Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Edge Craft Odyssey: Navigating guided super-resolution with a fast, precise, and lightweight network

Armin Mehri [a], Parichehr Behjati [a], Dario Carpio [b], Angel D. Sappa [a,b,*]

[a] Computer Vision Center, Autonomous University of Barcelona, Barcelona, 08193, Spain
[b] ESPOL Polytechnic University, FIEC-CIDIS, Guayaquil, EC090112, Ecuador

ARTICLE INFO

Keywords:
Guided thermal super-resolution
Lightweight network
Blending technique
Edge extraction

ABSTRACT

Thermal imaging technology is exceptionally valuable in environments where visibility is limited or nonexistent. However, the high cost and technological limitations of high-resolution thermal imaging sensors restrict their widespread use. Many thermal cameras are now paired with high-resolution visible cameras, which can help improve low-resolution thermal images. However, aligning thermal and visible images is challenging due to differences in their spectral ranges, making pixel-wise alignment difficult. Therefore, we present the Edge Craft Odyssey Network (ECONet), a lightweight transformer-based network designed for Guided Thermal Super-Resolution (GTSR) to address these challenges. Our approach introduces a Progressive Edge Prediction module that extracts edge features from visible images using an adaptive threshold within our innovative Edge-Weighted Gradient Blending technique. This technique provides precise control over the blending intensity between low-resolution thermal and visible images. Additionally, we introduce a lightweight Cascade Deep Feature Extractor that focuses on efficient feature extraction and edge weight highlighting, enhancing the representation of high-frequency details. Experimental results show that ECONet outperforms state-of-the-art methods across various datasets while maintaining a relatively low computational and memory requirements. ECONet improves performance by up to 0.20 to 1.3 dB over existing methods and generates super-resolved images in a fraction of a second, approximately 91 % faster than the other methods. The code is available at https://github.com/Rm1n90/ECONet.

1. Introduction

Thermal imaging is highly effective in low visibility conditions, but its high cost and technological limitation restricts its accessibility in consumer applications. Despite these drawbacks, thermal cameras have been used in a wide range of specialized fields such as military affairs [1], surveillance [2], firefighting [3], and multi-object tracking [4].

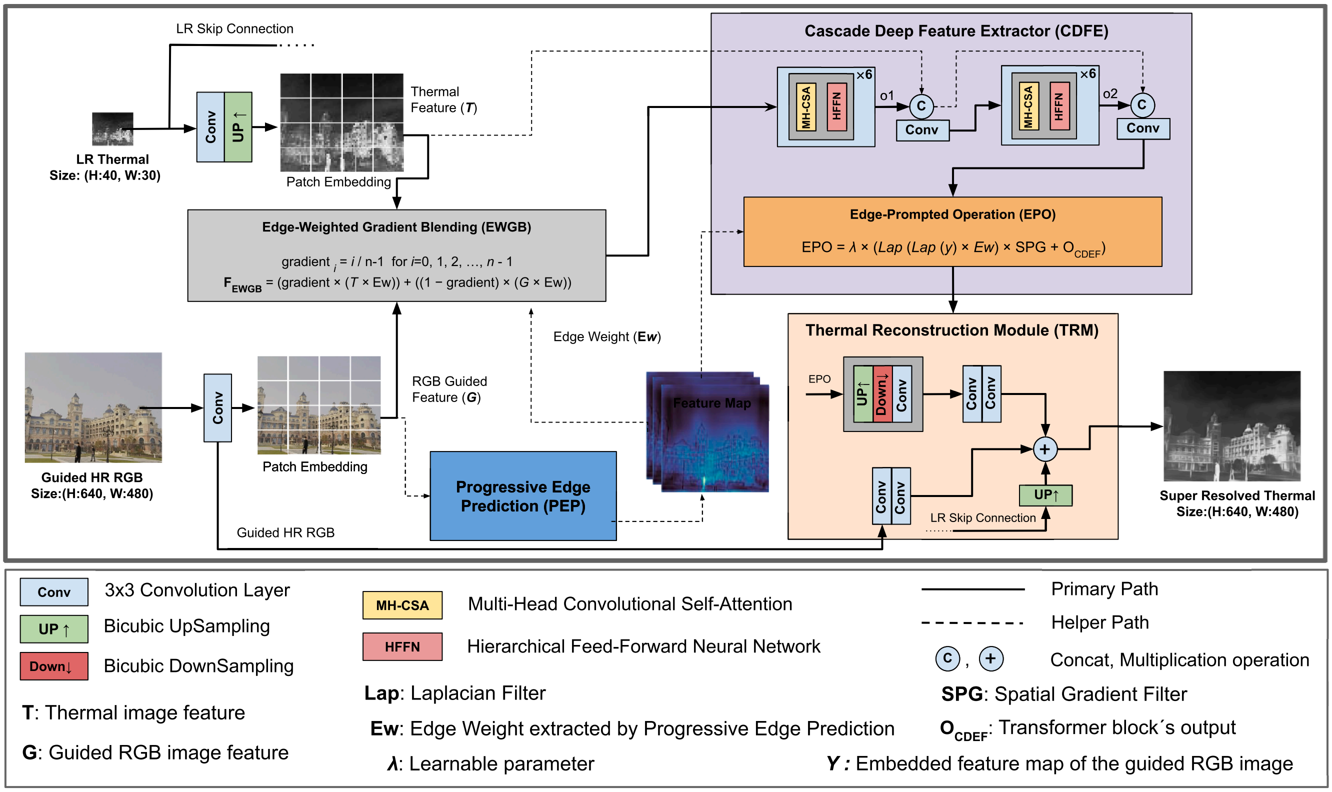
A practical way to expand consumer access to thermal cameras is to employ low-resolution thermal cameras, which offer a more budget-friendly alternative to the high-end models prevalent in specialized fields. Super-resolution (SR) techniques are particularly useful in such scenarios. Motivated by this, numerous single-image super-resolution (SISR) methods have been proposed and have shown strong performance [5,6]. However, when dealing with exceptionally small input images, SISR fails to accurately predict pixels due to the limited availability of high-frequency details. This results in a loss of texture and edges in the super-resolved image.

Recently, Guided Thermal Super-Resolution (GTSR) approaches have demonstrated remarkable performance in addressing low-resolution thermal imagery [7–10]. This success stems from pairing thermal cameras with high-resolution visible-range RGB cameras, which can serve as guiding references for enhancing the quality of low-resolution thermal images. Unlike the SISR task, which relies solely on a single low-resolution image, GTSR leverages the complementary information from high-resolution RGB images of the same scene. These RGB images offer detailed insights into edges and structures, while thermal images provide crucial temperature distribution data. This allows the network to learn how to adaptively select structures to transfer and thus have the ability to handle more complex scenarios.

Despite performance improvements, most existing GTSR methods still have some drawbacks. A critical aspect of SR involves the reconstruction of fine details and textures. Existing GTSR techniques, such as [7,11,12], achieve this with the help of Convolutional Neural Networks (CNNs) and end-to-end learning. However, for texture-rich RGB images, irrelevant edges may be transferred to thermal images (known as texture over-transferred). Additionally, low-resolution thermal images have different spectral characteristics than RGB images, making them difficult to use as guidance for GTSR [7]. These spectral differences between

* Corresponding author.
E-mail addresses: amehri@cvc.uab.cat (A. Mehri), pbehjati@cvc.uab.cat (P. Behjati), dncarpio@espol.edu.ec (D. Carpio), asappa@cvc.uab.cat, asappa@espol.edu.ec (A.D. Sappa).

https://doi.org/10.1016/j.patcog.2025.112392
Received 5 December 2024; Received in revised form 18 August 2025; Accepted 29 August 2025
Available online 2 September 2025



