Multi-Scale Feature Fusion using Channel Transformers for Guided Thermal Image Super Resolution

Raghunath Sai Puttagunta [1], Birendra Kathariya [1], Zhu Li [1], George York [2]
[1] University of Missouri-Kansas City
[2] US Air Force Academy
rpyc8@umsystem.edu, bkkvh8@umsystem.edu, lizhu@umkc.edu, george.york@afacademy.af.edu

Abstract

Thermal imaging, leveraging the infrared spectrum, offers a compelling alternative to visible spectrum (VIS) imagery in challenging environmental conditions like low-light, occlusions, and adverse weather. However, its widespread adoption in computer vision tasks is hampered by lower spatial resolution. We address this challenge by proposing a novel framework titled Multi-Scale Feature Fusion using Channel Transformers (MSFFCT) for Guided Thermal Image Super-Resolution (GTISR).

GTISR tackles the resolution limitations of thermal imagery. It leverages high-resolution RGB information as a guide to reconstruct high-resolution thermal imagery from low-resolution thermal inputs. At the core of MSFFCT lies a novel deep learning architecture that combines the strengths of two powerful approaches: channel-based transformers and multi-scale fusion.

MSFFCT overcomes inherent limitations of Convolutional Neural Networks (CNNs) typically used in super-resolution tasks. CNNs often suffer from restricted receptive fields, limiting their ability to capture long-range dependencies within the image. Additionally, computational cost grows significantly with larger inputs. MSFFCT addresses these shortcomings by enabling efficient processing of global information and offering superior scalability. MSFFCT achieved state-of-the-art results on the ×8 and ×16 GTISR tasks of the 2024 Perception Beyond Visual Spectrum (PBVS) challenge, winning 2nd place in both tasks and demonstrating its effectiveness in real-world scenarios.

1. Introduction

Computer vision has become an indispensable technology across diverse applications, from self-driving cars and robotics to medical imaging and security systems. At the forefront of this revolution lie RGB cameras, capturing rich visual information in the visible spectrum. However, their reliance on illumination conditions presents a fundamental limitation. Low-light scenarios can dramatically reduce image clarity, while occlusions (objects blocking the view) and adverse weather (rain, fog) further hinder accurate image analysis. For instance, blurry details in low-light and weather conditions images can lead to misidentification of objects in autonomous vehicles [33],

To overcome these limitations and expand the reach of computer vision, researchers have explored alternative imaging modalities. While active sensors like near-infrared or depth cameras address some limitations, passive sensors offer distinct advantages. Thermal infrared imaging stands out as a versatile modality, capturing mid-to-longwave radiation emitted as heat from all objects [14]. This unique characteristic allows thermal cameras to ”see” in complete darkness, penetrate obscurants like smoke or fog, and detect inherent thermal signatures that are invisible to RGB cameras. Even state-of-the-art computer vision algorithms struggle with object recognition in unconstrained environments with weather variations, shadows, and background clutter. In contrast, thermal sensors leverage robust thermal cues to facilitate accurate perception under these challenging real-world conditions.

The advantages of thermal imaging have fostered its growing adoption in diverse real-world applications, including agriculture [20], autonomous driving [11, 30], medical imaging [18, 40] , military applications [16], pedestrian detection [25], and surveillance systems [21]. However, despite these advantages, a key challenge remains: the resolution of thermal sensors is typically lower than that of RGB cameras. This limitation can hinder the ability to discern fine details crucial for accurate image interpretation. While higher resolution thermal sensors are available, their cost ranges anywhere between $200 to $20,000 [34] which significantly limits widespread adoption. Super-resolution (SR) emerges as a promising computer vision technique that

This work is supported in part by National Science Foundation Award 2148382.

addresses the challenge of limited resolution in thermal images.

SR aims to enhance the spatial resolution of an image, essentially creating a high-resolution version from a lower-resolution input. The recent surge in deep learning has made CNNs the cornerstone of many SR approaches. Dong et al. [12] pioneered the use of CNNs for SR, paving the way for a multitude of successful CNN-based SR methods [26, 28, 48, 49]. Following this success, researchers have begun to explore CNN-based SR specifically for thermal images [8, 9, 31, 32, 37, 38, 42]. However, CNNs have inherent limitations, particularly a restricted receptive field that hinders their ability to capture long-range dependencies within the image. Additionally, computational complexity increases significantly with larger input sizes.

In GTISR, a high-resolution RGB image acts as a guide for a low-resolution thermal image. By incorporating the rich details from the RGB image, GTISR can learn the fine textures and edges crucial for reconstructing a high-resolution thermal image. While recent CNN-based approaches have explored GTISR [19, 50, 52], CNN limitations motivate our exploration of transformers. These powerful deep learning architectures have achieved state-of-the-art performance in natural language processing (NLP) tasks and are increasingly being adapted for image restoration [44, 46] and super-resolution [27] tasks . Notably, transformers overcome the limitations of CNNs by offering a wider receptive field and improved computational efficiency for handling large input sizes.

Our efficient channel-based transformer, inspired by MST++ [3], employs a channel-wise self-attention to learn the interdependencies between features within the image channels. Our novel MSFFCT framework achieved state-of-the-art results on the ×8 and ×16 GTISR tasks of the 2024 Perception Beyond Visual Spectrum (PBVS) challenge [1], winning 2$^{\text{nd}}$ place in both tasks and demonstrating its effectiveness in real-world scenarios.

