Regular Article

E2DSR: Edge-Enhanced Representation for Deep Super-Resolution in Machine Vision Applications

Xiem HoangVan[1], Long Luong Hai[1], Thanh Nguyen Canh[2]

[1] University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam
[2] School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, Japan

Correspondence: Xiem HoangVan, xiemhoang@vnu.edu.vn
Communication: received 11 November 2025, revised 11 December 2025, accepted 22 December 2025
Online publication: 25 December 2025, Digital Object Identifier: 10.21553/rev-jec.427

Abstract– **While deep-based super-resolution (SR) has achieved remarkable progress, state-of-the-art models such as EDSR often rely solely on pixel-level information, resulting in overly smooth outputs that often fail to reconstruct the fine-grained edge details essential for downstream machine vision tasks. To address this challenge, we propose the Edge-Enhanced Deep Super-Resolution (E2DSR) model, a task-aware framework that leverages explicit edge guidance to enhance the reconstruction process with high-frequency edge information. E2DSR integrates a novel Edge Feature Enhancement (EFE) Block into a deep residual architecture, which learns to extract and fuse salient edge features from the low-resolution input. We demonstrate the effectiveness of our approach within a gesture recognition, where E2DSR significantly enhances input quality for a state-of-the-art YOLOv10 detector. Experimental results show that our method substantially outperforms the original EDSR and other approaches, increasing the mean average precision (mAP) from 0.776 to 0.822 on average across four representative gesture action types. Our work demonstrates that explicit edge guidance is a crucial component for developing super-resolution models that excel in practical machine vision applications.**

Keywords– **Edge-guided super-resolution, edge extraction, residual network, super-resolution.**

1 Introduction

Image super-resolution (SR) has long been a critical task in low-level computer vision, aiming to reconstruct high-resolution (HR) images from their low-resolution (LR) counterparts. This problem has significant applications across various domains, including medical imaging [1], digital forensics [2], surveillance [3], and machine vision [4, 5]. However, SR is inherently ill-posed, as a single LR image can correspond to multiple HR versions, making the recovery of lost high-frequency details a significant challenge. To achieve SR images, extensive research has been conducted, primarily along two main approaches: model-based and learning-based methods. The model-based approach relies on mathematical formulations and optimization techniques, incorporating prior knowledge, such as image smoothness, sparsity, and edge continuity, into the reconstruction process.

Traditional interpolation-based methods, such as Bicubic [6] and Lanczos [7, 8], regularization-based optimization [9, 10], and sparse coding [11, 12], fail to recover fine details, especially edges, resulting in blurred reconstructions. With the advent of deep learning, convolutional neural networks (CNNs) [13, 14] have revolutionized the field, leading to a variety of architectures that have advanced reconstruction quality. Early models like SRCNN [15] established an end-to-end mapping from LR to HR images, and subsequent innovations such as residual learning in VDSR [13] and EDSR [16] have significantly improved the performance of SR systems by leveraging hierarchical feature representations. More recently, transformer-based models [17, 18] have leveraged self-attention mechanisms to effectively model long-range dependencies between pixels, further advancing the state-of-the-art. However, these models are typically trained with pixel-wise loss functions, which tend to produce overly smooth outputs that fail to reconstruct the critical high-frequency details constituting edges and textures. This oversmoothing limitation is particularly problematic for downstream machine vision tasks where structural fidelity is essential. In addition, many of these high-performing models are computationally intensive, creating a need for more efficient, lightweight architectures suitable for real-time applications on edge devices.

To address the oversmoothing limitation, recent works have introduced edge-aware strategies to better guide the SR process. Research in visual perception has established that edge detection is one of the earliest stages of human visual processing [19], and object recognition depends primarily on edge-based structural information rather than surface properties [20]. These findings motivate the integration of explicit edge priors into SR networks. For instance, Nazeri *et al.* [21] proposed an edge-informed loss that uses precomputed edge maps as priors during training. Similarly, Wang *et al.* [22] introduced an Edge-Enhanced Feature Distillation Network (EFDN) that integrates edge-preserving modules and loss functions for efficient yet effective SR. Some methods [23, 24] reformulate the SR problem as an image inpainting task, where the edge generator first predicts the high-resolution edge map, which then guides the texture and color reconstruction.

Methods such as SRREdgeNet [25] and EdgeSR [26] demonstrate that by leveraging edge-specific modules, SR networks can better restore intricate boundaries and features. Moreover, techniques like the Soft-Edge Assisted Network [27] show that adaptively learning edge representations, rather than relying on fixed filters, can yield more robust and generalizable models. However, these approaches often introduce significant computational overhead through auxiliary edge prediction networks or require HR edge supervision during training, limiting their practicality for real-time deployment.