LR Skip Connection
LR Thermal
Size: (H:40, W:30)
Conv
UP ↑
Thermal Feature (T)
Patch Embedding
Edge-Weighted Gradient Blending (EWGB)
gradient $_i$ = i / n-1 for i=0, 1, 2, ..., n - 1
F_{EWGB} = (gradient × (T × Ew)) + ((1 − gradient) × (G × Ew))
Cascade Deep Feature Extractor (CDFE)
MH-CSA
HFFN
×6
o1
o2
C
Conv
Edge-Prompted Operation (EPO)
EPO = λ × (Lap (Lap (y) × Ew) × SPG + O_{CDEF})
Edge Weight (Ew)
Thermal Reconstruction Module (TRM)
EPO
UP↑
Down↓
Conv
+
UP↑
LR Skip Connection
Guided HR RGB
Size:(H:640, W:480)
RGB Guided Feature (G)
Patch Embedding
Progressive Edge Prediction (PEP)
Feature Map
Guided HR RGB
Super Resolved Thermal
Size:(H:640, W:480)
Conv　3x3 Convolution Layer
UP ↑　Bicubic UpSampling
Down↓　Bicubic DownSampling
T: Thermal image feature
G: Guided RGB image feature
MH-CSA　Multi-Head Convolutional Self-Attention
HFFN　Hierarchical Feed-Forward Neural Network
Lap: Laplacian Filter
Ew: Edge Weight extracted by Progressive Edge Prediction
λ: Learnable parameter
Primary Path
Helper Path
C , +　Concat, Multiplication operation
SPG: Spatial Gradient Filter
O_{CDEF}: Transformer block´s output
Y : Embedded feature map of the guided RGB image

Fig. 1. illustration of different components of Edge Craft Odyssey (ECONet).

the two modalities can introduce inconsistencies and inaccuracies during the super-resolution process. As a result, most of the existing GTSR methods struggle to effectively utilize RGB information to enhance the resolution of low-resolution thermal images.

To tackle the aforementioned problems, we first introduce a Progressive Edge Prediction (PEP) module to obtain edge attention weights beneficial for GTSR. This process activates a subset of intensity edges, thereby enabling adaptive transfer of structural textures from visible RGB images. Second, we propose a novel technique called Edge-Weighted Gradient Blending (EWGB). EWGB utilizes the output of PEP as an adaptive threshold to control the blending intensity between low-resolution thermal and RGB images. As a result, the proposed techniques can effectively preserve unique properties such as detailed texture and smooth segmentation across the two modalities. Finally, we present a lightweight Cascade Deep Feature Extractor (CDFE), a transformer-based method [13] that employs adaptive edge enhancement to improve high-frequency features and spatial consistency, thereby balancing feature enhancement and content preservation.

To validate the effectiveness of the proposed approaches, we developed a deep but lightweight architecture for GTSR named Edge Craft Odyssey Network (ECONet), illustrated in Fig. 1. Extensive experiments on the different benchmark datasets demonstrate the superiority of the proposed architecture over state-of-the-art models in terms of qualitative results, visual quality, running time, and memory usage. In summary, the main contributions of this paper are outlined as follows:

- A Progressive Edge Prediction (PEP) module that highlights attention weights in a lightweight and efficient way, allowing the network to focus on more informative features to improve discriminative capabilities.
- A novel blending technique, Edge-Weighted Gradient Blending (EWGB), that computes a linear gradient to control blending intensity between the images. This technique improves the preservation of relevant features while suppressing irrelevant textures.
- A new lightweight Cascade Deep Feature Extractor (CDFE) that introduces adaptive edge enhancement to improve high-frequency features and spatial consistency while keeping the architecture lightweight.

2. Related work

Over the past few decades, researchers have explored a variety of guided super resolution methods, learning mechanisms, and network architectures, each contributing unique advancements to the field. In the following section, we'll focus on the approaches most closely related to our work, highlighting their key contributions and progress in the field.

2.1. Conventional methods

Conventional methods for guided super resolution can be divided into three categories: filter-based, optimization-based, and learning-based methods. Filter-based methods utilize RGB images to guide joint filters to detect the edges in other domains. These methods employ techniques, such as joint bilateral upsampling [14,15], weighted mode filters [16,17], and guided filtering [18]. Optimization-based methods use models such as Markov Random Fields [19], non-local means filtering [20], and others [21] to solve the problem. Learning-based methods include approaches such as bimodal co-sparse analysis [22] and joint dictionary learning [23]. Gu et al. [24] introduced a task-driven learning method for dynamic guidance, while [25] developed an HR edge map inference method from high-resolution and low-resolution image pairs.

2.2. Deep learning-based methods

Recently, Convolutional Neural Networks (CNNs) have achieved unprecedented success in various tasks [26,27]. Choi et al. [28] made significant advances in the field of single thermal image super-resolution

(SR) by introducing a three-layer CNN model, which surpassed the performance of traditional algorithms. Subsequently, [29] integrated compressive sensing with deep learning techniques to enhance infrared SR capabilities. These advances in improving thermal images have led to the development of several methods for enhancing near-infrared image quality [30,31]. However, these methods often neglect super-resolution for extremely low-resolution inputs, such as those from affordable thermal cameras. Later, [28] and [32] proposed that employing visible SR methods or leveraging pre-trained models could yield favorable results for thermal images. A great number of deep CNN-based single-image SR techniques [6,33–36] have demonstrated remarkable efficacy on visible images in different areas. For example, MAda-SR [37] enhances medical image super-resolution across diverse modalities using an adaptive optimizer and weighted MSE loss, achieving superior stability. EDiffSR [38] employs a diffusion probabilistic model with an efficient activation network and a conditional prior enhancement module to achieve high-quality, computationally efficient super-resolution for remote sensing images. PLGNet [39], a two-stage prior-guided face super-resolution approach, leverages a hybrid network with multiscale feature extraction and prior interaction modules to enhance facial reconstruction accuracy and visual quality. TTST [40] improves transformer-based image super-resolution for earth observation by adaptively selecting crucial tokens with RTSG, integrating multi-scale features via MFL, and leveraging GCA. All these methods suggest their potential applicability to thermal imagery. Nevertheless, a common concern with single-image methods is their ability to accurately reconstruct high-resolution (HR) images from low-resolution (LR), noisy sensor inputs.

Numerous GTSR models have been proposed to alleviate the aforementioned problems [7,10,29]. For example, [41] proposed a multi-modal sensor fusion model to enhance the thermal images with the help of RGB images. [42] proposed a U-Net-based model architecture incorporating two encoders to capture both thermal and visible information. [10] introduced a Transformer in the Transformer network, called TnTViT-G, an efficient method to extract image features hierarchically and fuse them at different architectural levels. Meanwhile, other methods, such as [7,43,44], proposed using edges of the visible image to produce high-frequency details. Edge-based guidance helps reconstruct better high-frequency details. However, these methods assume pixel-to-pixel alignment between the low-resolution measurement and the guide image, which is difficult to achieve in the case of thermal images due to significant mismatches between low-resolution thermal images and high-resolution visible images. To this end, [45] proposed an edge attention mechanism to highlight the contours informative for guided upsampling. Gupta and Mitra [9] proposed a misalignment-map estimation block as a part of an end-to-end framework that adequately aligns the input images for performing guided super-resolution.

In contrast to previous approaches, our proposed method introduces a more efficient strategy to handle misalignment between feature maps extracted from the input and guide images enhancing performance and accuracy in GTSR.

3. Proposed method

Guided Thermal Super-Resolution aims to predict a high-resolution thermal image $\mathcal{H} \in \mathbb{R}^{M\times N}$ from a low-resolution thermal image $\tilde{\mathcal{T}} \in \mathbb{R}^{m\times n}$ with the guidance of the visible image $\mathcal{G} \in \mathbb{R}^{M\times N\times 3}$ of the same scene. Here, $M \times N$ and $m \times n$ represent the height and width of the input guided and thermal images, respectively. To perform GTSR, we first preprocess $\tilde{\mathcal{T}}$ to obtain $\mathcal{T} \in \mathbb{R}^{m\times n\times 3}$ by expanding its channel dimensions to match those of HR RGB-guided image. The $ECONet(\cdot)$ processes both images by taking into account the interactions between two different modalities. In the next section, we will explain the ECONet in detail.