2. Related Work

Deep learning techniques have revolutionized image restoration and SR, enabling the creation of high-resolution images from their lower-resolution counterparts. This section delves into relevant research areas that inform our work on GTISR. We begin by reviewing advancements in visible image SR and restoration techniques, which have laid the groundwork for applying deep learning to image enhancement tasks. Next, we focus on thermal image SR, exploring how researchers have addressed the inherent challenge of lower resolution in thermal sensors compared to their visible counterparts. Finally, we examine existing GTISR approaches, along with their limitations.

2.1. Visible Image Super Resolution and Restoration

The emergence of deep learning, particularly CNNs, revolutionized image restoration and SR. The pioneering Super-Resolution Convolutional Neural Network (SRCNN) by Dong et al. [12] demonstrated the remarkable capability of CNNs to learn the complex mapping between Low-Resolution (LR) and High-Resolution (HR) images, significantly surpassing traditional methods. Subsequent research focused on improving CNN architectures for SR, exploring strategies like increasing network depth (VDSR [26]) and incorporating residual connections (EDSR [28]). These advancements significantly enhanced SR performance. However, CNNs have inherent limitations, such as restricted receptive fields, which can hinder their ability to capture long-range dependencies within images.

Fueled by their success in NLP, transformers have made significant inroads into computer vision. Vision Transformer (ViT) by Dosovitskiy et al. [13] pioneered their effectiveness for image classification. Subsequently, transformers were adapted for various image restoration tasks, including denoising, deblurring, and super-resolution. For instance, Image Processing Transformer (IPT) by Chen et al. [4] employed a ViT-based approach, while U-Former by Wang et al. [44] and Restormer by Zamir et al. [46] utilized different self-attention mechanisms (window-based and channel-based, respectively). Notably, SwinIR by Liang et al. [27] leveraged the Swin Transformer architecture [29] with a shifted-window attention mechanism for image super-resolution.

2.2. Thermal Image Super Resolution

The success of deep learning models for SR has motivated researchers to explore their application in thermal image enhancement. Inspired by the pioneering SRCNN model [12], Choi et al. [7] proposed the Thermal Image Enhancement (TEN) network for thermal image SR. However, due to the limited availability of large-scale thermal image datasets, they resorted to using RGB images for training. Rivadeneria et al. [34] proposed a thermal SR network utilizing deep convolution layers with residual and dense connections. Rivadeneria et al. [35] also explored a CycleGAN-based model for thermal SR. Chudasama et al. [9] presented TherISURNet, a residual block-based progressive upscale strategy that emerged as the winner of the evaluation 1 of 2020 PBVS CVPR challenge [36].

Priya et al. [22] introduced a multi-level architecture with residual blocks for thermal SR, incorporating multi-level supervision with feature concatenation and an attention block inspired by [45]. This work emphasizes the importance of attention mechanisms for focusing on relevant features during reconstruction. Building upon Priya et al.'s [22] work, Nathan et al. [31] presented a multi-scale, multi-supervision architecture, utilizing a Res2Net [15] backbone instead of residual blocks for improved performance. Prajapati et al. [32] proposed ChasNet, featuring a channel splitting block with residual blocks and convolution layers with dense connections, aiming to preserve high-frequency details crucial for thermal image fidelity.

2.3. Guided Thermal Image Super Resolution

Despite advancements in thermal image SR, GTISR poses a new set of challenges. Early works in GTISR leveraged Generative Adversarial Networks (GANs) to guide the SR process. Almasri et al. [2] proposed a GAN-based model where features extracted from RGB images guided the thermal image super-resolution. Addressing the misalignment challenge, Gupta et al.[17] introduced an unaligned guided thermal SR method. Their approach utilizes two models: one to reduce misalignment in the feature space and another to estimate a misalignment map between the input thermal image and the guiding image. This work highlights the importance of handling misalignment for effective GTISR.

The winning solution for the GTISR task of the 2023 PBVS challenge [39] concatenates features from the RGB image and the low-resolution thermal image after a shallow feature extraction stage. These concatenated features are then processed through multiple NAF Blocks [5], which form the core of the network. Kasliwal et al.[23] proposed an encoder-decoder architecture for GTISR. Their work encodes both the low-resolution thermal image and the high-resolution RGB image, then combines the encoded features using a max operation before feeding them into the decoder to learn the high-resolution thermal image. Additionally, they introduced a contrastive loss function that acts as a regularizer. Suarez et al. [41] proposed a novel approach that involves creating a synthetic thermal image using a CycleGAN architecture [53]. This synthetic thermal image is then used as guidance for the low-resolution thermal image SR process. This work explored the potential of generative models for creating informative guidance suitable for GTISR tasks.

3. Proposed Method

3.1. Network Architecture

The MSFFCT architecture shown in Figure 1 receives two input images: a high-resolution RGB image $I_{rgb} \in \mathbb{R}^{H\times W\times 3}$ and a low-resolution thermal image $I_{lrth} \in \mathbb{R}^{h\times w\times 1}$. The resolution of the high-resolution RGB image I_{rgb} is either a factor of $\times 8$ or $\times 16$ larger than that of the low-resolution thermal image I_{lrth}. To address the resolution disparity, MSFFCT commences by applying bicubic upsampling to the low-resolution thermal image I_{lrth}. This creates an upsampled thermal image I_{upth} with the same resolution as the high-resolution RGB image I_{rgb}. The high-resolution RGB image I_{rgb} and the upsampled thermal image I_{upth} are then concatenated. The concatenated image is notated as I_{cat}.

$$I_{upth} = bicubic(I_{lrth}) \quad I_{cat} = concat(I_{rgb}, I_{upth}) \tag{1}$$

The concatenated images I_{cat} are downsampled using $\times 2$ and $\times 4$ pixel unshuffling. Downsampling reduces computational complexity while simultaneously capturing multi-scale features within the data. The $\times 2$ downsampled image is notated as I_{2dn} and the $\times 4$ downsampled image is notated as I_{4dn}. The shapes of I_{2dn} and I_{4dn} are $I_{2dn} \in \mathbb{R}^{H/2\times W/2\times 16}$ and $I_{4dn} \in \mathbb{R}^{H/4\times W/4\times 64}$, respectively.