Among learning-based techniques, the Enhanced Deep Super-Resolution (EDSR) model is widely recognized as a state-of-the-art method for ISR [16]. By eliminating batch normalization layers, EDSR reduces undesirable artifacts and enhances the reconstruction of fine image details. Furthermore, its residual learning architecture mitigates the vanishing gradient problem, enabling the effective training of deeper networks. However, the standard EDSR model comprises approximately 43 million parameters, resulting in substantial computational demands that hinder its deployment in resource-constrained environments. While a reduced baseline version with 16 residual blocks offers a more feasible alternative, this simplification often leads to a drop in reconstruction quality due to limited model capacity. These challenges highlight the need for an efficient enhancement to the EDSR architecture that can improve reconstruction quality, particularly for high-frequency structural details, without substantial computational overhead.

The importance of structural detail preservation becomes particularly evident in machine vision applications where fine textures and boundaries directly impact recognition performance. Gesture recognition systems, for example, rely heavily on the accurate reconstruction of hand contours and finger positions to distinguish between different gestures. When input images are captured from long distances or with low-cost sensors, the resulting low resolution can obscure these critical details, significantly degrading recognition accuracy. This motivates the development of SR methods that prioritize the reconstruction of structurally significant features for downstream recognition tasks.

To address these challenges, this paper proposes the Edge-Enhanced Deep Super-Resolution (E2DSR) model, a task-aware framework that integrates explicit edge guidance with a lightweight architecture. Our approach incorporates a novel Edge Feature Enhancement (EFE) block into a streamlined EDSR backbone with 16 residual blocks. Unlike existing edge-guided methods that employ separate auxiliary networks or require HR edge supervision, the EFE block is a shallow, learnable module that extracts edge information directly from the LR input using classical operators (Sobel or Canny) and processes it through a multi-path architecture with learnable fusion. This design enables the network to explicitly preserve and reconstruct structural features vital for machine perception while adding only 86K parameters (4.8 increase) to the baseline. As demonstrated in Figure 1, this focus on structural detail translates directly to improved performance on downstream tasks, achieving higher confidence in gesture recognition compared to existing SR methods. In summary, the key contributions of this paper are summarized as follows:

(a) Low resolution image (b) Bicubic

(c) EDSR (d) Our proposed

Figure 1. Comparison of gesture recognition results on a $\times 4$ super-resolved image using different methods. The YOLOv10 detector assigns varying confidence scores depending on the input quality. Our proposed method achieves the highest confidence (0.86), outperforming Bicubic (0.74), EDSR (0.69), and the original low-resolution input.

- A task-aware super-resolution framework (E2DSR) that integrates an explicit edge-enhancement module with a deep residual network to prioritize the reconstruction of structurally significant features.
- The design of a lightweight EFE Block to achieve high-fidelity edge reconstruction without significant computational overhead.
- A modern gesture recognition pipeline that demonstrates the practical benefits of our SR model in improving the accuracy of a state-of-the-art YOLOv10 model.
- A comprehensive evaluation on multiple standard benchmarks that validates the superior performance of our proposed model against state-of-the-art methods, particularly in machine vision applications.

The remainder of this paper is organized as follows. Section **Related Works** describes a review of related works in super-resolution, edge-guided image reconstruction, and gesture recognition. Section **Proposed Method** details the architecture of our proposed super-resolution network - E2DSR, including the edge feature

enhancement module and SR subnetworks. Section **Experimental Results** presents a comprehensive evaluation of our method on various datasets, comparing it with state-of-the-art approaches. Finally, Section **Conclusion** concludes the paper and discusses potential directions for future research.