3.1. ECONet: Edge Craft Odyssey Network

ECONet is designed to perform guided thermal super-resolution (GTSR) by integrating information from a low-resolution thermal image and a high-resolution RGB-guided image. The architecture is composed of several interconnected modules that collaboratively enhance the resolution and quality of the thermal image while preserving critical edge and texture details. The key components of ECONet are: Progressive Edge Prediction (PEP), Edge-Weighted Gradient Blending (EWGB), Cascade Deep Feature Extractor (CDFE) with Edge-Prompted Operation (EPO), and Thermal Reconstruction Module (TRM). Given the low-resolution thermal image $\mathcal{I}^{\mathcal{T}}_{LR}$ and the high-resolution guided RGB image $\mathcal{I}^{G}$, we first apply a 3×3 convolution layer followed by bicubic upscaling to $\mathcal{I}^{\mathcal{T}}_{LR}$, and a 3×3 convolution layer to $\mathcal{I}^{G}$. The resulting feature maps are then passed through a patch embedding layer to obtain $\mathcal{I}^{\mathcal{T}}_{emb}$ and $\mathcal{I}^{\mathcal{G}}_{emb}$, respectively:

$$\mathcal{I}^{\mathcal{T}}_{emb} = \Phi\Big(\mathcal{U}^{\uparrow}_{\text{bic}}\Big(\mathcal{C}(\mathcal{I}^{\mathcal{T}}_{LR})\Big)\Big), \quad \mathcal{I}^{\mathcal{G}}_{emb} = \Phi\big(\mathcal{C}(\mathcal{I}^{G})\big) \tag{1}$$

where Φ denotes the patch embedding operation as used in Vision Transformers [46], $\mathcal{C}$ is a convolution layer, and $\mathcal{U}^{\uparrow}_{\text{bic}}$ is bicubic upscaling. This step ensures that both embeddings have the same spatial resolution, mapping the input images to a high-dimensional feature space suitable for subsequent processing.

3.1.1. Progressive edge prediction

Zhao et al. [45] introduced Guided Weight Prediction (GESA) to mitigate overly strong texture transfer from RGB-guided images. While effective to some extent, GESA lacked the capacity to capture fine and abstract edge features. To address this, we propose Progressive Edge Prediction (PEP).

The PEP module takes the guided RGB image $\mathcal{I}^{G}$ as input and produces an edge attention map **Ew**. The module is designed to extract and refine edge information from the high-resolution RGB-guided image, which is crucial for guiding thermal super-resolution. By focusing on edges—regions of significant intensity transitions, the module enhances boundary and texture details that are essential for high-quality reconstruction. The architecture of the PEP module employs a series of convolutions with varying kernel sizes and dilation rates to capture edge details across multiple scales. Small kernels target fine, localized edges, while larger and dilated kernels expand the receptive field to detect broader structures. These features are aggregated to form a comprehensive edge representation. The module operates through a progressive refinement process, iterating over multiple blocks. In each block, features are downsampled to capture larger-scale edges, refined via convolutional layers, and then upsampled back to the original resolution using bicubic interpolation. Skip connections integrate the original input features with the refined outputs at each stage, preserving high-frequency details that might otherwise be lost during the downsampling and upsampling processes. The final refined features are concatenated, passed through a convolutional layer, and activated with a sigmoid activation function to produce an edge attention map in the range [0, 1]. A key aspect of the PEP module is its adaptive thresholding mechanism, facilitated by the sigmoid activation. This function maps the aggregated edge features to a continuous range, effectively learning a threshold during training. Unlike fixed thresholding methods, this adaptive approach dynamically adjusts the sensitivity of edge detection based on the input data and the specific requirements of the super-resolution task. The PEP module can be formulated as follows:

$$\tilde{E}_w = \mathcal{P}\big(\mathcal{I}^{G}\big) \in \mathbb{R}^{M\times N\times C}. \tag{2}$$

where $\mathcal{P}$ represents the operations within the PEP module, as illustrated in Fig. 2. The resulting edge attention map **Ew** is subsequently used in the EWGB and EPO modules to emphasize edge-related features, ensuring that its spatial resolution matches that of the embedded features.

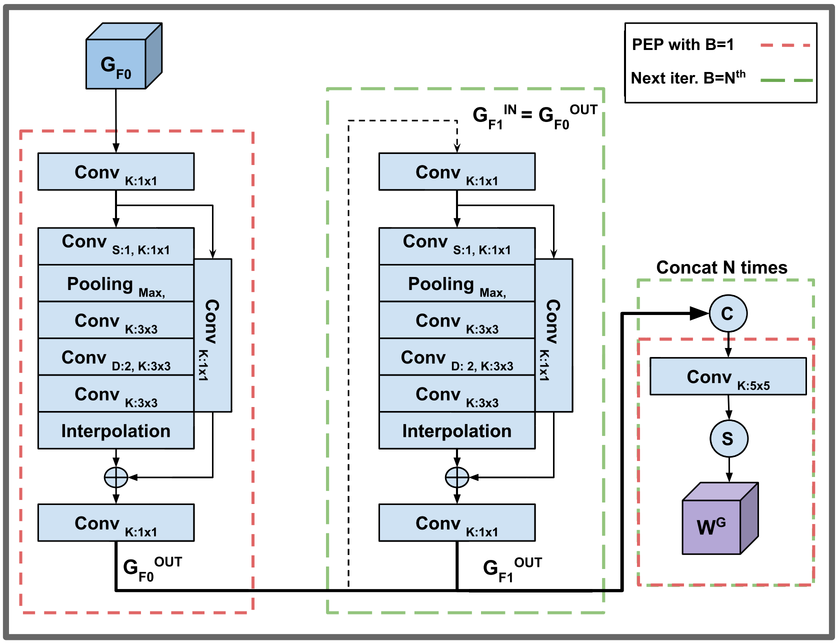


Fig. 2. Details illustration of Progressive Edge Prediction module (PEP).

3.1.2. Edge-weighted gradient blending

A central challenge in GTSR is fusing information from different modalities while retaining the distinct characteristics of the first modality while taking advantage of the RGB-guided image without empowering it during the process. This is naturally challenging since the input images come from different domains and present variations in terms of resolution and quality. However, the RGB-guided and thermal images obtained from the same scenario can provide various types of valuable information that can be shared or correlated between the two modalities, such as information related to the shape and edges of the objects. Therefore, we present EWGB as a novel gradient-based blending strategy for integrating two images from different modalities, focusing primarily on adaptively highlighting edges during the process. The EWGB can be represented as follows:

$$WEGB = \mathrm{G}_r \times (\mathcal{I}^{T}_{emb} \times \mathrm{Ew})) + \mathrm{G}_r \times (\mathcal{I}^{G}_{emb} \times \mathrm{Ew})), \tag{3}$$

where **Ew** denotes the edge attention map obtained by the PEP module. $\mathbf{G}_r$ denotes linear gradient vector $\mathrm{G}_r = [g_0, g_1, \dots, g_{n-1}]$ where $g_i = \frac{i}{n-1}$ for $i = 0, 1, \dots, n-1$. $\mathbf{G}_r$ is linearly interpolated between 0 and 1 across the channel dimensions, which acts as a weighting factor that determines the contribution of pixel values from two different modalities, allowing direct control of the blending ratio at each pixel along a dimension between regions in feature space. Unlike other methods such as summation or alpha blending, which treat all pixels equally, EWGN effectively combines the details of the input images by using the rich boundary information encoded in their gradients. This highlights the regions with a high gradient intensity that are often associated with edges and discontinuities in the images.

3.1.3. Cascade deep feature extractor

The CDFE module processes the blended feature $\mathcal{F}_{\mathrm{EWGB}}$ through a series of transformer-based blocks, each comprising a Multi-Head Convolutional Self-Attention (MH-CSA) and a Hierarchical Feed-Forward Neural Network (HFFN). These blocks, inspired by Restormer [47], efficiently encode local and global dependencies at a low computational cost.

The multi-Dconv head transposed attention (MDTA) in [47] is an efficient self-attention layer with multiple heads, where each head handles a specific subset of embedding dimensions. However, this can negatively impact the network's performance, particularly when the token embedding dimension is insufficient. In such cases, the dot product of the query and key becomes ineffective as a discriminative capability. Therefore, we introduced Multi-Head Convolutional Self-Attention (MH-CSA) and Hierarchical Feed-Forward Neural Network (HFFN) with key modifications to address the problem.

