IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 January 2023, accepted 29 January 2023, date of publication 2 February 2023, date of current version 8 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3241852

RESEARCH ARTICLE

TnTViT-G: Transformer in Transformer Network for Guidance Super Resolution

ARMIN MEHRI[1], PARICHEHR BEHJATI[1], AND ANGEL DOMINGO SAPPA[1,2], (Senior Member, IEEE)
[1]Computer Vision Center, Autonomous University of Barcelona, 08193 Barcelona, Spain
[2]ESPOL Polytechnic University, Guayaquil EC090112, Ecuador

Corresponding author: Armin Mehri (amehri@cvc.uab.es)

This work supported in part by the Air Force Office of Scientific Research under Award FA9550-22-1-0261, in part by the Grant PID2021-128945NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", in part by the "CERCA Programme/Generalitat de Catalunya," and in part by the ESPOL Project CIDIS-12-2022.

ABSTRACT Image Super Resolution is a potential approach that can improve the image quality of low-resolution optical sensors, leading to improved performance in various industrial applications. It is important to emphasize that most state-of-the-art super resolution algorithms often use a single channel of input data for training and inference. However, this practice ignores the fact that the cost of acquiring high-resolution images in various spectral domains can differ a lot from one another. In this paper, we attempt to exploit complementary information from a low-cost channel (visible image) to increase the image quality of an expensive channel (infrared image). We propose a dual stream Transformer-based super resolution approach that uses the visible image as a guide to super-resolve another spectral band image. To this end, we introduce Transformer in Transformer network for Guidance super resolution, named TnTViT-G, an efficient and effective method that extracts the features of input images via different streams and fuses them together at various stages. In addition, unlike other guidance super resolution approaches, TnTViT-G is not limited to a fixed upsample size and it can generate super-resolved images of any size. Extensive experiments on various datasets show that the proposed model outperforms other state-of-the-art super resolution approaches. TnTViT-G surpasses state-of-the-art methods by up to $0.19 \sim 2.3dB$, while it is memory efficient.

INDEX TERMS Single image super resolution, guidance super resolution, transformers, convolutional neural network.

I. INTRODUCTION

In recent years, image super resolution, has achieved significant interest from both academic and industrial communities. The process of reconstructing a high-resolution (HR) image from its low-resolution (LR) counterpart is referred to as the super resolution (SR) problem in the field of image processing. Due to the fact that a single LR image might have numerous mappings from LR to SR, SR is an ill-posed problem, which is also known as a one-to-many problem. Thus, numerous SR methods have been introduced to reconstruct a high-resolution image from its low-resolution ones, such as traditional approaches like the self-exemplars approach [1], anchoring neighborhood regression [2], sparse representation [3] and random forest [4].

More recently, by advancing the deep learning approaches, several Convolutional Neural Networks (CNNs) and Transformer networks are being used as a solution for the ill-posed SR problem. This is largely attributable to the recent successes of deep learning approaches in a variety of vision tasks, such as object detection, image recognition, semantic segmentation, image classification, and many others. The first work in this direction has been presented by Dong et al. [5], who developed a three-layer CNN model to train a nonlinear LR-to-HR mapping function called SRCNN, which greatly outperforms the traditional machine

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate.

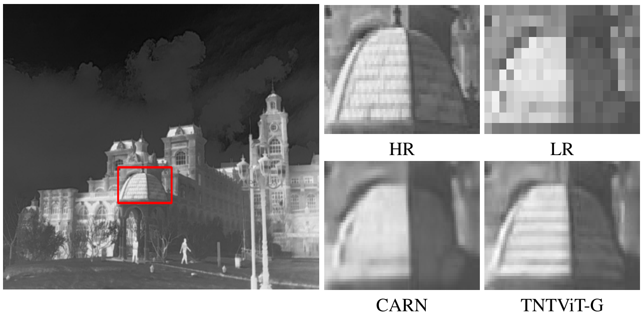


FIGURE 1. An example of super-resolved results comparison on M3FD dataset for scale factor ×4.

learning-based methods. The majority of later expansions of SRCNN enhance SR accuracy by employing more complicated network designs (such as RDN [6], EDSR [6], RCAN [7], among others) or by utilizing a training dataset with better quality.

However, in real-world applications, the environment around us is dynamic and changing all the time due to many known and unknown reasons, which require dealing with various challenging conditions such as rain, fog, occlusions, poor lighting, low resolution, and many others. All these factors make it difficult for an algorithm that uses only a visible-band sensor (RGB) to achieve high performance under these conditions [8]. Therefore, the visible image is found to be insufficient for such cases, and cross-spectral images have become increasingly necessary in many applications as they are robust against obstacles in visual environments and provide support to the RGB images. Cross-spectral images (e.g., infrared images) have been used in many ranges of specialized fields such as surveillance [9], military affairs [10], pedestrian tracking [11], firefighting [12] and many others. However, the associated costs of having such images grow significantly with the increase in their resolution.

Various approaches and algorithms have been proposed to improve the resolution of different infrared images using hardware or software. Employing low-resolution cameras that are less expensive than high-end cross-spectral domain cameras and using SR methods to increase the resolution of such images is one strategy to boost the consumer applicability of such cameras to deal with a challenging situation at a lower cost. However, as previously stated, single image SR (SISR) is a tricky operation that becomes even more difficult when the input image has a very poor resolution (such as the ones produced by inexpensive cross/multi-spectral sensors), which SISR techniques may hallucinate missing details from low-resolution inputs and therefore yielding to artifacts [13].

To address the aforementioned problems, a fundamental solution is to take advantage of any additional information that can be found with the low-resolution infrared images as most cross-spectral cameras are accompanied by an inbuilt visible RGB camera with higher resolution. As a result, it is permissible to use low-cost visible images as additional information to considerably improve the accuracy of the SR results obtained from the costly infrared images. For example, long-wave infrared (LWIR) detectors, required to capture thermal images, are sealed inside their own separate vacuum packages in order to carry out the high-precision thermal measurement, which is a procedure that is both time-consuming and costly [14]. As a result, the cost of LWIR sensors is much higher than that of RGB ones with comparable spatial resolutions. The majority of commercially available LWIR cameras capture LR images (for example, 160 × 120 or even 80 × 60 pixels) [15], in which significant information is severely lost.

In this paper, we attempt to boost the performance of image restoration in the expensive channel by taking into account the complementary information captured by an additional low-cost visible sensor. The primary focus of this work is to build a deep learning model that applies multimodal sensor fusion using visible cross-spectral images. The proposed approach is evaluated with two different schemes (i.e., thermal infrared (LWIR), and near infrared (NIR)), but is also valid for any other input data. The proposed model accepts two images as inputs and integrates them in such a way as to enhance the generated infrared image resolution with fine detail with the help of the corresponding visible image. Thus, a guidance super resolution network ($TnTViT - G$) is proposed to enhance the LR infrared image by integrating the rich information from the HR visual image. We show that HR visual images can help the model fill in the missing values and generate higher frequency details in the reconstructed SR infrared image, as shown in Fig. 1.