The downsampled images I_{2dn} and I_{4dn} are fed into the core network, which was inspired by TSFNet [24]. The core network comprises three key components: a shallow feature extractor, a fusion block with a channel-wise transformer, and a reconstruction block.

3.1.1 Shallow Feature Extractor

Multi-scale features are extracted from I_{2dn} and I_{4dn} using a two-stream architecture. I_{2dn} is inputted into one stream and I_{4dn} is inputted into a second steam. In parallel, each stream employs a sequence of two deformable convolutions [10] with a kernel size of 3×3. An activation layer utilizing the Parametric ReLU (PReLU) function is used in between the deformable convolutions. Together, these two parallel processes are referred to as a shallow feature extractor.

$$F_{2x} = H_{DC2}(I_{2dn}) \quad \text{and} \quad F_{4x} = H_{DC4}(I_{4dn}) \tag{2}$$

H_{DC2} and H_{DC4} represent the parallel processes within the shallow feature extractor for I_{2dn} and I_{2dn}, respectively. $F_{2x} \in \mathbb{R}^{H/2\times W/2\times C_{out}}$ and $F_{4x} \in \mathbb{R}^{H/4\times W/4\times C_{out}}$ represent the $\times 2$ and $\times 4$ downsampled features from the deformable convolution. c_{out} represents the feature channels of the deformable convolution.

Deformable convolutions are specifically chosen in the shallow feature extractor due to their ability to handle potential misalignments between the concatenated RGB and thermal features, which can arise from sensor discrepancies or variations in object pose.

3.1.2 Fusion Block

The next key component of our proposed network architecture is a series of N fusion blocks. Each fusion block comprises three key components: two parallel residual blocks, one transposed convolution, and one channel-wise transformer.

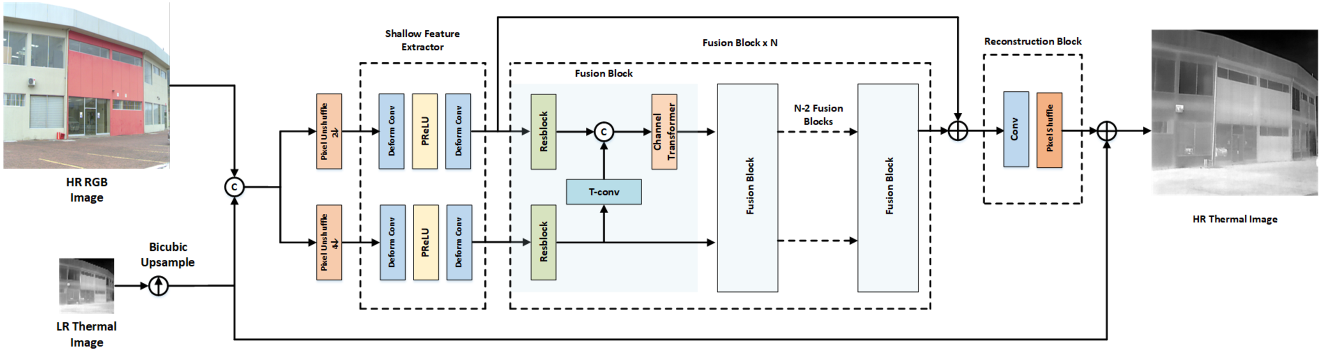

Figure 1. The MSFFCT Architecture

The inputs to the first fusion block are the outputs F_{2x} and F_{4x} from the shallow feature extractor. F_{2x} and F_{4x} are inputted into two parallel residual blocks.

Each residual block has an efficient channel attention inspired by [43]. This approach assigns importance to different feature channels through a global weighting scheme, allowing the model to focus on informative features crucial for reconstruction. The effectiveness of channel attention for transformer models in achieving superior visual representations has been demonstrated by Chen et al. [6]. By leveraging this strategy within our residual blocks, we aim to enhance the feature extraction capabilities of MSFFCT.

$$F_{2x} = H_{Res2}(F_{2x}) \quad \text{and} \quad F_{4x} = H_{Res4}(I_{4x}) \tag{3}$$

H_{Res2} and H_{Res4} represent the residual blocks with enhanced channel attention. The ×4 downsampled feature F_{4x} is now upsampled with a transposed convolution to match the spatial resolution of F_{2x}.

$$F_{4xup} = H_{tconv}(F_{4x}) \tag{4}$$

H_{tconv} represents the transposed convolution. F_{4xup} now has the same spatial resolution as F_{2x}.

$$F_{2x} = concat(F_{2x}, F_{4x}) \tag{5}$$

The F_{4xup} and F_{2x} features are concatenated with each other and passed to the channel transformer.