2 Related Works

2.1 Edge-Guided Super-Resolution

While deep learning has significantly advanced the field of single-image super-resolution (SISR), a primary challenge remains: models trained with standard pixel-wise losses tend to produce overly smooth results, failing to reconstruct the critical high-frequency details that constitute edges and textures. State-of-the-art models like the EDSR model [16] remain a cornerstone in single-image SR, utilizing deep residual networks and pixel-shuffle upsampling for improved fidelity. While effective in distortion-based metrics, it often produces overly smoothed textures due to its sole reliance on pixel-wise MSE loss. To address this, SRGAN [28] introduced a perceptual loss and adversarial training to prioritize perceptual quality over PSNR. Similarly, RCAN [29] leveraged channel attention to preserve fine details in challenging regions. A significant challenge in single-image SR is losing the high-frequency details that define object structures. To address this, many researchers have proposed edge-guided or edge-enhanced methods that use edge information as an explicit prior to guide the reconstruction process by incorporating high-frequency structure into SR pipelines. Ye *et al.* [30] presents an edge-guided depth filling method to interpolate depth values on the HR image grids constrained by the acquired edges to prevent predicting across the depth boundaries. The first and most common approach is to use an auxiliary edge prediction network. In this paradigm, a dedicated subnetwork first extracts or reconstructs an edge map from the low-resolution (LR) input, which is then fused with features in the main SR network. For instance, SeaNet [27] introduced a "soft-edge reconstruction network" (Edge-Net) to generate edge priors that assist in the final image refinement. SRREdgeNet [25] proposed a sequential pipeline where a dense edge detection network processes the output of an initial SR model before a final merge network combines them. Others, like SESR [31], integrate Laplacian filters directly into the loss function, guiding the model to recover spatial gradients more accurately. These approaches recognize that accurate edge recovery is critical for high-quality SR. This strategy has been applied across different domains, including models that predict depth edges from LR depth and color images and those that use an edge detection auxiliary network for infrared image SR.

A second strategy involves integrating edge processing directly into the network architecture. Rather than using a separate network, these methods incorporate edge-aware components into their core building blocks. For example, EIPNet [24] embeds a lightweight "edge block" at multiple scales within the SR network to progressively provide structural information during the upscaling process. A particularly efficient approach is the re-parameterizable Edge-oriented Convolution Block (ECB) [32] proposed by ECBSR, which uses a multi-branch design during training to learn 1st and 2nd-order spatial derivatives that are then merged into a single, fast 3×3 convolution for inference. The third strategy enforces edge fidelity through the training objective. Rather than relying solely on architectural changes, these methods introduce specialized loss functions that explicitly penalize edge inaccuracies. EFDN [33] introduced an "edge-enhanced gradient loss" to explicitly penalize inaccuracies in the gradient domain, forcing the network to preserve high-frequency information during training better. Edge-SR [26] presents a set of one-layer architectures designed for image SR on edge devices. Hu *et al.* [34] comprises the SR backbone network, which includes a shallow features extraction module, a deep feature extraction module, and a reconstruction module with the edge detection auxiliary network (EDAN).

While these methods validate the importance of edge priors, they often introduce significant trade-offs. Approaches with auxiliary networks such as SeaNet [27] and SRREdgeNet [25] substantially increase model complexity and inference time. Additionally, several methods require HR edge maps as supervision during training, limiting data preparation flexibility, while fixed filter approaches lack adaptability for task-specific edge representations. Our proposed E2DSR addresses these limitations through a different design philosophy. Unlike auxiliary network approaches, our Edge Feature Enhancement (EFE) block is a shallow, learnable module. The EFE block extracts edge information directly from the LR input using Sobel or Canny operators, eliminating the need for HR edge supervision. The multi-path architecture with learnable fusion enables adaptive combination of edge cues with contextual information. This design achieves competitive reconstruction quality with minimal inference overhead, making our method suitable for real-time machine vision applications.

2.2 Super-Resolution for Gesture Recognition

The performance of vision-based gesture recognition systems is fundamentally dependent on the quality of the input image. In many practical scenarios, such as long-range human-robot interaction (HRI) or when using low-cost sensors, images are often of low resolution, which can obscure the fine details necessary for accurate classification. To address this, super-resolution (SR) and image enhancement techniques have been employed as crucial pre-processing steps to improve the clarity and detail of gesture images. Recent work has highlighted the necessity of SR for enabling gesture recognition at extended distances. Bamani *et al.* [35] addresses the "Ultra-Range Gesture Recognition (URGR)" problem, aiming for effective recognition at distances up to 25 meters using only

an RGB camera. They identify low resolution as the primary challenge and propose a novel SR model, HQ-Net, to specifically enhance the cropped image of the user before classification. This approach is motivated by the observation that general-purpose SR models like ESRGAN, while effective for some tasks, may struggle with gestures by over-smoothing and distorting indistinguishable features such as fingers. Similarly, SR has been explored in the context of specialized sensors. Chen *et al.* [36] tackles gesture recognition using ultra-low-resolution infrared thermopile sensors, where environmental temperatures can cause blurry or missing finger contours. They employ a diffusion model for image reconstruction to enhance gesture area features and improve subsequent recognition accuracy. In recognition-based applications [37], SR enhances spatial details and texture cues critical for robust feature extraction—particularly when hardware limitations, compression artifacts, or long-range imaging compromise input quality. Low-resolution inputs can severely impair the effectiveness of object detection algorithms. By reconstructing fine-grained structures such as edges, contours, and textures, SR facilitates the accurate detection of small or distant objects, which is particularly valuable in domains like surveillance, UAV imagery, and autonomous systems [38].