The main contributions can be summarized as follows:

- We present TnTViT-G, an efficient dual-stream Transformer-based network for guidance super resolution (GSR) task. The TnTViT-G Transformer blocks are built on top of the idea of the Vision Transformer (ViT) by completely revising the self-attention layer.
- We present a lightweight Dual Attention layer that significantly improves the reconstruction quality by generating a global attention map from two local attention weights, which are obtained individually by two branches in parallel while it is not memory hunger.
- We present a high-quality arbitrary upsampling module, which is capable of producing SR images at any scale factor.
- Extensive experiments show that TnTViT-G outperforms CNN and transformer-based networks on different benchmark datasets for the GSR task.

The rest of the paper is organized as follows: Section II discusses the related work. Section III describes the proposed TnTViT-G and its core components in detail. Experimental comparisons against several state-of-the-art methods are presented in Section IV. The model investigation is presented in section V. Section VI concludes the paper.

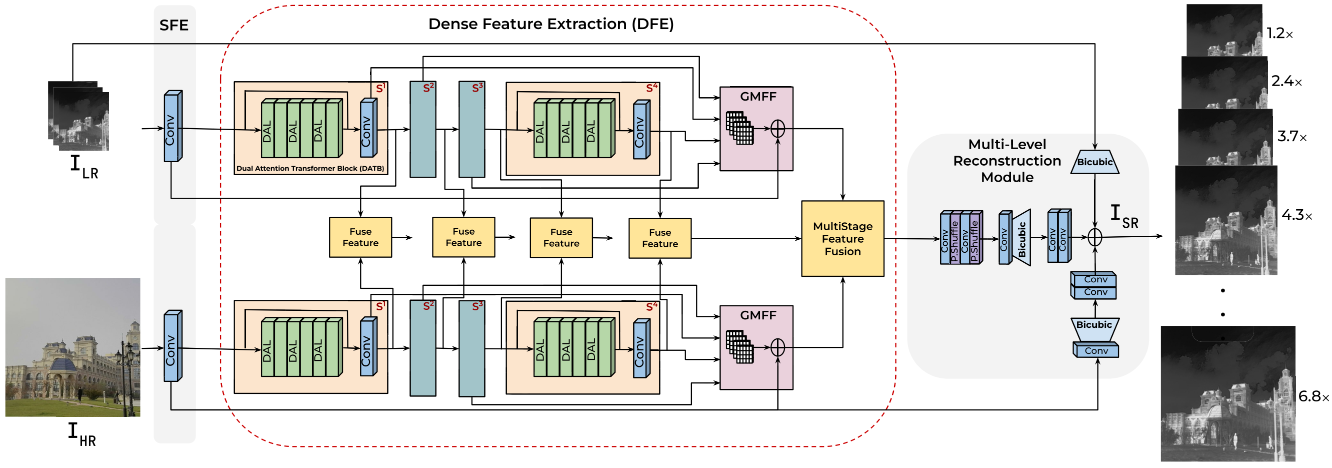


FIGURE 2. The overall network architecture of the proposed TnTViT-G.

II. RELATED WORK

In this section, the most recent state-of-the-art SR deep learning CNN and Transformer-based approaches are detailed.

A. DEEP LEARNING BASED SINGLE IMAGE SUPER-RESOLUTION

Single Image Super Resolution aims to restore the well-detailed image from its low-quality counterpart. The first deep learning-based work has been introduced by Dong et al. [5] (SRCNN), which uses a convolutional neural network to tackle the SR task. The SRCNN presents a shallow neural network that receives an upsampled image as an input that costs extra computation. Later on, to address this drawback, FSRCNN [16] and ESPCN [17] have been proposed by receiving the LR image as input to reduce the large computational and run-time cost and upsampling the features near the output of the network by a single transposed convolution layer. Even though the strength of deep learning comes from deep layers, the above-mentioned methods are referred to as shallow networks. Therefore, Kim et al. [18] use residual learning to ease the training challenges and increase the depth of their network by adding 20 convolutional layers. Then, [19] proposes a memory block in MemNet for deeper networks and solves the problem of long-term dependency with 84 layers. Lim et al. [20] introduces EDSR by expanding the network size and enhancing the residual block by omitting the batch normalization from the residual block. Zhang et al. [6] propose RDN with residual and dense skip connections to fully use hierarchical features.

Moreover, in recent years, there has been a rise in interest in developing lightweight approaches for super resolution tasks in order to lower the high computing cost of SR task. Ahn et al. [21] design an efficient network that is suitable for the mobile scenario. Later, [22] introduces MAFFSRN by proposing multi-attention blocks to improve the performance. LatticeNet [23] introduces an economical structure to adaptively combine residual blocks. Recently, OverNet presented by [24], designs an efficient network structure with a multi-loss function to boost the network's performance. Also, a neural architecture search (NAS)-based strategy has been proposed in SISR to construct efficient networks. MoreMNA-S [25] and FALSR [26] are two examples of networks that use NAS strategies in their network. However, due to the limitations of the NAS strategy, the performance of these models is limited.

B. VISION TRANSFORMER

Natural Language Processing (NLP) is revolutionized by the groundbreaking performance of Transformer networks. Transformer networks, unlike Convolutional Neural Network approaches, have the benefit of being able to capture long-range dependencies of the input sequences by using a self-attention layer. The "self-attention" layer is the fundamental concept of the Transformer network. The Computer Vision community was inspired to modify the Transformer to use in Vision problems. The initial research in this field was carried out by Alex et al., who proposed the Vision Transformer known as ViT [27], which replaces the conventional CNN with the Transformer. ViT is directly trained on the medium-sized flattened patches with large-scale data pre-training.

Since introducing the first work, many Transformer based architectures have been proposed for the Vision tasks such as in image recognition [28], object detection [29], [30], segmentation [31], [32], and action recognition [33], [34]. In addition, Transformer based models have been studied for low-level vision problems such as super resolution [35], [36], image colorization [37], denoising [38], and image restoration [39]. For instance, DETR [29] is a transformer network designed for object detection, which can predict a set of objects and model their relationships. SwinIR [35], introduced by Jingyun et al. for low-level vision tasks, uses Swin Transformer [30] by applying self-attention within local image regions to solve the low-level vision problems.

C. GUIDANCE SUPER RESOLUTION

Guidance Super Resolution (GSR) techniques have been used to upsample images from a different domain to generate more accurate SR images by using the information of other domain images (e.g., visible images), while having such high-resolution images is expensive (e.g., thermal images). Traditional GSR approaches, such as joint bilateral upsampling [40] and rapid bilateral filtering [41], have already been studied for this task; however, these methods frequently over-smooth the reconstructed image. Recently, by advancing deep learning methods, several approaches have been introduced to boost the performance of the GSR task [42]. GSR techniques have been studied in different super resolution domains, such as depth-map SR, infrared SR, thermal SR, hyperspectral SR, and some others. MSG-Net [43], employs CNNs to accomplish guidance super resolution, which is the first CNN model that attempts to upsample depth images under multi-scale guidance from the corresponding HR visible images.

