zha. A battle of network structures: An empirical study of cnn, transformer, and mlp. *arXiv preprint arXiv:2108.13002*, 2021. 5