Multi-Head Convolutional Self-Attention (MH-CSA) Generates query (**Q**), key (**K**), and value (**V**) projections matrix by using 1×1 followed by 3×3 depth-wise convolutions. Then, restructure the queries and key projections so that their interaction via dot product produces a transposed-attention map **A** with dimensions $\mathbb{R}^{\hat{C} \times \hat{C}}$. The attention map is then passed to a conv layer before applying softmax, which can be formulated as follows:

$$\mathrm{Attention}(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) = \hat{\mathbf{V}} \cdot \mathrm{Softmax}(C_{1\times1}(\hat{\mathbf{K}} \cdot \hat{\mathbf{Q}}/\alpha)), \tag{4}$$

here, we introduce a 1×1 convolutional layer to model the interactions among different heads to enhance the discriminative power of the attention weights, which leads the attention map of each head to depend on all the keys and queries. As a result, this helps to better differentiate between relevant and irrelevant parts of the input patches. Also, additional depth-wise convolution is applied to (**V**) projections to allow the network to encode the information across spatial positions adaptively.

Hierarchical Feed-Forward Neural Network (HFFN) Is a FFN that introduces an improvement over traditional feed-forward networks in transformer blocks [10,46]. Previously proposed FFNs failed to perform well in the super-resolution task, which requires high-frequency detail reconstruction and maintaining spatial coherence. Therefore, HFFN is designed to address the limitations by learning diverse representations across different levels to recover accurate structural information. HFFN facilitates a better understanding of various aspects of the feature by integrating hierarchical processing and spatial context awareness into the core design of FFN. The HFFN can be represented as follows:

$$\begin{aligned} H_i &= GELU\big(C_{3\times3}\big(GELU\big(C_{1\times1}(X)\big)\big)\big), \quad \text{for } i = 1, \dots, L, \\ Y &= C_{1\times1}\big(\mathrm{Concat}\big(H_1, H_2, \dots, H_L\big)\big), \end{aligned} \tag{5}$$

where $GELU$ stands for Gaussian Error Linear Unit activation function [48]. L and $Concat$ denote the number of levels in the hierarchical structure and the concatenation operation along the channel dimension, respectively.

Each block of the proposed transformer then concatenates with unmodified information before passing to a 1×1 convolution layer to the Edge-Promoted Operation.

3.1.4. Edge-prompted operation

Constructs thermal feature maps with enhanced boundaries and edges by aggregating multi-channel information obtained from $\mathcal{I}^{\mathcal{G}}_{emb}$, O_{CDEF}, PEP (Ew), and Spatial Gradient operation. The EPO can be seen as follows:

$$EPO = \lambda \cdot (\mathcal{L}(\mathcal{L}(\mathcal{I}^{G}_{emb}) \odot Ew) \cdot \|\nabla G(O_{CDEF})\| + O_{CDEF}), \tag{6}$$

where $\mathcal{L}(\cdot)$ is the Laplacian filter by convolving with $\mathcal{I}^{G}_{emb}$ to highlight edges and sharp intensity variations, which correspond to object boundaries, texture edges, and other important features. $\|\nabla G(O_{CDEF})\|$ denotes the magnitude of the Spatial Gradient filter on O_{CDEF}, calculated as $\sqrt{G_x^2 + G_y^2}$, where G_x^2 and G_y^2 are the horizontal and vertical components of the gradient. λ is a learnable parameter that scales the contribution of the enhanced perceptual feature component. The EPO approach brings significant advantages by emphasizing perceptual clarity, maintaining detail, and adaptive edge enhancement. Therefore, EPO improves high-frequency features and spatial consistency by using Laplacian and spatial gradient operations to sharpen high-frequency details such as edges and textures.

3.1.5. Thermal reconstruction module

Finally, the thermal reconstruction module (TRM) predicts the high-resolution thermal image by combining the feature maps derived from the feature integration of three parallel paths (EPO, $\mathcal{F}^{G}$, and a skip connection). First, the extracted feature map from the CDFE pass-through set of operations can be represented as:

$$\begin{aligned}\mathcal{F}_{\text{EPO}} &= \mathcal{C}_w\Big(\mathcal{D}^{\downarrow}_{\text{bic}}\Big(\mathcal{U}^{\uparrow}_{\text{bic}}\big(\mathcal{F}_{\text{EWGB}}\big)\Big)\Big),\\ \mathcal{F}_{\text{out}} &= \mathcal{C}_w\big(\mathcal{R}\big(\mathcal{C}_w\big(\mathcal{F}_{\text{EPO}}\big)\big)\big),\end{aligned} \tag{7}$$

where $\mathcal{U}^{\uparrow}_{\text{bic}}$ and $\mathcal{D}^{\downarrow}_{\text{bic}}$ represent bicubic upscaling and downscaling operations, respectively. $\mathcal{F}_{EPO}$ denotes the feature obtained from the Edge-Prompted Function, whereas $\mathcal{R}$ stands for the LeakyReLU activation function. By doing so, the network enhances its discriminate ability to accurately enhance the pixels through different levels of upscaling, bringing it closer to its high-resolution counterpart while producing less noise and artifacts. The TRM module can be formulated as follows:

$$\mathcal{I}_{SR} = TRM(\mathcal{F}_{EPO} + \mathcal{F}_{G} + \mathcal{U}^{\uparrow}_{bic}(\mathcal{I}^{\mathcal{T}}_{LR})) \in \mathbb{R}^{M\times N}, \tag{8}$$

where $\mathcal{I}_{SR}$ denotes the final super-resolved image. $\mathcal{F}_{G}$ stands for the obtained feature map from the RGB-guided image, followed by two convolution layers and activation operations. $\mathcal{U}^{\uparrow}_{bic}(\mathcal{I}^{\mathcal{T}}_{LR})$ indicates the direct upsample connection on the low-resolution thermal image. The detailed investigation of our TRM is given in the ablation study section.

3.1.6. Training loss

Unlike prior works [7,10,45], which usually use MAE or MSE as a loss function, we propose a custom loss function built up by a combination of multiple weighted-loss terms that noticeably improve the learning capability of the network. The custom loss function $Loss_{final}$ can be formulated as:

$$Loss_{final} = \alpha_1 Loss_{mae} + \alpha_2 Loss_{ssim} + \alpha_3 Loss_{grad}, \tag{9}$$

where $Loss_{mae} = \frac{1}{\text{n}} \sum_{n}^{i=1} ||I_{sr} - I_{hr}||$, $Loss_{ssim} = 1 - SSIM(I_{sr} - I_{hr})$, and $Loss_{grad} = \frac{1}{HW} \sum_k |\nabla I_{sr} - \nabla I_{hr}|$ stand for the mean absolute error, the structural similarity index, and the spatial gradient loss function between the super-resolved image and ground truth, respectively. ∇ denotes the spatial gradient, computed using the Sobel and convolution operators. The α_1, α_2, and α_3 are the hyperparameters to tune the weight of each loss function. Using the $Loss_{final}$ loss function, the network generates enhanced and sharper reconstruction images, reducing pixel-level mismatch while helping to highlight edge details, resulting in perceptually superior SR images.

4. Experiments

We conduct extensive evaluations on multiple datasets using both quantitative and qualitative metrics to demostrate the superiority of ECONet. We compare our ECONet against the existing SOTA of guided super-resolution methods, including a simple non-guided bicubic upsampling technique as a baseline. The evaluated methods include: SVLRM [23], JIIF [49], DKN [50], FDKN [50], DAGF [51], PMBANet [43], DCTNet [45], and D2A2 [52]. All models were retrained using their open-source code and default settings for scale factors ×8, ×16, and ×32.

4.1. Experimental setup

4.1.1. Datasets

We evaluate ECONet on two thermal datasets. M3FD [53], and CIDIS [53]. The M3FD dataset consists of 870 paired images. We used 820, 30, and 20 images for the training, validation, and test phases. The CIDIS dataset contains 940 paired images, split into 700 and 200 for the training and validation phases and 40 for the test phase.