Most GSR methods are based on the Siamese algorithm, which lets the network accept two inputs and perform simultaneous feature extraction from both spectral images and visible images at the same time. These images are then fused at different levels of the network and upsampled to provide high-resolution images. Furthermore, GSR approaches with similar structures used in guidance hyperspectral SR methods include [44], [45]. Also, some models were proposed for guidance infrared SR such as [46], [47]. Feras et al. [13] propose a multimodal sensor fusion model to enhance the thermal images with help of RGB images. Riccardo et al. [48] propose an alternative interpretation of guided super resolution, where the roles of the source and guide images are swapped. Honey et al. [42] propose a network for GSR from low-resolution thermal images that do not require pixel-to-pixel alignment between the thermal and the guide image. Moreover, some approaches for cross-modal guidance super resolution extract edges from the visible images in order to obtain high-frequency features. The use of edge-based guiding facilitates the reconstruction of higher-frequency features such as [49], [50]. Despite that the aforementioned approaches achieve reasonable performance, these methods are limited to a fixed scale factor and not ideal for real-world applications due to the number of network parameters and their performance. Hence, in this paper, we propose a novel dual stream Transformer based network for GSR, which archives remarkable performance with a completely new design for an upsampling module to be able to reconstruct SR images in any arbitrary size while reminding computationally efficient.

III. PROPOSED METHOD

In this section, the overall network architecture of the proposed TnTViT-G is described. Following that, more information about the Dual Attention layer is provided. TnTViT is designed for Single Image Super Resolution and TnTViT-G is a siamese-based network of TnTViT, which designs for Guidance Super Resolution.

A. OVERALL PIPELINE

The main objective of the proposed model is to design an efficient Transformer-based network for Guidance Super Resolution that is capable of producing fine details high-quality images with the help of the guided images (e.g., visible images) to boost the performance of the network while staying computationally low. Thus, we employ the original Transformer structure but modify it in such a way that the model achieves a considerable performance advantage over existing CNN and Transformer networks. The overall architecture of the TnTViT-G is illustrated in Fig. 2. It consists of two streams to extract the features of LR infrared input images and HR visible images. In particular, the proposed TnTViT-G consists of three modules: Shallow Feature Extraction (SFE), Dense Feature Extraction (DFE), and Multi-Level Reconstruction Modules. We defined I_{LR}^{IR}, I_{HR}^{Vis}, and I_{SR}^{IR} as the low-quality infrared, high-quality RGB inputs, and high-quality output of our network, respectively.

B. SHALLOW FEATURE EXTRACTION

Given the input images to the network, we apply a single 3 × 3 convolutional layer on each network's streams to the provided LR and HR visible inputs in order to map these images space to a higher dimensional feature space and improve the performance of the network [51]. Therefore, we extract the shallow features as follows:

$$\boldsymbol{F}_0^{IR} = Conv_{3\times3}(\boldsymbol{I}_{LR}^{IR}), \boldsymbol{F}_0^{Vis} = Conv_{3\times3}(\boldsymbol{I}_{HR}^{Vis}), \tag{1}$$

where $F_0^{IR}(.)$ and $F_0^{Vis}(.)$ denotes the output of shallow feature extraction on both infrared and visible images.

C. DENSE FEATURE EXTRACTION

After mapping the inputs to a higher dimensional feature space, the features pass through the Dense Feature Extraction F_{DFE} to encode the information in order to understand the context of the sequences. The feature encoders of the proposed approach (i.e, Dense Feature Extraction) is a Transformer based network, which shares between both input images (I_{LR}^{IR} and I_{HR}^{Vis}) to keep the network computationally efficient. However, each stream receives the same patch of the input image with different sizes since LR images are relatively smaller than visible images. Particularly, Dense Feature Extraction design by using several Transformer blocks to extract abstract features and spotlights the high-level information. Each Transformer block consists of several Transformer layers and a 1 × 1 Conv layer with the benefit of cascade connections to transfer the information from the previous stage to the current stage and help the gradient flow of the network. Thus, we extract the feature as follows:

$$\boldsymbol{F}_{DFE} = H_{DFE}(F_0^{IR}; F_0^{Vis}), \tag{2}$$

where $H_{DFE}(.)$ is Dense Feature Extraction with several Transformer blocks which can be formulated as

$$\boldsymbol{F}_i = Conv_{1\times1}(C[H_{DATB}(F_{i-1}), X_{i-1}], i = 1, 2, \ldots, K, \quad (3)$$

where $H_{DATB}(.)$ denotes the i_{th} Dual Attention Transformer Blocks. C stands for the concatenation operation between the initial and output features of each *DATB* block. *Conv* denotes the convolutional layer after concat operation within each DATB. Using a convolutional layer in the Transformer block, help to transfer inductive bias from the convolution operation into the Transformer network and provide a more solid foundation for the later aggregation with shallow features.

After encoding the features through several DATB, the output feature maps of each DATB stage are concatenated together to highlight the positional information via the GMFF module, which stands for Gated multi-layer perceptron (MLP) Feature Fusion, before reconstructing the SR images. GMFF module is designed to generate a multi-stage representation feature map of Transformer blocks. Later, the feature map passes through a lightweight MLP network. However, unlike to standard MLP network, the GMFF's MLP module is designed by using a 3 × 3 depthwise Conv layer and gating mechanism technique to first, leak the spatial information since highlighting such features is important in SR tasks to achieve high performance. Second, allowing useful information to pass through the network and suppressing the less informative ones. The gating mechanism is used by applying the element-wise product of two parallel routes of linear transformation layer that one of which is activated with the GELU. Thus, Gated MLP Feature Fusion can be seen as follows:

$$\boldsymbol{F}_{GMFF} = MLP(GELU(Conv_{3\times3}(MLP(F_i)))) + F_0, \quad (4)$$

where F_{GMFF} is the output of DFE with aggregation of the initial features, which is later used by the Multi-Stage Feature Fusion Module.

D. MULTI STAGE FEATURE FUSION MODULE

After encoding the information of both LR infrared image I^{IR}_{LR} and HR visible image I^{Vis}_{HR} with a dual stream shared network, the LR features first scale up to the same spatial size as the HR visible image before fusing the information with a learnable bicubic upsampling that contains a conv layer before it; later, the aggregated features of all the stages are concatenating together to enhance the LR infrared images before upsampling them to the desired output size.

$$\begin{aligned}&\boldsymbol{F}_{MSFF}\\&= Conv_{1\times1}(C[H_{UP}(F^{IR}_{GMFF}), F^{Vis}_{GMFF}, FF_{S1}, \ldots, FF_{S4}]),\end{aligned} \quad (5)$$

where $F_{MSFF}(\cdot)$ denotes the output of the multi-stage feature fusion module of both TnTViT streams and the feature fusion of each stage.