Beyond single-frame enhancement, some frameworks have leveraged temporal information in image sequences. Li *et al.* [39] proposes a gesture recognition system based on multi-frame super-resolution, which fuses feature information from multiple consecutive frames to reconstruct a high-resolution gesture image. This approach is designed to mitigate issues like motion blur that occur during rapid gesture transformations in dynamic environments. Other research has explored super-resolution in alternative domains; for example, Kushwaha *et al.* [40] developed an adaptive super-resolution transform (ASLIT) to generate a high-resolution time-frequency representation of a gesture image, which is then used for classification. However, these methods typically require large datasets of corresponding low- and high-resolution images for training, which are often unavailable for specific sensor types, thus limiting the applicability of standard SR techniques. In addition, sophisticated reconstruction techniques can come with a high computational cost, which is a barrier for real-time applications.

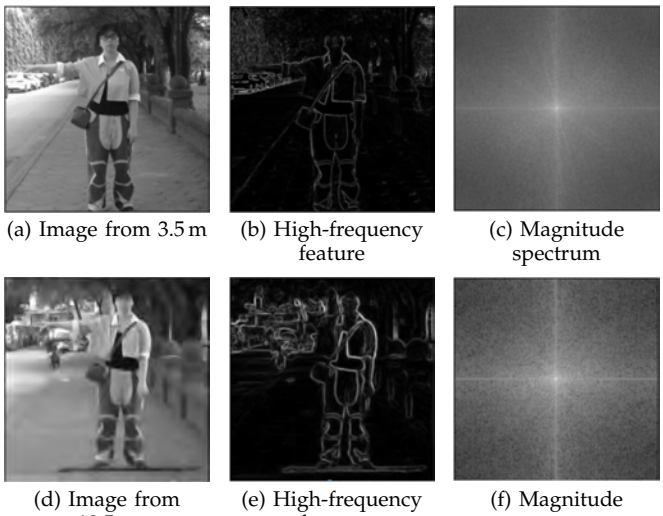
These limitations underscore the need for a task-aware SR model that effectively preserves structurally-critical features and is efficient enough for practical deployment. Our work addresses this gap by integrating edge-enhanced super-resolution with YOLOv10 [41], a state-of-the-art real-time object detector, for gesture recognition on the HaGRID dataset [42]. Unlike previous approaches using general-purpose SR models [35] or sensor-specific architectures [36], our E2DSR explicitly preserves structural features critical for recognition. This enables us to demonstrate the direct relationship between edge-enhanced reconstruction and improved recognition accuracy.

3 Proposed Method

3.1 Overview of the E2DSR Network

The goal of our proposed Edge-Enhanced Deep Super-Resolution (E2DSR) network is to reconstruct an HR image I_{hr}^{*} from an LR input I_{lr}, unlike standard models that primarily optimize for pixel-level accuracy. E2DSR is designed to be task-aware, prioritizing the restoration of high-frequency structural details that are critical for downstream machine vision tasks. Figure 2 illustrates two images of a similar scene captured from different distances, along with their corresponding edge maps and 2D Fourier transform representations. As observed, the image captured from a greater distance exhibits a notable reduction in high-frequency components, resulting in degraded visual quality compared to the image taken from a closer proximity. Although deep learning-based SR models are capable of learning high-frequency features implicitly through end-to-end training and loss function optimization, our experiments indicate that capturing such features often incurs significant computational overhead during training. In some cases, this added complexity can even degrade output quality due to ineffective feature learning or overfitting.

The overall architecture builds upon the proven effectiveness of the EDSR model, utilizing a streamlined deep residual backbone to balance performance and computational efficiency. The core of our contribution is the integration of a novel Edge Feature Enhancement (EFE) Block that operates in parallel with the main feature-extraction path. This block explicitly extracts, processes, and injects edge information into the network, thereby providing direct guidance for reconstructing sharp, coherent object boundaries. A fundamental consideration in our approach is that edge maps are extracted directly from the low-resolution input, which inherently contains degraded high-frequency information compared to the original high-resolution image. Although downsampling reduces high-frequency



(a) Image from 3.5 m (b) High-frequency feature (c) Magnitude spectrum

(d) Image from 10.5 m (e) High-frequency feature (f) Magnitude spectrum

Figure 2. Illustration of images with various amounts of high-frequency information due to the captured distance.

details, the fundamental structural information encoded in edges, such as object boundaries, dominant contours, and spatial relationships, is largely preserved at a coarse level. In addition, the human visual system relies heavily on edge information for object recognition, and these salient structural cues remain identifiable even at reduced resolutions [20]. The Sobel and Canny operators can still capture the approximate locations and orientations of prominent edges from LR inputs. Importantly, our approach does not aim to recover an exact HR edge map; rather, the extracted serve as spatial priors that guide the network's attention toward structurally significant regions requiring enhanced reconstruction.