3.1.3 Channel Transformer

The proposed channel transformer draws inspiration from the channel-wise multi-head self-attention concept introduced in [3]. In the realm of transformer-based models, the computation of self-attention is typically performed on tokens. Notably, models such as Vision Transformer (ViT) [13] employ global attention, where each pixel in the feature map is treated as a token. For a feature map $F \in \mathbb{R}^{H \times W \times C}$, the time complexity of global self-attention is $\mathcal{O}(H^2W^2C)$. In contrast, other transformer-based architectures, including UFormer [44] and SwinIR [27], adopt window-based or shifted-window-based self-attention approaches for token generation. Rather than treating each pixel as a token, these models partition the feature map into non-overlapping $M \times M$ windows, where each pixel within a window is considered a token. This strategy reduces the time complexity from $\mathcal{O}(H^2W^2C)$ to $\mathcal{O}(M^2HWC)$. In our work, we deviate from these approaches by treating each channel in the feature map as a token. This entails computing self-attention at the channel level, circumventing the complexity associated with spatial dimensions. This approach demonstrates computational efficiency, as it focuses on feature channels rather than spatial dimensions, reducing the time complexity of global self-attention from $\mathcal{O}(H^2W^2C)$ to $\mathcal{O}(HWC^2)$. Contingent upon the window size and the number of channels in the feature map, window-based self-attention can achieve efficiency comparable to, or better than, channel-based self-attention. However, the receptive field size in window-based self-attention is constrained by the window size, whereas in channel-based self-attention, the receptive field size encompasses the entire spatial dimension.

The input for our channel transformer will be from the 2× downsampled branch of the multi-scale network. The input to our channel transformer is $F_{2x} \in \mathbb{R}^{H/2 \times W/2 \times 2C}$, where C represents the feature channels of the 2× branch. First, we flatten $F_{2x} \in \mathbb{R}^{H/2 \times W/2 \times 2C}$ into $X \in \mathbb{R}^{HW/4 \times 2C}$. Next, we project X into three fully connected layers to get Query Q, Key K, and Value V with $Q, K, V \in \mathbb{R}^{HW/4 \times C}$.

$$Q = XW_Q, K = XW_K, V = XW_V \tag{6}$$

W_Q, W_K, and W_V represent the weights of the fully connected layer and are learnable. The key K is transposed into $K^{\intercal}$ and multiplied with Query Q to obtain the attention matrix A. Attention matrix $A \in \mathbb{R}^{C \times C}$.

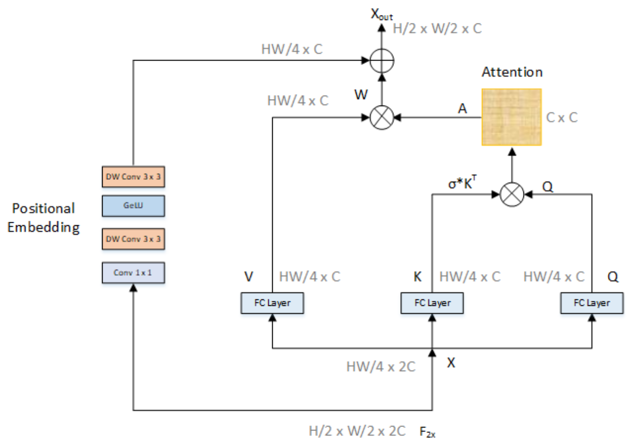

Figure 2. Channel Transformer

$$A = softmax(\sigma * K^{\mathsf{T}} * Q) \tag{7}$$

σ is a learnable parameter, which learns how much weight to give to each feature channel in the attention map. The self-attention is calculated by the following:

$$SA = V * A \tag{8}$$

To calculate multi-head self-attention, each channel is divided into k heads and then k self-attention maps are learned in parallel. The self-attention maps are then linearly projected via fully connected layers and then a positional embedding is added as a residue.

$$X_{out} = W * (SA) + PE(V) \tag{9}$$

W represents the weight of the fully connected layer and PE represents the positional embedding. The positional embedding has two 3×3 depthwise convolution layers with a GELU activation function in between them. The final feature map will have the dimensions $X_{out} \in \mathbb{R}^{H/2 \times W/2 \times C}$.

The output from the channel transformer, and the output from the residual block with input F_{4x}, are the inputs into the subsequent N - 1 fusion blocks. In the final N fusion block, the output from the channel transformer is concatenated with F_{2x}. The output from that concatenation will be the input in the next key component, the reconstruction block.

$$X_{out,i} = H_{fusion,i}(X_{out,i-1}), \quad i = 2, ..., N \tag{10}$$
$$X_{out,N} = X_{out,N} + F_{2x} \tag{11}$$

where $H_{fusion,i}$ is the i^{th} fusion block and $X_{out,i-1}$ is the output from $(i-1)^{th}$ fusion block.

3.1.4 Reconstruction Block

The output $X_{out,N}$ is passed through a convolution and then upsampled $\times 2$ using pixel shuffling. There is a residue from the original bicubic upsampled image to the pixel shuffled upsampled image to learn the reconstructed high resolution thermal image I_{rhrth}.

$$I_{rhrth} = H_{Rec}(X_{out}) + I_{upth} \tag{12}$$

H_{Rec} represents the convolution layer and the $2\times$ pixel shuffle operation. I_{upth} represents the bicubic upsampled thermal image.

3.2. Loss Function

The 2024 PBVS competition [1] ranked submissions based on the PSNR and SSIM metrics. To optimize those metrics, we used a combination of L_1 loss, SSIM loss, and perceptual loss.