E. MULTI LEVEL RECONSTRUCTION MODULE

Later, to upsample the LR infrared image after fusing the information, we propose a new inductive bias in GSR architectures to generate SR images more accurately with fewer artifacts compared with the other methods or naive interpolation techniques. To do so, we first pass the information through two-pixel shuffle layers and a conv layer before each of them. Second, the upsampled features with pixel shuffle layers feed to a learnable bicubic interpolation to upscale the features to any arbitrary size. Later, the information aggregated with the shallow features of the HR-guided image. These features are also upscaled with learnable bicubic interpolation.

$$\boldsymbol{I}_{SR} = H^{\uparrow}_{Rec}(H_{UP}(F^{IR}_0) + F_{MSFF} + H_{UP}(I_{LR})), \quad (6)$$

where $H_{Rec}(\cdot)$ and I_{SR} denote the up-sampling module and high-quality reconstructed image respectively. Hence, the proposed module can learn how to refine the pixels more correctly via different levels of upscaling to bring it closer to its actual high-resolution counterpart and beyond. Extensive experiments have been detailed in the ablation study to show the efficiency of the proposed reconstruction module over other approaches.

F. LOSS FUNCTION

To keep the consistency with previous works, we use L_1 loss as a cost function during training to optimize the parameters of the proposed TnTViT-G.

$$L_1(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \|I_{SR} - I_{HR}\|_1 \,, \quad (7)$$

where I_{SR} is obtained by taking a low-quality infrared image as the input of our model and I_{HR} is the corresponding ground truth.

In the next subsections, we provide more details about our Transformer layer.

G. DUAL ATTENTION LAYER

This section presents the proposed Dual Attention layer, an architecture abstracted from the general multi-head Transformer layer [52] with revising the self-attention layer. As is generally known, self-attention is critical to achieve excellent performance in Transformer-based networks. However, self-attention might be troublesome for a variety of reasons. For example, the computational complexity of self-attention grows quadratically with the number of tokens to mix. Also, due to its nature, self-attention does not take into account the local contextual information and treats the images as flattened sequences that ignore the structure of the image. Thus, we propose the Dual Attention layer to address the mentioned limitations by constructing a global attention map at a lower computational cost. The dual Attention layer creates a global attention map by combining two local attention maps, which are obtained in parallel by using a CNN-based Attention Module and a Transformer self-attention layer. Unlike the

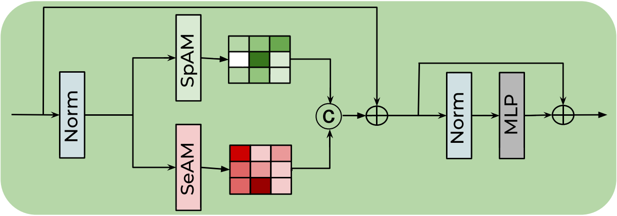


FIGURE 3. Illustration of dual attention layer (DAL).

prior token mixer, Dual Attention is able to take into account both long-range dependency and local contextual information with less computing cost.

As shown in Fig 3, we design our Dual Attention such that the channel information is distributed evenly across both attention module branches (SpAM and SeAM). Both attention branches get half of the input tensor from the Norm layer tensor X to generate the local attention map independently. The SeAM is a self-attention Transformer, which first generates the query (Q), key (K), and value (V) projections, enriched with the local context. Inspired by [39], we apply SeAM only across the channels rather than the spatial dimensions. Our SeAM uses only depth-wise convolutions to emphasize the channel-wise spatial context before computing feature covariance to produce the attention map. Thus, Q, K, V are computed as:

$$Q = W_d^Q Y, K = W_d^K Y, V = W_d^V Y, \tag{8}$$

where $W_d^{(\cdot)}$ is the 3×3 depth-wise convolution. Next, query and key projections reshape in such a way that their dot-product interaction generates a transposed-attention map. Thus, the attention map generates the following:

$$Attention(Q, K, V) = W_d(V.Softmax(K.Q/\alpha)) + X, \tag{9}$$

where X is the input feature map and α is a learnable scaling parameter that is used to regulate the magnitude of the dot product of K and Q before applying the Softmax function. Similar to previous works [35], [52], [53] we perform the attention function for h times to learn separate attention maps in parallel in our SeAM module.

The second branch of our Dual Attention layer is the Spatial Attention Module (SpAM), which is an almost parameter-free attention mechanism. SpAM receives the other half of the input tensor to generate the local attention map. The goal of SpAM module is to encode the spatial information, which represents the importance of each pixel in the input feature with a negotiable cost. Given half of the input tensor information, the channels of the input tensor are reduced by mean and max operations, of which the shape is $1 \times H \times W$. The obtained features are concatenated, then pass through a dilated convolution layer with a kernel size of 3×3. After, a sigmoid activation layer applies to the output feature to generate the attention weights of shape $1 \times H \times W$, which are later multiplied with the input tensor to refined tensors of shape $C \times H \times W$. Thus, the SpAM can be formulated as follows:

$$X = Sigmoid(Conv_{3\times 3}[F_{Mean}(X), F_{Max}(X)]) * X, \tag{10}$$

where $F_{Mean}(\cdot)$ and $F_{Max}(\cdot)$ denotes for mean and max operations. Later, both locally generated attention maps from SpAM and SeAM are concatenated together to obtain a unified global attention map with less computational cost. Thus, the generated attention map contains both long-range dependency and local context information with enrich of spatial features.

Following that, a multi-layer perceptron (MLP) with two fully connected layers and a GELU non-linearity activation function between them is employed for further feature modifications. The norm layer is also added before MLP, and both modules contain the residual connection between them. As a result, the entire procedure within our Dual Attention is as follows:

$$\begin{aligned} X &= (Norm(SpAM(X/2) + SeAM(X/2))) + X \\ Y &= MLP(Norm(X)) + X \end{aligned} \tag{11}$$

where $Norm(\cdot)$ stands for normalization layer and Y for the output feature map.

IV. EXPERIMENTAL RESULTS

A. SETTING

1) DATASETS

Two datasets have been used to perform the experiments, namely M3FD [54] and RGB-NIR [55]. The first dataset is M3FD, which newly released by [54]. The M3FD dataset contains pairs of visible and thermal images. The dataset was built with a synchronized system of one binocular optical camera and one binocular thermal sensor to capture corresponding two modality images. We use the M3FD Fusion dataset, which consists of 300 aligned pair images from different scenarios in daytime, night, and overcast. Also, the dataset consists of images from different scenes, such as road, campus, street, forest, and many others.