4.1.2. Evaluation metric

We employ three widely used evaluation metrics to assess the quality of the super-resolved results of our ECONet compared with other approaches—i.e., peak signal-to-noise ratio (PSNR), Structural Similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [54] with VGG pre-trained network. Higher values indicate better performance for PSNR and SSIM, while lower values are preferable for LPIPS.

4.1.3. Implementation detail

The training samples are randomly selected with a batch size set to 8. The number of training steps is set to $60K$. We use the Adam

Table 1
Quantitative comparison between our proposed ECONet and state-of-the-art approaches on M3FD and CIDIS benchmark datasets for scale factor [×8, ×16, ×32]. We use the three metrics [PSNR, SSIM, LPIPS]. The best and the second-best values are **highlighted** and $\underline{\text{underline}}$, respectively.

Scale	Model	M3FD PSNR/SSIM/LPIPS	CIDIS PSNR/SSIM/LPIPS
×8	Bicubic	25.64/0.7696/0.4579	26.02/0.7867/0.4464
	SVLRM (2021) [23]	26.52/0.7992/0.3915	23.86/0.7434/0.4059
	JIIF (2021) [49]	27.75/0.8441/$\underline{0.3028}$	21.68/0.7401/0.3768
	DKN (2021) [50]	29.09/0.8277/0.3626	23.49/0.7504/0.3919
	FDKN (2021) [50]	28.81/0.8180/0.3783	23.33/0.7464/0.3985
	PMBANet (2020) [43]	28.40/$\underline{0.8510}$/0.3081	$\underline{28.81}$/$\underline{0.8681}$/$\underline{0.3165}$
	DCTNet (2022) [45]	28.30/0.8219/0.3811	24.59/0.7777/0.3949
	DAGF (2023) [51]	**29.54**/0.8382/0.3469	26.76/0.8237/0.3247
	D2A2 (2024) [52]	26.82/0.8293/0.3136	26.64/0.8341/0.3272
	ECONet (ours)	$\underline{29.38}$/**0.8763**/**0.2717**	**29.02**/**0.8788**/**0.3001**
×16	Bicubic	22.43/0.7056/0.5818	21.66/0.6824/0.5905
	SVLRM (2021) [23]	23.22/0.7253/0.4727	20.26/0.6697/0.4799
	JIIF (2021) [49]	25.47/0.7800/$\underline{0.3618}$	19.46/0.6662/0.4311
	DKN (2021) [50]	24.58/0.7510/0.4210	23.10/0.7014/0.4244
	FDKN (2021) [50]	25.29/0.7427/0.4705	22.76/0.6860/0.4705
	PMBANet (2020) [43]	24.85/0.7728/0.4137	$\underline{24.41}$/$\underline{0.7737}$/$\underline{0.4041}$
	DCTNet (2022) [45]	25.70/0.7632/0.4533	21.13/0.6907/0.4670
	DAGF (2023) [51]	**26.63**/$\underline{0.7823}$/0.4209	23.52/0.7269/0.3976
	D2A2 (2024) [52]	24.30/0.7766/0.3890	23.51/0.7700/0.3904
	ECONet (ours)	$\underline{26.51}$/**0.8280**/**0.3365**	**24.79**/**0.7911**/**0.3704**
×32	Bicubic	19.76/0.6705/0.6250	21.29/0.6983/0.6276
	SVLRM (2021) [23]	18.64/0.6411/0.5177	19.87/0.6716/0.5019
	JIIF (2021) [49]	**24.72**/$\underline{0.7633}$/$\underline{0.3698}$	21.80/0.7232/0.4633
	DKN (2021) [50]	23.74/0.7271/0.4649	21.79/0.6988/0.4759
	FDKN (2021) [50]	22.95/0.7056/0.5122	21.49/0.6925/0.5294
	PMBANet (2020) [43]	–/–/–	–/–/–
	DCTNet (2022) [45]	23.23/0.7119/0.4997	20.65/0.7044/0.4873
	DAGF (2023) [51]	24.35/0.7463/0.4654	$\underline{22.16}$/0.7230/0.4580
	D2A2 (2024) [52]	22.71/0.7495/0.4171	22.14/$\underline{0.7435}$/$\underline{0.4449}$
	ECONet (ours)	$\underline{24.43}$/**0.8042**/**0.3409**	**23.49**/**0.7644**/**0.4447**

optimizer and multi-step decay, with the initial learning rate set to 10^{-3} and decreasing the learning rate at steps $35K$ and $50K$. For the network hyper-parameters setting, the number of transformer blocks in CDFE is set to 2, with 6 attention layers, and 4 attention heads. The dimensions are set to 64 for the entire network. Following to [45], Lambda is a learnable parameter randomly initialized, where $0 \sim N(0.1, 0.3)$. The $\alpha_1, \alpha_2, \alpha_3$ are set to $3, 10,$ and 5, respectively. ECONet is implemented using the PyTorch framework.

4.2. Comparison with the state-of-the-arts

4.2.1. Qualitative comparison

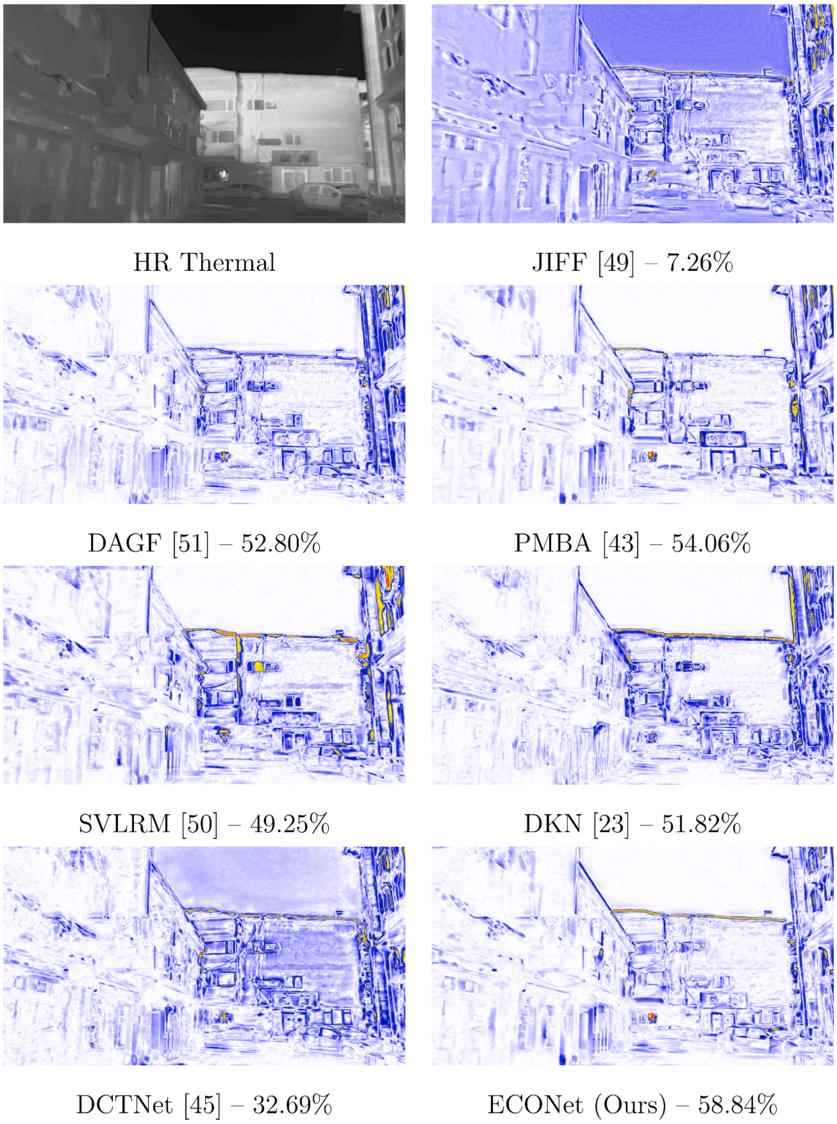
Table 1 reports the quantitative results on two benchmark datasets for scale factors of ×8, ×16, and ×32. As can be seen, ECONet shows superior performance, generalizing well across both datasets. In contrast, other approaches tend to perform well only on specific datasets or scale factors but fail to generalize effectively. ECONet achieves the best performance on all perceptual metrics and either the best or second-best results on pixel-wise metrics across various datasets and scale factors. It is worth mentioning that the PMBANet [43] and DAGF [51] have almost $113, 114K$ and $2, 440K$ parameters, respectively, while ECONet uses only $948K$ parameters. This highlights ECONet's ability to achieve superior performance with significantly fewer parameters.