3.2.1 L_1 Loss

L_1 loss measures the absolute difference between the ground truth image and the predicted image. In GTISR, the ground truth image is the given high-resolution thermal image I_{hrth} and the predicted image is the reconstructed high-resolution thermal image I_{rhrth}. L_1 loss is defined by the following equation:

$$L_1 = \frac{1}{n}\sum_{i=1}^{n} |I_{rhrth_i} - I_{hrth_i}| \tag{13}$$

3.2.2 SSIM Loss

Structural Similarity Index Measure (SSIM) calculates the similarity between two images and assigns a value between -1 and 1. A value of 1 indicates that the two images are exact matches, a value of 0 indicates no similarity, and value of -1 indicates the two images are exact inverses. The SSIM loss is calculated for a given high-resolution thermal image I_{hrth} and reconstructed high-resolution thermal image I_{rhrth}. The SSIM loss is defined by the following equation:

$$L_{SSIM} = 1 - SSIM(I_{rhth_i}, I_{hrth_i}) \tag{14}$$

3.2.3 Perceptual Loss

Perceptual loss measures the visual similarity between two images and is primarily used in GAN-based models. We calculate the mean absolute error between the VGG features of a given high-resolution thermal image I_{hrth} and reconstructed high-resolution thermal image I_{rhrth} at different layers. The perceptual loss is defined by the following equation:

Method	PSNR ×8	SSIM ×8	PSNR ×16	SSIM ×16	Params (M)	GMacs
Bicubic	25.17	0.8494	22.04	0.7901	-	-
Restromer [46]	28.72	0.8753	25.39	0.8059	15.08	83.94
AHMF [51]	28.38	0.8676	24.72	0.7790	3.36	11.75
NafNet [39]	29.16	0.8832	25.50	0.8069	116.35	86.51
MSFFCT	29.42	0.8879	25.90	0.8188	12.17	154.59

Table 1. Comparison with Different Models on PBVS 24 [1] GTISR Dataset

$$\mathcal{L}_{\text{perceptual}} = \frac{1}{N}\sum_{i=1}^{N} \|\Phi_i(I_{hrth}) - \Phi_i(I_{rhrth})\|^2 \qquad (15)$$

$\Phi_i(.)$ represents the VGG feature map at layer i for given image.

Our final loss function is a weighted average of L_1 loss, SSIM loss, and perceptual loss:

$$L_{final} = \alpha * L_1 + \beta * L_{SSIM} + \gamma * L_{perceptual} \qquad (16)$$

4. Experiments and Results

4.1. Dataset

We evaluated the performance of MSFFCT on the PBVS 24 GTISR dataset [1]. The dataset includes ×8 and ×16 downscaled low-resolution thermal images with paired high-resolution RGB images of the same scene, both in daylight conditions, to be used as a guide for GTISR. This dataset has 700 images for training, 100 images for validation, and 40 images for testing. Since the ground truth labels are not released, we report our results on the validation dataset.

4.2. Experimental Setup

During training, we randomly cropped the low-resolution thermal images to either 32 × 32 (for the ×8 GTISR task) or 16 × 16 (for the ×16 GTISR task). We trained the model for 100 epochs using a batch size of 8, Adam optimizer with default parameters, and initial learning rate of 1e-4. We gradually decreased the learning rate to 1e-6 using a cosine annealing scheduler. We augmented the data with flipping and mixup [47]. Mixup augmentaton acts as a regularizer to the network during the training process.

We achieved optimal performance with fusion blocks of size 48. We used 64 feature channels for both 2× and 4× feature branches. The weight for the loss function α was 7, β was 1, and γ was 0.15. We evaluated MSFFCT performance using PSNR and SSIM to be in alignment with the metrics used to rank submissions in the PBVS 24 GTISR tasks [1].

We implemented MSFFCT in PyTorch and trained it for over 2 days on 2 NVIDIA RTX A6000 GPUs.

4.3. Quantitative Results on Validation Dataset

We comprehensively evaluated MSFFCT against several state-of-the-art approaches, including Restromer [46], Attention-based Hierarchical Multi-modal Fusion (AHMF) [51], and the winning PBVS 23 GTISR challenge approach [39], which was based on NAFNet [5]. Restromer [46] employs a channel-based self-attention mechanism for image restoration. AHMF [51] is a state-of-the-art solution for guided depth super-resolution tasks. It is important to note that for the Restromer model to be applicable to the GTISR task, we implemented a pre-processing step involving thermal image upsampling using the corresponding RGB image. The resulting features were then concatenated and fed as input to the network.

As shown in Table 1, MSFFCT surpasses the PSNR and SSIM values of several state-of-the-art approaches. It demonstrates a significant PSNR improvement of 0.26 dB over the previous year's winner, NAFNet [39], on the ×8 GTISR task. This improvement is even more pronounced (0.4 dB) for the ×16 GTISR task. When compared to Restromer, another channel-wise self-attention model, MSFFCT achieves a PSNR gain of 0.70 dB for ×8 GTISR and 0.51 dB for ×16 GTISR. It is noteworthy that MSFFCT achieves this superior performance with a significantly lower number of trainable parameters compared to NAFNet.

Table 2 shows the effects of the size of the fusion block. We experimented with fusion block sizes of 16, 24, 32, and 48. The best performance was achieved with a fusion block size of 48. A fusion block size of 32 outperforms the NAFNet-based model [5] on both ×8 and ×16 GTISR task while having a significantly lower number of parameters and slightly more GMac operations. We can also observe that fusion block sizes of 16 and 24 outperform Restromer [46] with lower numbers of parameters and GMac operations.