The second dataset is the RGB-NIR Scene [56] dataset. The RGB-NIR Scene dataset contains aligned pairs of 477 RGB and near-infrared images, divided into 9 categories such as country, field, forest, indoor, mountain, old building, street, urban, and water. The images were acquired by utilizing different exposures from customized SLR cameras equipped with visible and near-infrared filters.

2) EVALUATION PROTOCOL

Two widely used quantitative metrics have been considered to measure the performance of our TnTViT compared to other approaches. We used the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) to measure reconstructed SR image accuracy. PSNR assesses the image by statistically calculating distortion levels between the reconstructed and ground-truth images. SSIM measures the structural similarity between two images based on luminance, contrast, and structure, which has a value range between [0-1]. Higher value, better for both PSNR and SSIM.

TABLE 1. Average PSNR, SSIM comparisons with SOTA CNN- and Transformer-based methods with the same range of network parameters on the Bicubic (BI) degradation for scale factors [×2, ×4, ×8]. Best results are highlighted.

Scale	Method	DM	G/S	M3FD		RGB-NIR	
				PSNR↑	SSIM↑	PSNR↑	SSIM↑
×2	Bicubic	BI	Single	37.74	0.9465	31.91	0.8792
×2	CARN [21]	BI	Single	38.82	0.9538	33.05	0.8982
×2	SwinIR [35]	BI	Single	37.89	0.9498	31.84	0.8863
×2	TNTViT [OURS]	BI	Single	38.91	0.9542	33.14	0.9002
×2	PixTransform [48]	BI	Guided	25.59	0.7274	29.09	0.8255
×2	UGSR [42]	BI	Guided	33.75	0.9354	30.74	0.8959
×2	TNTViT-G [OURS]	BI	Guided	**39.01**	**0.9556**	**34.49**	**0.9202**
×4	Bicubic	BI	Single	30.79	0.8435	26.63	0.7129
×4	CARN [21]	BI	Single	31.58	0.8336	27.33	0.7284
×4	SwinIR [35]	BI	Single	30.82	0.8457	26.12	0.7177
×4	TNTViT [OURS]	BI	Single	31.64	0.8646	27.40	0.7395
×4	PixTransform [48]	BI	Guided	25.24	0.7086	27.85	0.7900
×4	UGSR [42]	BI	Guided	28.43	0.8323	28.28	0.8232
×4	TNTViT-G [OURS]	BI	Guided	**32.00**	**0.8735**	**29.59**	**0.8252**
×8	Bicubic	BI	Single	26.77	0.7594	24.10	0.6142
×8	CARN [21]	BI	Single	27.41	0.7787	24.79	0.6348
×8	SwinIR [35]	BI	Single	26.81	0.7621	24.18	0.6153
×8	TNTViT [OURS]	BI	Single	27.50	0.7607	24.90	0.6348
×8	PixTransform [48]	BI	Guided	23.78	0.6643	26.10	0.7454
×8	UGSR [42]	BI	Guided	27.76	**0.7875**	**26.34**	0.7453
×8	TNTViT-G [OURS]	BI	Guided	**27.88**	0.7628	26.21	**0.7635**

3) DEGRADATION MODELS

Degradation models have been created to replicate LR images in order to demonstrate the effectiveness of our proposed approach. The degradation model is bicubic downsampling (BI), which simulates LR images with the scale factors [×2, ×4, ×8] by applying bicubic downsampling to HR images.

4) IMPLEMENTATION DETAILS

We randomly select 70%, 20%, and 10% of images of each dataset for the training, validation, and test phases respectively. In the training phase, we provide the image patches as inputs with different sizes based on the size of each dataset from LR images and corresponding RGB images. The batch size has been set to 32 for the training. Horizontal random flips and 90 degree rotation data augmentation were applied to patches of images. Adam optimizer has been used with the initial learning rate 10^{-3}. $L1$ is used as a loss function to optimize the model and the network has been trained for $150K$ iterations. Also, the configurations of our transformer encoder are as follows: we use 4 Transformer blocks within 6 Transformer layers for each block, Embedding dimension set to 64, and MLP ratio to 2 for all Transformer blocks. TnTViT-G is developed using the PyTorch framework and trained on a single NVIDIA RTX 3090 GPU to achieve its performance.

B. COMPARISON WITH STATE-OF-THE-ART METHODS

In this section, we compare our proposed GSR (TnTViT-G) and SISR (TnTViT) with other lightweight state-of-the-art approaches on different datasets with different scale factors.

1) EXPERIMENTS ON BICUBIC DEGRADATION

Table 1 shows comparisons between the proposed approaches (TnTViT and TnTViT-G) and SOTA CNN- and Transformer-based models, CARN [21], SwinIR [35], PixTransform [48], and UGSR [42] on the Bicubic (BI) degradation model for scale factors [×2, ×4, ×8]. It is worth mentioning that these networks contain almost the same number of network parameters, allowing for a fair comparison. As can be observed, when the proposed method is compared to the approaches mentioned above, TnTViT achieves better