4.2.2. Error maps visualization

Fig. 3 shows the error maps of several methods on the CIDIS dataset at scale factor ×16. The error maps are color-coded: white, and blue, indicate lower error values, while yellow, orange, and red represent higher errors. The ECONet produces lower error rates, shown by visualizing more white and blue colors, compared to other methods, which demonstrate higher error rates, shown by the greater presence of yellow, orange, and red colors. This indicates that the ECONet can enhance the visual perception of the image while minimizing reconstruction errors and preserving distinctive characteristics of thermal imagery.

4.2.3. Quantitative comparison

Fig. 4 presents visual results. ECONetsubstantially improves the reconstruction of super-resolved thermal images by integrating LR thermal data with structural cues from visible images. As can be seen, the objects located in areas with low light intensity are clearly illuminated and allow easy differentiation between foreground objects and the surrounding environment. Moreover, the background details that



HR Thermal

JIFF [49] – 7.26%

DAGF [51] – 52.80%

PMBA [43] – 54.06%

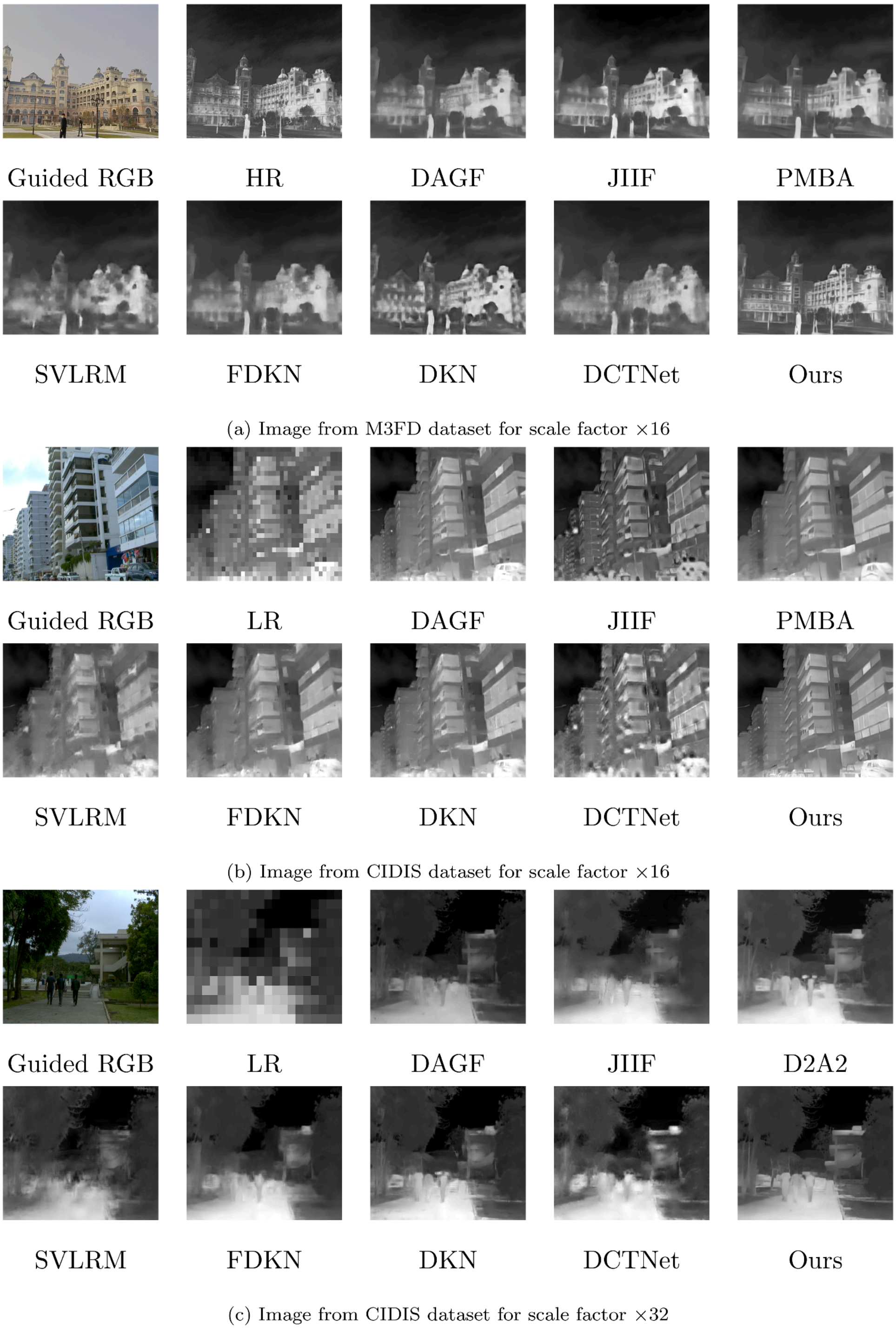
SVLRM [50] – 49.25%

DKN [23] – 51.82%

DCTNet [45] – 32.69%

ECONet (Ours) – 58.84%

Fig. 3. Visual comparison of error maps on the M3FD dataset for ×16 super-resolution. The error maps are shown in different color ranges: white, blue, yellow, orange, and red, representing the error range from zero to one. A higher proportion of white pixels indicates fewer errors (zoom in and see in color for the best view).



(a) Image from M3FD dataset for scale factor ×16

(b) Image from CIDIS dataset for scale factor ×16

(c) Image from CIDIS dataset for scale factor ×32

Fig. 4. Visual results on both the M3FD and CIDIS datasets. Note: The test set for the CIDIS dataset was not available (zoom in for best view).

are challenging to identify due to the insufficient illumination have distinct boundaries and enough contour information, thereby improving the overall understanding of the image, particularly when it comes to larger-scale factors that cause difficulties for other state-of-the-art methods.

5. Ablation study

We conduct ablation experiments to analyze the contribution of each module in ECONet. Best and second-best values in the tables are **highlighted** and underlined respectively.

5.1. Qualitative comparison with PBVS challenge winners

Table 2 compares the performance of various models, including AIR, GUIDEDSR, UMKC MCC, and VISION IC, the winners of the PVBS 2024 challenge [55], and our proposed method ECONet, across two scale factors (×8 and ×16) using the PSNR and SSIM metrics. Although ECONet achieves slightly lower PSNR and SSIM values than other models, such as GUIDEDSR and UMKC MCC, it significantly outperforms them in terms of model efficiency and compactness. For instance, ECONet has only $948K$ parameters, which is approximately **632** times fewer than GUIDEDSR, which has a massive $600M$ network parameters. Despite this

Table 2
Comparison between ECONet and PBVS 2024 challenge winners.

Scale	Model	Params	CIDIS	
			PSNR	SSIM
×8	AIR	3.4M	–	–
	GUIDEDSR	600M	31.52	0.9127
	UMKC MCC	12.17M	30.05	0.8947
	VISION IC	3.30M	29.34	0.8824
	ECONet(ours)	**948K**	29.02	0.8788
×16	AIR	3.4M	24.77	0.7878
	GUIDEDSR	600M	25.99	0.8266
	UMKC MCC	12.17M	25.67	0.8167
	VISION IC	3.30M	24.69	0.7928
	ECONet(ours)	**948K**	24.79	0.7911

Table 3
Investigation on the impact of EWGB on network performance for scale factor ×16. The best and second-best values are **highlighted** and <u>underlined</u>.