4.4. Quantitative Results on Test Dataset

Table 3 presents the results for GTISR tasks at scaling factors of ×8 and ×16. We employed self ensemble learning during testing, which involved flipping the test images horizontally and vertically and then averaging the predicted images. This learning strategy increased PSNR by 0.36 dB

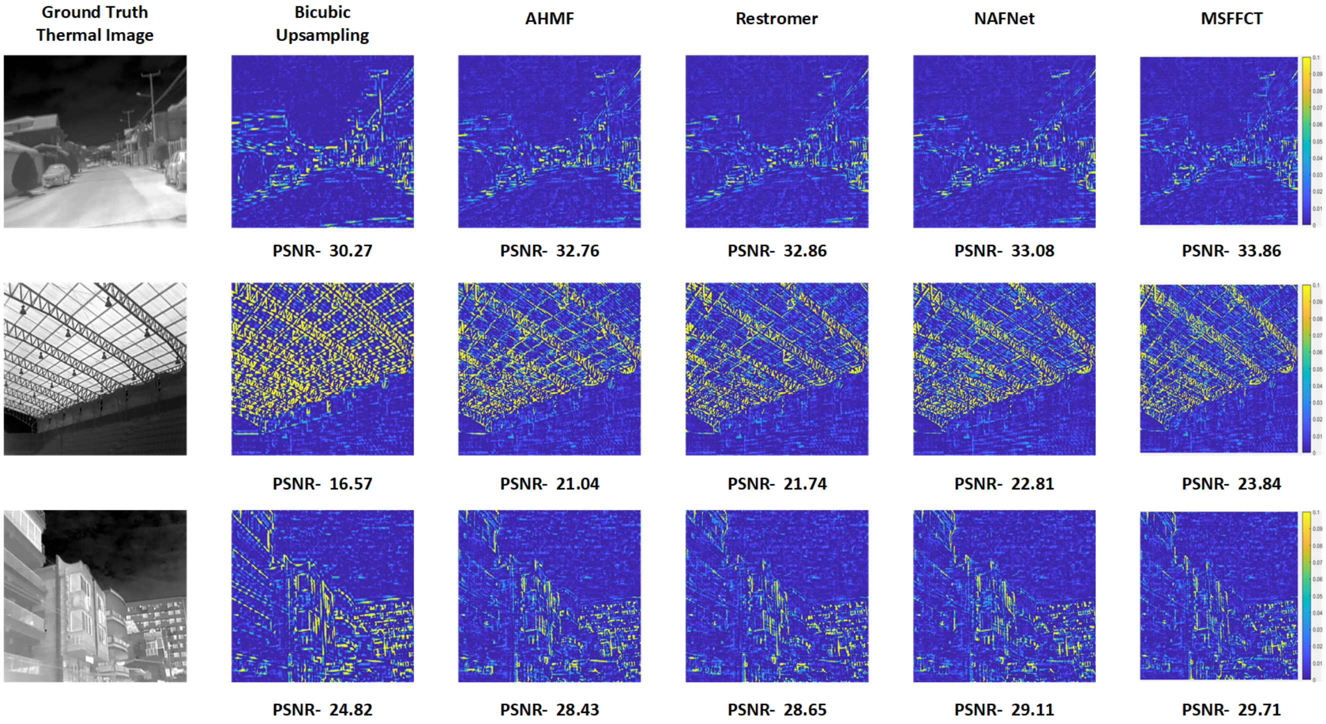

Figure 3. Qualitative Results on PBVS 24 [1] for ×8 GTISR Task

Number of Fusion Blocks	PSNR ×8	SSIM ×8	PSNR ×16	SSIM ×16	Params (M)	GMacs
16	28.87	0.8774	25.52	0.806	4.12	51.75
24	29.01	0.8795	25.66	0.8106	6.14	77.46
32	29.20	0.8839	25.70	0.8124	8.15	103.17
48	29.42	0.8879	25.90	0.8188	12.17	154.59

Table 2. Comparison of Fusion Block Sizes on PBVS 24 [1] GTISR Dataset

for the ×8 task and 0.25 dB for the ×16 task. We further enhanced performance through model ensemble learning. This strategy involved taking the weighted average of predictions from multiple models. The model ensemble learning included the following models: MSFFCT, MSFFCT without deformable convolutions, and the winning model from the PBVS 23 GTISR challenge [39]. This ensemble learning led to increasing PSNR by 0.37 dB for the ×8 task. To further improve our results on the testing dataset, we combined model ensemble and self ensemble learning. That approach increases the PSNR by 0.49 dB compared to the proposed method for the ×8 task. On the testing dataset, the model achieved PSNR values of 30.05 dB and 25.67 dB on the ×8 and ×16 GTISR tasks respectively, which placed us 2nd for PBVS 24 GTISR task [1].

4.5. Qualitative Results on Validation dataset

Figure 3 illustrates the absolute difference maps between the predicted thermal images and the ground truth images for various reconstruction methods on a validation dataset for the ×8 GTISR task. The first column shows the ground truth thermal image. The remaining columns depict the absolute difference maps for each method compared to the ground truth. In these difference maps, deeper blue regions signify better reconstruction fidelity, indicating a smaller absolute difference between the predicted and ground truth images.

Based on the qualitative comparison, MSFFCT achieves superior performance compared to other reconstruction techniques. Notably, it outperforms NAFNet and Restormer, which are both models that have channel attention in their architecture. Restromer also has channel-wise self-attention. This observation suggests that the proposed method establishes more effective feature space correlations, enabling superior reconstruction quality.