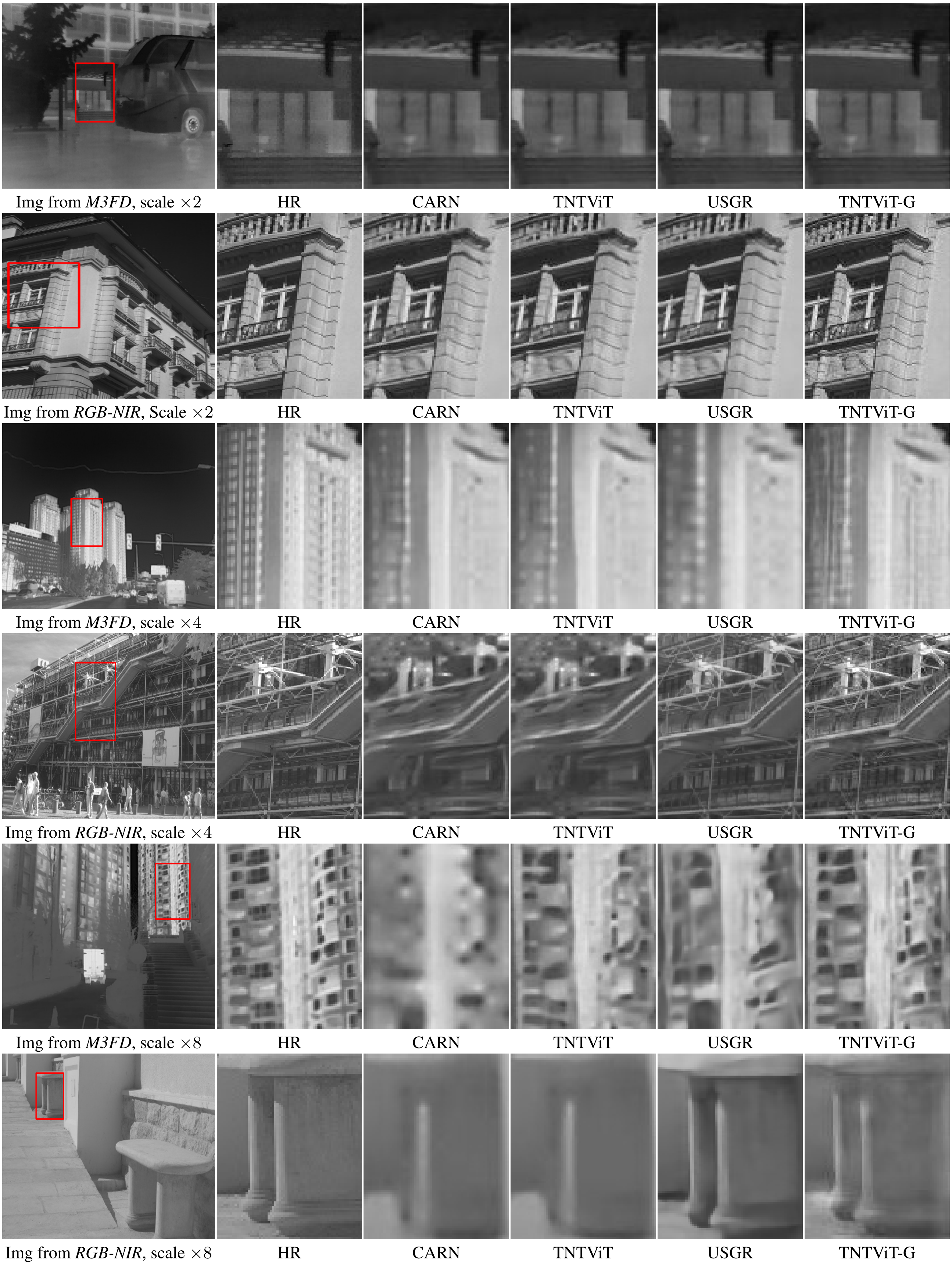

Img from *M3FD*, scale ×2
HR
CARN
TNTViT
USGR
TNTViT-G
Img from *RGB-NIR*, Scale ×2
HR
CARN
TNTViT
USGR
TNTViT-G
Img from *M3FD*, scale ×4
HR
CARN
TNTViT
USGR
TNTViT-G
Img from *RGB-NIR*, scale ×4
HR
CARN
TNTViT
USGR
TNTViT-G
Img from *M3FD*, scale ×8
HR
CARN
TNTViT
USGR
TNTViT-G
Img from *RGB-NIR*, scale ×8
HR
CARN
TNTViT
USGR
TNTViT-G

FIGURE 4. Visual results for scale factors [×2, ×4, ×8] on M3FD an RGB-NIR datasets respectively.

TABLE 2. Average PSNR results on RGB-NIR dataset for different upscaling methods with arbitrary scales. Best results are **highlighted**, second best underlined.

Experiment	Scale				
	×2	×2.2	×2.4	×2.6	×2.8
Pixel Shuffle	34.06	–	–	–	–
P.S. Bicubic	34.21	34.89	34.73	34.67	34.53
TNTViT MLUP	**34.49**	**35.24**	**35.07**	**34.90**	**34.71**
	×3	×3.1	×3.3	×3.5	×3.7
Pixel Shuffle	32.18	–	–	–	–
P.S. Bicubic	32.31	32.44	32.08	31.95	31.74
TNTViT MLUP	**32.57**	**32.45**	**32.29**	**32.15**	**32.02**
	×4	×4.2	×4.4	×4.6	×4.8
Pixel Shuffle	29.11	–	–	–	–
P.S. Bicubic	29.44	29.77	29.62	29.54	29.48
TNTViT MLUP	**29.59**	**29.85**	**29.81**	**29.73**	**29.64**

results without the help of any guided image (visible image). Furthermore, the proposed method (TnTViT-G) with the guidance of visible image information achieves superior results in almost most of the cases with major margins. This demonstrates that TnTViT-G continually accumulates this hierarchical information from different spectral images in order to construct more robust representative features that are well-focused on spatial context information since that is the key to an accurate SR image. This claim is validated by the derived SSIM scores, which are based on the visible structures in the image and hence are more accurate. Figure 4 shows some qualitative results on M3FD and RGB-NIR datasets on different scale factors for SOTA methods of both SISR and GSR. As can be seen, TnTViT produces images better than the existing method in the SISR since the network is able to focus better on spatial information. However, TnTViT-G is able to reconstruct high-frequency details significantly better than all the existing methods and generates more accurate SR infrared images, which are more similar to ground truth images.

V. ABLATION STUDY

The proposed model is further studied in an extended ablation investigation to demonstrate its efficiency. The ablation study is intended to offer further information about the performance of the proposed approach.

A. VISUALIZATION ON IMPACT OF GUIDED IMAGE

Figure 5 shows the average feature maps of each stage of our Dense Feature Extraction module to investigate the impact of the guidance image (i.e., visible image) when it is stacked up with the LR feature map in each stage of DFE. Each average feature map reflects the output of the Transformer block at each stage in the Dense Feature Extraction module. The average feature maps without guidance images are presented on the top row, while those with guidance images are shown on the bottom row. We can observe from the feature maps that using a guidance image helps the network acquire sharper representations. Second, as the network focuses more on high-level information, feature maps tend to include more negative values at each stage, showing a greater influence of suppressing the smooth area of the input image, which yields a more accurate SR output.

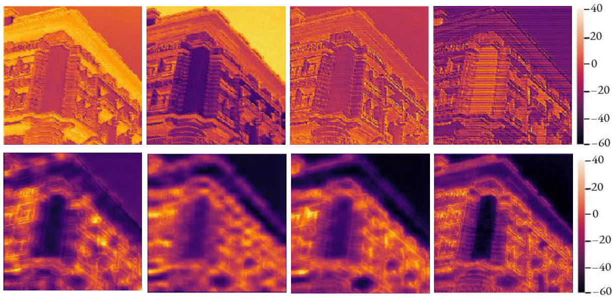


FIGURE 5. Average feature maps of TnTViT (*top*) and TnTViT-G (*bottom*) on different stages of dense feature extraction.

TABLE 3. Avargae LPIPS comparison between proposed method and the other methods on benchmark datasets for scale factors [×2, ×4]. The lower is better.

Methods	Scale	M3FD	RGB-NIR
CARN [21]	×2	0.1127	0.1365
SwinIR [35]	×2	0.2076	0.2291
UGSR [42]	×2	0.2879	0.1327
TNTViT (Ours)	×2	0.1013	0.1224
TNTViT-G (Ours)	×2	**0.0916**	**0.0934**
CARN [21]	×4	0.2418	0.3371
SwinIR [35]	×4	0.3176	0.3985
UGSR [42]	×4	0.3361	0.2533
TNTViT (Ours)	×4	0.2322	0.3262
TNTViT-G (Ours)	×4	**0.2119**	**0.2202**

B. INFLUENCE OF MULTI-LEVEL RECONSTRUCTION MODULE

We investigate the advantages of using the proposed Multi-Level Reconstruction Module, as well as the impact

TABLE 4. Average running time (s), memory consumption (MB), number of parameters (K), FLOPs (G), and PSNR comparison on RGB-NIR dataset for ×4.