Setting	M3FD		CIDIS	
	PSNR	SSIM	PSNR	SSIM
Sum	26.18	0.7920	24.42	0.7716
ConCat	<u>26.36</u>	<u>0.8115</u>	<u>24.58</u>	<u>0.7841</u>
Alpha Blending	25.97	0.7853	24.21	0.7689
EWGB (ours)	**26.51**	**0.8280**	**24.79**	**0.7911**

dramatic reduction in complexity, ECONet achieves PSNR of 29.02 and SSIM of 0.8788 at scale ×8, 24.79 and 0.7911 at scale ×16. This demonstrates that ECONet is a highly efficient and lightweight model, making it practical for real-world applications where heavy models such as GUIDEDSR or UMKC MCC are impractical due to their resource-intensive requirements.

5.2. Impact of EWGB

Table 3 evaluates the proposed Edge-Weighted Gradient Blending (EWGB) against other techniques, i.e., summation, concatenation, alpha-blending, and weighted alpha-blending. Unlike other methods, EWGB is proposed using the edge information obtained from the proposed PEP module. This allows for a more fine-blending process, focusing on edge features while ensuring smooth transitions from the LR thermal image to the HR RGB-guided image. As illustrated in Table 3, ECONet achieves superior outcomes by using EWGB with about **0.18** dB performance boosts compared to ConCat operations.

5.3. Effect of progressive edge prediction

We evaluate the effect of the proposed Progressive Edge Prediction (PEP) compared with existing methods and without any edge-guided modules as a baseline on the CIDIS dataset. As shown in Table 4, the proposed method performed poorly without using any of the modules (baseline). We can see that the proposed PEP enhances the network performance by $0.22dB$ and $0.16dB$ compared to GESA and baseline, respectively. Unlike GESA, PEP uses progressive feature extraction techniques that enable access to abstract edge features by iteratively concatenating intermediate feature maps across multiple levels. Therefore, PEP maximizes the ability of the network to access the deeper feature maps, which contain complex patterns and high-level representations, without adding significant computational costs.

Table 4
ECONet performance on different edge-guided modules for scale ×16.

	Baseline	GESA	PEP	Params (K)	PSNR (dB)
ECONet	✓			784	24.57
		✓		801	24.63
			✓	948	**24.79**

Table 5
Performance impact of different settings on the CIDIS dataset.

#	Settings	Params(K)	X16	X32
1	w/o EPO	–	24.61	23.29
1	w/o TRM	836	24.28	23.96
2	w/o $C(Up/Down)$	871	24.67	23.38
3	w/o LR Skip-Connection	–	24.71	23.44
4	w/o $L_{ssim} + L_{grad}$	–	24.63	23.36
	ECONet	948	**24.79**	**23.49**

5.4. Setting investigation on ECONet

Extensive evaluations were performed to confirm the design choices of our ECONet. Various experiments were executed to evaluate the impact of each component of our network. As shown in Table 5, **#1**, ECONet performance decreased by almost $0.18dB$ without using EPO, which shows the effectiveness of our EPO in constructing the multi-channel feature with enhanced edge and boundary features. **#2**, We carried out an experiment using the ECONet without the TRM module and each component of TRM. We reconstructed the SR image directly after the CDFE module without the TRM module. The results show that the low-cost TRM module can boost the network by $0.51dB$, adding only $112K$ parameters. **#3**, A simple yet effective technique in the TRM module can help the network increase its performance. Upscaling, downscaling, and a convolution operation in the TRM can bring $0.12dB$ performance gain while being almost cost-free. By doing so, the network can generalize better to unseen features and give the network a bigger vision than it needs. **#4**, The results from settings 4 demonstrate the impact of the proposed loss terms. The results indicate that the proposed custom loss can bring marginally satisfactory performance gains, while only using $L1$ or $L2$ loss functions can produce poor results. **#6**, we also removed the LR skip connection integrated with bicubic upscaling operations from ECONet. The results show that residual connections are important in boosting the network's overall performance.

5.5. Model complexity and running time

Table 6 compares ECONet with other SOTA methods in terms of model complexity (K), running time (s), and memory usage (MB) on the CIDIS dataset for ×16. To provide a fair comparison, all methods were tested using the same setup, with their public source code and default hyper-parameters, on an Intel Core i9-10900K CPU and an NVIDIA RTX 3090 GPU. The proposed method generates super-resolved images significantly faster than other SOTA approaches. This comparison shows that our proposed model properly balances performance and

Table 6
Average running time (s) and memory consumption (MB) comparison on CIDIS dataset for ×16.

Methods	Parameters(K)	Memory(MB)	Running Time(s)	PSNR(dB)
JIFF [49]	10831	7999	0.8929	19.46
DKN [50]	1160	16352	0.4587	23.10
FDKN [50]	693	3611	0.3571	22.76
DAGF [51]	2440	7974	0.2564	23.52
PMBANet [43]	113144	5053	0.2019	24.41
DCTNet [45]	483	3486	0.3322	21.13
ECONet (ours)	948	**1824**	**0.0280**	**24.79**

running-time requirements, which makes it suitable for real-time applications or devices with low computation capacity.

6. Conclusions and future work

In this paper, we introduce ECONet, a novel Edge Craft Odyssey Network designed for Guided Thermal Super-Resolution (GTSR), addressing challenges arising from spectral disparities between low-resolution thermal and high-resolution RGB images. Through the integration of the Progressive Edge Prediction (PEP) module, Edge-Weighted Gradient Blending (EWGB) technique, and Cascade Deep Feature Extractor (CDFE), ECONet enhances the resolution of low-quality thermal images while preserving essential features and suppressing irrelevant textures. These modules work together to exploit edge information, control blending intensity, and extract deep features, thereby improving the overall quality and visual fidelity of super-resolved thermal images. A series of ablation experiments demonstrated the effectiveness of the proposed components. Empirical results confirmed that our method reconstructs high-frequency details, preserves object structures, and outperforms state-of-the-art GTSR approaches in both distortion and perceptual metrics, while maintaining a lightweight and efficient architecture.

In the future, we plan to address challenges in guided super-resolution when low-quality thermal images and high-quality RGB-guided images are misaligned, a frequent issue in multi-sensor systems. Such misalignment disrupts pixel-wise correspondence between the two modalities, reducing the effectiveness of conventional super-resolution approaches. To overcome this, we aim to develop blind-guided super-resolution techniques that do not depend on precise sensor alignment. Additionally, we will investigate the use of generative adversarial networks (GANs) to mitigate registration issues. Our ultimate goal is to enhance the robustness of our methods in real-world scenarios, where perfect image alignment is rarely achievable, thereby ensuring stronger performance and broader practical applicability.

CRediT authorship contribution statement

Armin Mehri: Writing – original draft, Investigation, Formal analysis, Conceptualization; **Parichehr Behjati:** Writing – review & editing, Visualization, Data curation; **Dario Carpio:** Software, Data curation; **Angel D. Sappa:** Writing – review & editing, Supervision, Funding acquisition.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Angel Sappa reports financial support was provided by Air Force Office of Scientific Research. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the Air Force Office of Scientific Research Under Award FA9550-24-1-0206, in part by the Agencia Estatal de Investigación, Ministerio de Ciencia e Innovación, under Grant PID2021-128945NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by “ERDF Away of making Europe,” in part by the ESPOL project CIDIS-003-2024, in part by the Generalitat de Catalunya CERCA Program to CVCs general activities, and the Departament de Recerca i Universitats from Generalitat de Catalunya under Grant 2021SGR01499.

References

[1] R. Elgohary, A.K. Mahmoud, O. Khaled, S. Salah, M.B. Badawi, Enhanced thermal human detection in military applications using deep learning, in: 2024 International Telecommunications Conference (ITC-Egypt), IEEE, 2024, pp. 91–98.

[2] Y. Ling, Z. Luo, D. Lin, S. Li, M. Jiang, N. Sebe, Z. Zhong, Cross-modality average precision optimization for visible thermal person re-identification, Pattern Recognit. 164 (2025) 111489.

[3] L. Wang, O. Doukhi, D.J. Lee, FCDNet: a lightweight network for real-time wildfire core detection in drone thermal imaging, IEEE Access 13 (2025) 14516–14530.

[4] W. El Ahmar, A. Sappa, R. Hammoud, Thermal pedestrian multiple object tracking challenge (TP-MOT), in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 4602–4609.

[5] B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 136–144.