Method	PSNR ×8	SSIM ×8	PSNR ×16	SSIM ×16
MSFFCT	29.54	0.8869	25.42	0.8092
MSFFCT with Self Ensemble	29.90	0.8929	25.67	0.8167
Model Ensemble	29.91	0.8919	-	-
Self Ensemble and Model Ensemble	30.05	0.8947	-	-

Table 3. Results on PBVS 24 [1] GTISR ×8 and ×16 Test Dataset

5. Conclusion and Future Work

5.1. Conclusion

Thermal imaging, leveraging the infrared spectrum, offers a compelling alternative to VIS imagery in challenging environmental conditions like low-light, occlusions, and adverse weather. However, its widespread adoption in computer vision tasks is hampered by lower spatial resolution. GTISR tackles the resolution limitations of thermal imagery by leveraging high-resolution RGB information to reconstruct high-resolution thermal imagery from low-resolution thermal inputs.

Existing CNN-based SR methods [26, 28, 48, 49] often suffer from restricted receptive fields, limiting their ability to capture long-range dependencies within the images. To address this, we propose MSFFCT. MSFFCT is a novel GTISR architecture that combines the strengths of multi-scale fusion and channel-based transformers, inspired by [3] and [24], respectively. This combination enables effective capture of rich feature information and long-range dependencies.

MSFFCT achieved state-of-the-art results on the ×8 and ×16 GTISR tasks of the 2024 PBVS challenge [1], winning 2^{nd} place in both tasks and demonstrating its effectiveness in real-world scenarios. It demonstrates a significant PSNR improvement of 0.26 dB over the previous year's winner, NAFNet [39], on the ×8 GTISR task. This improvement is even more pronounced (0.4 dB) for the ×16 GTISR task. Specifically, MSFFCT achieved a PSNR of 30.05 dB and SSIM of 0.8947 for the ×8 GTISR task test dataset and a PSNR of 25.67 dB and SSIM of 0.8167 for ×16 GTISR task test dataset. MSFFCT also outperforms other state-of-the-art benchmarks on the 2024 PBVS challenge dataset [1] for both ×8 and ×16 downscaling factors.

5.2. Future Work

While MSFFCT demonstrates promising results under the assumption of near perfect alignment between thermal and RGB images, real-world scenarios often present misalignment challenges. To address this, future work will explore incorporating loss functions that promote feature-space alignment, similar to the approach proposed by Gupta et al. [17]. Additionally, we will investigate the potential of incorporating a synthetic thermal image generation module inspired by Suarez et al. [41] within the GTISR pipeline. This approach has the potential to further enhance the effectiveness of our framework by providing informative synthetic guidance for scenarios with misaligned inputs.

References

[1] PBVS 24. Thermal image super-resolution challenge (gtisr) - track2. https://codalab.lisn.upsaclay.fr/competitions/17014, 2023. Accessed on 2024-03-07. 2, 5, 6, 7, 8

[2] Feras Almasri and Olivier Debeir. Rgb guided thermal super-resolution enhancement. In *2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech)*, pages 1–5, 2018. 3

[3] Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte, and Luc Van Gool. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *CVPRW*, 2022. 2, 4, 8

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer, 2021. 2

[5] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration, 2022. 3, 6

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer, 2023. 4

[7] Yukyung Choi, Namil Kim, Soonmin Hwang, and In So Kweon. Thermal image enhancement using convolutional neural network. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 223–230, 2016. 2

[8] Yukyung Choi, Namil Kim, Soonmin Hwang, and In So Kweon. Thermal image enhancement using convolutional neural network. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 223–230, 2016. 2

[9] Vishal Chudasama, Heena Patel, Kalpesh Prajapati, Kishor Upla, Raghavendra Ramachandra, Kiran Raja, and Christoph Busch. Therisurnet - a computationally efficient thermal image super-resolution network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 388–397, 2020. 2

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks, 2017. 3

[11] Xuerui Dai, Xue Yuan, and Xueye Wei. Tirnet: Object detection in thermal infrared images for autonomous driving. *Applied Intelligence*, 51:1244–1261, 2021. 1

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks, 2015. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2, 4

[14] Rikke Gade and Thomas B Moeslund. Thermal cameras and applications: a survey. *Machine vision and applications*, 25: 245–262, 2014. 1

[15] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, 2021. 3

[16] 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

[17] Honey Gupta and Kaushik Mitra. Toward unaligned guided thermal super-resolution. *IEEE Transactions on Image Processing*, 31:433–445, 2022. 3, 8

[18] Aayesha Hakim and RN Awale. Thermal imaging-an emerging modality for breast cancer detection: a comprehensive review. *Journal of Medical systems*, 44:1–18, 2020. 1

[19] Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline, 2021. 2

[20] Roselyne Ishimwe, K Abutaleb, Faruk Ahmed, et al. Applications of thermal imaging in agriculture—a review. *Advances in remote Sensing*, 3(03):128, 2014. 1

[21] Prashanth Kannadaguli. Yolo v4 based human detection system using aerial thermal imaging for uav based surveillance applications. In *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pages 1213–1219. IEEE, 2020. 1

[22] Priya Kansal and Sabari Nathan. A multi-level supervision model: A novel approach for thermal image super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 2

[23] 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 (CVPR) Workshops*, pages 507–514, 2023. 3