Methods	Parameters (K)	FLOPs (G)	Memory (MB)	Running Time (s)	PSNR
CARN [21]	1,592	90.9	1,230	0.072	27.33
SwinIR [35]	929	49.6	3,110	0.185	26.12
USGR [42]	4,500	–	2,150	0.424	28.28
TnTViT (OURS)	1,238	43.7	2,324	0.116	27.40
TnTViT-G (OURS)	1,317	79.5	2,549	0.204	**29.59**

of two widely used upsampling and interpolation approaches on reconstruction results. We carried out the following experiments: *i*) Directly employing Pixel Shuffle layer to produce images after fusing the information of both network's streams instead of our MLUP; *ii*) Using Pixel Shuffle layer followed by a conv layer and bicubic interpolation to scale the generated SR image to arbitrary scales. As can be seen in Table 2, when the proposed MLUP module is used for upscaling, superior results are obtained by a large margin compared to other upsampling techniques. These studies demonstrate that, opposite to common practice, the MLUP significantly improves reconstruction accuracy since the module is able to generate the SR images in multi-level with the access of both direct and indirect shallow and abstract features and yields consistent improvements on benchmark datasets.

C. LEARNED PERCEPTUAL IMAGE PATCH SIMILARITY

In Table 3, we provide the Learned Perceptual Image Patch Similarity (LPIPS) [57] evaluation metric to evaluate the quality of the generated super-resolved images. LPIPS has been demonstrated to correlate better with human perceptual similarity of image quality than other evaluation metrics (i.e., PSNR and SSIM). LPIPS is a deep-feature-based evaluation metric that calculates the perceptual distance between two images. As can be seen, the proposed model achieves a lower value than other approaches (lower is better). PixTransform [48] was excluded from the table since it could not even outperform bicubic interpolation. This shows the effectiveness of the proposed TnTViT-G to generate more accurate super-resolved IR images with the help of HR visible images.

D. MODEL COMPLEXITY AND INFERENCE TIME ANALYSIS

Table 4 compares the proposed TnTViT and TnTViT-G architectures with existing CNN and Transformer-based architecture approaches on RGB-NIR test images in terms of network Parameters (M), FLOPs (G), Memory consumption (MB), and Running Time (s). To provide a fair comparison, all models are tested using the same setup, including their public source code and default hyper-parameters, on an Intel Core i9-10900K CPU and an NVIDIA RTX 3090 GPU. As can be seen, TnTViT generates the SR images faster than other Transformer methods. This comparison shows that our proposed model properly balances performance and running time requirements.

TABLE 5. Summary of abbreviations.

List of abbreviations and their associated meanings	
TnTViT	Transformer in Transformer Network for Single Image Super Resolution
TnTViT-G	Transformer in Transformer Network for Guidance Super Resolution
SR	Super Resolution
SISR	Single Image Super Resolution
GSR	Guided Super Resolution
CNN	Convolutional Neural Network
ViT	Vision Transformer
MLP	Multi-layer Perceptron
SOTA	State-of-the-art
SFE	Shallow Feature Extraction
DFE	Dense Feature Extraction
MLUP	Multi Level Reconstruction Module
DATB	Dual Attention Transformer Block
DAL	Dual Attention Layer
GMFF	Gated MLP Feature Fusion
MSFF	Multi Stage Feature Fusion
SeAM	Self-Attention Layer
SpAM	Spatial Attention Module
LR	Low Resolution
HR	High Resolution
NIR	Near Infrared
BI	Bicubic Degradation
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure
LPIPS	Learned Perceptual Image Patch Similarity

VI. CONCLUSION AND FUTURE WORK

This paper introduces TnTViT-G, a novel approach for guidance super resolution based on Transformer architecture. TnTViT-G is designed to accept two images of different domains, extract the information from each domain (infrared and corresponding visible image) with a separate stream, and fuse them efficiently at different stages while remaining memory efficient. We propose a dense feature extraction, which contains both a transformer self-attention layer and a convolutional attention module that can capture both global dependency and local context information at a lower computational cost while its well focusing on spatial features compared to other Transformer models. Furthermore, unlike other GSR methods, TnTViT-G is able to generate SR images in arbitrary sizes, while other methods only generate SR images in fixed sizes. Our experiments highlight that a high-cost, low-resolution spectral image (IR image) can be enhanced by a corresponding high-resolution, low-cost visible image (visible image). We have demonstrated that our approach achieves superior performance compared to other lightweight state-of-the-art methods on all benchmark datasets.

In the future, we will expand our approach for unsupervised guidance super resolution when a paired dataset is not available. To do so, we will attempt to change the methodology of our proposed architecture to use a Generative Adversarial Network. Finally, despite the fact that there has been experimental proof that a low-cost channel can be

TABLE 6. List of datasets used for multimodel guided image super-resolution.

Dataset	Amount	Format	Category Keywords
M3FD	300	PNG	road, campus, street, forest, etc.
RGB-NIR	477	TIFF	country, field, forest, indoor, mountain, old building, street, urban, and water

used to increase the resolution of an expensive channel, this strategy relies on a well-registered paired dataset, which is difficult to obtain since there can be misalignment between multi-model sensors, and a simple feed-forward network cannot deal with the mismatch problem. Thus, the image alignment technique is required as a pre-process to match the counterparts before encoding features of both domains and generating a super-resolved image. Therefore, we will try to integrate the feature alignment method into our forward pass network to address the aforementioned problem.

APPENDIX

In Table 5, we provide a full list of the abbreviations and acronyms, which have been used in this paper. Table 6 contains a list of datasets which has been used to carry out the experiments of this study. We provide the number of HR images, image formats, and category keywords.

REFERENCES

[1] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.

[2] R. Timofte, V. De, and L. V. Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1920–1927.

[3] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467–3478, Aug. 2012.

[4] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3791–3799.

[5] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 184–199.

[6] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[7] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 286–301.

[8] F. Qingyun, H. Dapeng, and W. Zhaokui, "Cross-modality fusion transformer for multispectral object detection," 2021, *arXiv:2111.00273*.

[9] W. K. Wong, H. L. Lim, C. K. Loo, and W. S. Lim, "Home alone faint detection surveillance system using thermal camera," in *Proc. 2nd Int. Conf. Comput. Res. Develop.*, May 2010, pp. 747–751.

[10] A. C. Goldberg, T. Fischer, and Z. I. Derzko, "Application of dual-band infrared focal plane arrays to tactical and strategic military problems," *Proc. SPIE*, vol. 4820, pp. 500–514, Jan. 2003.