[6] A. Mehri, P.B. Ardakani, A.D. Sappa, MPRNet: Multi-path residual network for lightweight image super resolution, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2021, pp. 2704–2713.

[7] H. Gupta, K. Mitra, Pyramidal edge-maps and attention based guided thermal super-resolution, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, 2020, pp. 698–715.

[8] R.E. Rivadeneira, A.D. Sappa, B.X. Vintimilla, D. Bin, L. Ruodi, L. Shengye, Z. Zhong, X. Liu, J. Jiang, C. Wang, Thermal image super-resolution challenge results-PBVS 2023, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 470–478.

[9] H. Gupta, K. Mitra, Toward unaligned guided thermal super-resolution, IEEE Trans. Image Process. 31 (2021) 433–445.

[10] A. Mehri, P. Behjati, A.D. Sappa, TnTViT-G: Transformer in transformer network for guidance super resolution, IEEE Access 11 (2023) 11529–11540.

[11] P. Song, X. Deng, J.F.C. Mota, N. Deligiannis, P.L. Dragotti, M.R.D. Rodrigues, Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries, IEEE Trans. Comput. Imaging 6 (2019) 57–72.

[12] R.E. Rivadeneira, A.D. Sappa, B.X. Vintimilla, S. Nathan, P. Kansal, N. Gutierrez, B. Mojra, W.J. Beksi, Thermal image super-resolution challenge-PBVS 2021. In 2021 IEEE, in: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 1–11.

[14] J. Kopf, M.F. Cohen, D. Lischinski, M. Uyttendaele, Joint bilateral upsampling, ACM Trans. Graphics (ToG) 26 (3) (2007) 96–es.

[15] M. Camplani, T. Mantecón, L. Salgado, Depth-color fusion strategy for 3-D scene modeling with Kinect, IEEE Trans. Cybern. 43 (6) (2013) 1560–1571.

[16] D. Min, J. Lu, M.N. Do, Depth video enhancement based on weighted mode filtering, IEEE Trans. Image Process. 21 (3) (2011) 1176–1190.

[17] K. He, J. Sun, Fast guided filter, arXiv:1505.00996 (2015).

[18] X. Tan, C. Sun, T.D. Pham, Multipoint filtering with local polynomial approximation and range guidance, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2941–2948.

[19] J. Diebel, S. Thrun, An application of Markov random fields to range sensing, Adv. Neural Inf. Process. Syst. 18 (2005) 1–8.

[20] J. Park, H. Kim, Y.-W. Tai, M.S. Brown, I. Kweon, High quality depth map upsampling for 3D-TOF cameras, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 1623–1630.

[21] Y. Li, D. Min, M.N. Do, J. Lu, Fast guided global interpolation for depth and motion, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14, Springer, 2016, pp. 717–733.

[22] M. Kiechle, S. Hawe, M. Kleinsteuber, A joint intensity and depth co-sparse analysis model for depth map super-resolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1545–1552.

[23] J. Dong, J. Pan, J.S. Ren, L. Lin, J. Tang, M.-H. Yang, Learning spatially variant linear representation models for joint filtering, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2021) 8355–8370.

[24] S. Gu, W. Zuo, S. Guo, Y. Chen, C. Chen, L. Zhang, Learning dynamic guidance for depth image enhancement, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3769–3778.

[25] J. Xie, R.S. Feris, M.-T. Sun, Edge-guided single depth image super resolution, IEEE Trans. Image Process. 25 (1) (2015) 428–438.

[26] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[28] Y. Choi, N. Kim, S. Hwang, I.S. Kweon, Thermal image enhancement using convolutional neural network, in: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2016, pp. 223–230.

[29] X. Zhang, C. Li, Q. Meng, S. Liu, Y. Zhang, J. Wang, Infrared image super resolution by combining compressive sensing and deep learning, Sensors 18 (8) (2018) 2587.

[30] F. Liu, P. Han, Y. Wang, X. Li, L. Bai, X. Shao, Super resolution reconstruction of infrared images based on classified dictionary learning, Infrared Phys. Technol. 90 (2018) 146–155.

[31] C. Sun, J. Lv, J. Li, R. Qiu, A rapid and accurate infrared image super-resolution method based on zoom mechanism, Infrared Phys. Technol. 88 (2018) 228–238.

[32] W.-S. Lai, J.-B. Huang, N. Ahuja, M.-H. Yang, Deep Laplacian pyramid networks for fast and accurate super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 624–632.

[33] P. Behjati, P. Rodriguez, C. Fernández, I. Hupont, A. Mehri, J. Gonzàlez, Single image super-resolution based on directional variance attention network, Pattern Recognit. 133 (2023) 108997.

[34] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, SwinIR: Image restoration using swin transformer, in: Proceedings of the IEEE Int. Conference on Computer Vision, 2021, pp. 1833–1844.

[35] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, H. Li, Uformer: a general U-shaped transformer for image restoration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 17683–17693.

[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE Int. Conference on Computer Vision, 2021, pp. 10012–10022.

[37] Z. Wu, F. Zhu, K. Guo, R. Sheng, L. Chao, H. Fang, Modal adaptive super-resolution for medical images via continual learning, Signal Process. 217 (2024) 109342.

[38] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, L. Zhang, EDiffSR: an efficient diffusion probabilistic model for remote sensing image super-resolution, IEEE Trans. Geosci. Remote Sens. 62 (2023) 1–14.

[39] L. Li, Y. Zhang, L. Yuan, X. Gao, PLGNet: prior-guided local and global interactive hybrid network for face super-resolution, IEEE Trans. Circuits Syst. Video Technol. 34 (10) (2024) 10166–10181.

[40] Y. Xiao, Q. Yuan, K. Jiang, J. He, C.-W. Lin, L. Zhang, TTST: a top-k token selective transformer for remote sensing image super-resolution, IEEE Trans. Image Process. 33 (2024) 738–752.

[41] F. Almasri, O. Debeir, Multimodal sensor fusion in single thermal image super-resolution, in: Computer Vision–ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers 14, Springer, 2019, pp. 418–433.

[42] A. Kasliwal, P. Seth, S. Rallabandi, S. Singhal, CoReFusion: contrastive regularized fusion for guided thermal super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 507–514.

[43] X. Ye, B. Sun, Z. Wang, J. Yang, R. Xu, H. Li, B. Li, PMBANet: progressive multi-branch aggregation network for scene depth super-resolution, IEEE Trans. Image Process. 29 (2020) 7427–7442.

[44] N. Metzger, R.C. Daudt, K. Schindler, Guided depth super-resolution by deep anisotropic diffusion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 18237–18246.

[45] Z. Zhao, J. Zhang, S. Xu, Z. Lin, H. Pfister, Discrete cosine transform network for guided depth map super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 5697–5707.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: transformers for image recognition at scale, arXiv:2010.11929 (2020).

[47] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.-H. Yang, Restormer: efficient transformer for high-resolution image restoration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 5728–5739.

[48] D. Hendrycks, K. Gimpel, Gaussian error linear units (GELUs), arXiv:1606.08415 (2016).

[49] J. Tang, X. Chen, G. Zeng, Joint implicit image function for guided depth super-resolution, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4390–4399.

[50] B. Kim, J. Ponce, B. Ham, Deformable kernel networks for joint image filtering, Int. J. Comput. Vision 129 (2) (2021) 579–600.

[51] Z. Zhong, X. Liu, J. Jiang, D. Zhao, X. Ji, Deep attentional guided image filtering, IEEE Trans. Neural Netw. Learn. Syst. 35 (9) (2023) 12236–12250.

[52] X. Jiang, Z. Kuang, C. Guo, R. Zhang, L. Cai, X. Fan, C. Li, The devil is in the details: boosting guided depth super-resolution via rethinking cross-modal alignment and aggregation, arXiv:2401.08123 (2024).

[53] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, Z. Luo, Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 5802–5811.

[54] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: CVPR, 2018.

[55] R.E. Rivadeneira, A.D. Sappa, C. Wang, J. Jiang, Z. Zhong, P. Chen, S. Wang, Thermal image super-resolution challenge results-PBVS 2024, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2024, pp. 3113–3122.