[24] Birendra Kathariya, Zhu Li, and Geert Van der Auwera. Joint pixel and frequency feature learning and fusion via channel-wise transformer for high-efficiency learned in-loop filter in vvc. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023. 3, 8

[25] My Kieu, Andrew D Bagdanov, Marco Bertini, and Alberto Del Bimbo. Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In *European Conference on Computer Vision*, pages 546–562. Springer, 2020. 1

[26] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks, 2016. 2, 8

[27] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer, 2021. 2, 4

[28] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution, 2017. 2, 8

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 2

[30] Farzeen Munir, Shoaib Azam, Muhammd Aasim Rafique, Ahmad Muqeem Sheri, Moongu Jeon, and Witold Pedrycz. Exploring thermal images for object detection in underexposure regions for autonomous driving. *Applied Soft Computing*, 121:108793, 2022. 1

[31] Sabari Nathan and Priya Kansal. Leveraging multi scale backbone with multilevel supervision for thermal image super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4332–4338, 2021. 2

[32] Kalpesh Prajapati, Vishal Chudasama, Heena Patel, Anjali Sarvaiya, Kishor Upla, Kiran Raja, Raghavendra Ramachandra, and Christoph Busch. Channel split convolutional neural network (chasnet) for thermal image super-resolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4363–4372, 2021. 2, 3

[33] Qingpao Qin, Kan Chang, Mengyuan Huang, and Guiqing Li. Denet: detection-driven enhancement network for object detection under adverse weather conditions. In *Proceedings of the Asian Conference on Computer Vision*, pages 2813–2829, 2022. 1

[34] Rafael Rivadeneira, Patricia Suarez, Angel Sappa, and Boris Vintimilla. *Thermal Image SuperResolution Through Deep Convolutional Neural Network*, pages 417–426. 2019. 1, 2

[35] Rafael E Rivadeneira and Angel D Sappa. Thermal image super-resolution: A novel architecture and dataset. 2020. 2

[36] Rafael E. Rivadeneira, Angel D. Sappa, Boris X. Vintimilla, Lin Guo, Jiankun Hou, Armin Mehri, Parichehr Behjati Ardakani, Heena Patel, Vishal Chudasama, Kalpesh Prajapati, Kishor P. Upla, Raghavendra Ramachandra, Kiran Raja, Christoph Busch, Feras Almasri, Olivier Debeir, Sabari Nathan, Priya Kansal, Nolan Gutierrez, Bardia Mojra, and William J. Beksi. Thermal image super-resolution challenge - pbvs 2020. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 2

[37] Rafael E. Rivadeneira, Angel D. Sappa, Boris X. Vintimilla, Sabari Nathan, Priya Kansal, Armin Mehri, Parichehr Behjati Ardakani, Anurag Dalal, Aparna Akula, Darshika

Sharma, Shashwat Pandey, Basant Kumar, Jiaxin Yao, Rongyuan Wu, Kai Feng, Ning Li, Yongqiang Zhao, Heena Patel, Vishal Chudasama, Kalpesh Prajapati, Anjali Sarvaiya, Kishor P. Upla, Kiran Raja, Raghavendra Ramachandra, Christoph Busch, Feras Almasri, Thomas Vandamme, Olivier Debeir, Nolan B. Gutierrez, Quan H. Nguyen, and William J. Beksi. Thermal image super-resolution challenge - pbvs 2021. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4359–4367, 2021. 2

[38] Rafael E Rivadeneira, Angel D Sappa, Boris X Vintimilla, and Riad Hammoud. A novel domain transfer-based approach for unsupervised thermal image super-resolution. *Sensors*, 22(6):2254, 2022. 2

[39] Rafael E. Rivadeneira, Angel D. Sappa, Boris X. Vintimilla, Dai Bin, Li Ruodi, Li Shengye, Zhiwei Zhong, Xianming Liu, Junjun Jiang, and Chenyang Wang. Thermal image super-resolution challenge results - pbvs 2023. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 470–478, 2023. 3, 6, 7, 8

[40] Roslidar Roslidar, Aulia Rahman, Rusdha Muharar, Muhammad Rizky Syahputra, Fitri Arnia, Maimun Syukri, Biswajeet Pradhan, and Khairul Munadi. A review on recent progress in thermal imaging and deep learning approaches for breast cancer detection. *IEEE Access*, 8:116176–116194, 2020. 1

[41] Patricia L Suárez, Dario Carpio, and Angel D Sappa. Enhancement of guided thermal image super-resolution approaches. *Neurocomputing*, 573:127197, 2024. 3, 8

[42] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 341–348, 2022. 2

[43] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks, 2020. 4

[44] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration, 2021. 2, 4

[45] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module, 2018. 2

[46] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration, 2022. 2, 6

[47] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018. 6

[48] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks, 2018. 2, 8

[49] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 2, 8

[50] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution, 2022. 2

[51] Zhiwei Zhong, Xianming Liu, Junjun Jiang, Debin Zhao, Zhiwen Chen, and Xiangyang Ji. High-resolution depth maps imaging via attention-based hierarchical multi-modal fusion. *IEEE Transactions on Image Processing*, 31:648–663, 2022. 6

[52] Zhiwei Zhong, Xianming Liu, Junjun Jiang, Debin Zhao, Zhiwen Chen, and Xiangyang Ji. High-resolution depth maps imaging via attention-based hierarchical multi-modal fusion. *IEEE Transactions on Image Processing*, 31: 648–663, 2022. 2

[53] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020. 3