[11] J. Baek, S. Hong, J. Kim, and E. Kim, "Efficient pedestrian detection at nighttime using a thermal camera," *Sensors*, vol. 17, p. 1850, Aug. 2017.

[12] B. C. Arrue, A. Ollero, and J. R. M. de Dios, "An intelligent system for false alarm reduction in infrared forest-fire detection," *IEEE Intell. Syst. Appl.*, vol. 15, no. 3, pp. 64–73, May/Jun. 2000.

[13] F. Almasri and O. Debeir, "Multimodal sensor fusion in single thermal image super-resolution," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 418–433.

[14] A. Rogalski, P. Martyniuk, and M. Kopytko, "Challenges of small-pixel infrared detectors: A review," *Rep. Prog. Phys.*, vol. 79, no. 4, Apr. 2016, Art. no. 046501.

[15] Y. Cao, F. Wang, Z. He, J. Yang, and Y. Cao, "Boosting image super-resolution via fusion of complementary information captured by multi-modal sensors," *IEEE Sensors J.*, vol. 22, no. 4, pp. 3405–3416, Feb. 2022.

[16] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 391–407.

[17] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Sep. 2016, pp. 1874–1883.

[18] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[19] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4539–4547.

[20] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.

[21] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 252–268.

[22] A. Muqeet, J. Hwang, S. Yang, J. H. Kang, Y. Kim, and S.-H. Bae, "Ultra lightweight image super-resolution with multi-attention layers," 2020, *arXiv:2008.12912*.

[23] X. Luo, Y. Xie, Y. Zhang, Y. Qu, C. Li, and Y. Fu, "Latticenet: Towards lightweight image super-resolution with lattice block," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K.: Springer, Aug. 2020, pp. 272–289.

[24] P. Behjati, P. Rodriguez, A. Mehri, I. Hupont, C. F. Tena, and J. Gonzalez, "OverNet: Lightweight multi-scale super-resolution with overscaling network," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2694–2703.

[25] X. Chu, B. Zhang, R. Xu, and H. Ma, "Multi-objective reinforced evolution in mobile neural architecture search," 2019, *arXiv:1901.01074*.

[26] X. Chu, B. Zhang, H. Ma, R. Xu, and Q. Li, "Fast, accurate and lightweight super-resolution with neural architecture search," 2019, *arXiv:1901.07261*.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[28] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 558–567.

[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[31] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[32] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–13.

[33] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Decoupled spatial–temporal attention network for skeleton-based action recognition," 2020, *arXiv:2007.03263*.

[34] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Comput. Vis. Image Understand.*, vols. 208–209, Jul. 2021, Art. no. 103219.

[35] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[36] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Efficient transformer for single image super-resolution," 2021, *arXiv:2108.11084*.

[37] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," 2021, *arXiv:2102.04432*.

[38] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," 2021, *arXiv:2106.03106*.

[39] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," 2021, *arXiv:2111.09881*.

[40] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, Jul. 2007.

[41] J. T. Barron and B. Poole, "The fast bilateral solver," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 617–632.

[42] H. Gupta and K. Mitra, "Toward unaligned guided thermal superresolution," *IEEE Trans. Image Process.*, vol. 31, pp. 433–445, 2021.

[43] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, Sep. 2016, pp. 353–369.

[44] Z. Shi, C. Chen, Z. Xiong, D. Liu, Z.-J. Zha, and F. Wu, "Deep residual attention network for spectral image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Sep. 2018, pp. 214–229.

[45] F. Lahoud, R. Zhou, and S. Susstrunk, "Multi-modal spectral image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Sep. 2018, pp. 35–50.

[46] X. Chen, G. Zhai, J. Wang, C. Hu, and Y. Chen, "Color guided thermal image super resolution," in *Proc. Visual Commun. Image Process. (VCIP)*, Nov. 2016, pp. 1–4.

[47] P. Song, X. Deng, J. F. Mota, N. Deligiannis, P. L. Dragotti, and M. R. Rodrigues, "Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 57–72, 2019.

[48] R. d. Lutio, S. D'aronco, J. D. Wegner, and K. Schindler, "Guided super-resolution as pixel-to-pixel transformation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8829–8837.

[49] J. Xie, R. S. Feris, and M.-T. Sun, "Edge-guided single depth image super resolution," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 428–438, Jan. 2016.

[50] D. Zhou, R. Wang, J. Lu, and Q. Zhang, "Depth image super resolution based on edge-guided method," *Appl. Sci.*, vol. 8, no. 2, p. 298, 2018.

[51] T. Xiao, P. Dollar, M. Singh, E. Mintun, T. Darrell, and R. Girshick, "Early convolutions help transformers see better," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 30392–30400.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[53] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," 2021, *arXiv:2111.11418*.

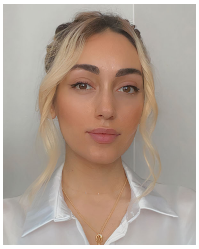
[54] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 5802–5811.

[55] M. Brown and S. Süsstrunk, "Multi-spectral sift for scene category recognition," in *Proc. CVPR*, Jun. 2011, pp. 177–184.

[56] M. Brown and S. Süsstrunk, "Multispectral SIFT for scene category recognition," in *Proc. Comput. Vis. Pattern Recognit.(CVPR)*, Jun. 2011, pp. 177–184.

[57] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

ARMIN MEHRI received the B.Sc. and M.Sc. degrees in computer science from Eastern Mediterranean University, in 2014 and 2017, respectively. He is currently pursuing the Ph.D. degree in deep learning and computer vision with the Computer Vision Center, Universitat Autònoma de Barcelona. He is also the Computer Vision and Deep Learning Lead with Camaleonic Analytics, working on AI-based software for sponsorship in sports. His research interests include computer vision and image processing under cross-modal frameworks resulting in cross-spectral domains.



PARICHEHR BEHJATI received the bachelor's and master's degrees in computer science from Eastern Mediterranean University and the Ph.D. degree (cum laude) in deep learning and computer vision from the Computer Vision Center, Universitat Autònoma de Barcelona. She worked as a Research Assistant with the Computer Science Department, Eastern Mediterranean University, from 2014 to 2016. Her research interests include deep learning and computer vision.

ANGEL DOMINGO SAPPA (Senior Member, IEEE) received the degree in electromechanical engineering from the National University of La Pampa, General Pico, Argentina, in 1995, and the Ph.D. degree in industrial engineering from the Polytechnic University of Catalonia, Barcelona, Spain, in 1999. In 2003, after holding research positions in France, the U.K., and Greece, he joined the Computer Vision Center, Barcelona, where he currently holding a senior scientist position. Since 2016, he has been a Full Professor at ESPOL Polytechnic University, Guayaquil, Ecuador, where he leads the computer vision team at the CIDIS Research Center. He is also the Director of the Electrical Engineering Ph.D. Program. His research interests include cross-spectral image processing and representation, 3D data acquisition, processing, modeling, and computer vision applications.