Furthermore, our multi-path feature processing module is specifically designed to compensate for LR edge limitations. The feature extraction branch applies learnable convolution to enhance coarse edge information, while the feature augmentation branch provides complementary contextual information. The learned fusion mechanism adaptively combines both branches, enabling the network to refine the initial edge cues beyond what is directly observable in the LR input. This explicit guidance mechanism distinguishes our approach from conventional models that learn such features only implicitly.

3.2 Edge Feature Enhancement Block

The Edge Feature Enhancement (EFE) Block is the core architectural contribution of our E2DSR model. It is a lightweight module designed to explicitly extract, process, and inject salient high-frequency edge features from the initial LR input into the main network, providing direct structural guidance for the reconstruction. The process is composed of two primary stages: edge extraction and multi-path feature processing. Figure 3 illustrates the Edge Feature Enhancement Block, which integrates alongside the residual convolutional blocks to inject additional high-frequency information into the low-resolution input, thereby enhancing the overall reconstruction quality.

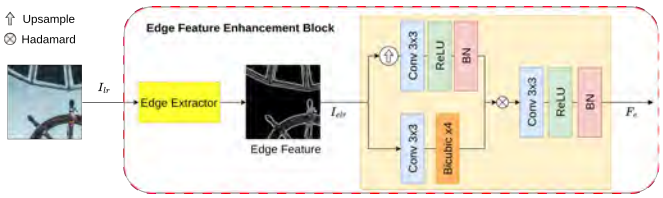


Figure 3. Architecture of our proposed EFE module for edge feature learning.

3.2.1 Edge Extraction: Given a low-resolution input image I_{lr}, the first step is to extract an initial edge map, denoted as I_{elr}. To further investigate the impact of different edge detection techniques, two versions of our model are implemented: 1) E2DSR_C, which integrates edge information using the Canny algorithm, and 2) E2DSR_S, which utilizes the Sobel algorithm for edge enhancement. Firstly, the Sobel algorithm is applied to our model, which is a gradient-based edge detection technique that calculates the gradient of image intensity at each pixel, emphasizing regions of rapid intensity change, which are typically edges. This method approximates the image gradient by convolving the input with two 3×3 kernels, G_x for horizontal changes and G_y for vertical changes

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \quad (1)$$

The corresponding gradient images, I_x and I_y are computed as

$$I_x = I_{LR} * G_x, \quad I_y = I_{LR} * G_y, \quad (2)$$

where $*$ denotes the 2D convolution operation. The final edge map I_{elr} is the magnitude of the gradient

$$I_{ELR} = \sqrt{I_x^2 + I_y^2}. \quad (3)$$

Secondly, the Candy is a more refined, multi-stage algorithm that produces cleaner and more continuous edges. The first step is noise reduction, which begins by smoothing the image to reduce noise and interference with gradient calculation. This is typically done by convolving the input image I_{lr} with a 2D Gaussian filter kernel, G_σ

$$\begin{aligned} G_\sigma(x,y) &= \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \\ I_{smooth} &= I_{lr} * G_\sigma. \end{aligned} \quad (4)$$

After that, the gradient magnitude M and direction θ are calculated from the smoothed image, typically using Sobel operators, similar to the method described as

$$\begin{aligned} M &= \sqrt{(I_{smooth} * G_x)^2 + (I_{smooth} * G_y)^2}, \\ \theta &= atan2((I_{smooth} * G_y), (I_{smooth} * G_x)). \end{aligned} \quad (5)$$

To achieve thin, single-pixel-wide edges, the non-maximum suppression step examines each pixel and suppresses its values to zero if its magnitude is not the maximum compared to its two neighbors along the gradient direction θ. This ensures that only the sharpest local peaks of the gradient remain. Finally, two thresholds, a high (T_h) and a low (T_l), are used to distinguish between strong, weak, and non-edges. The final edge map I_{elr} is formed by a conditional process

$$I_{elr}(x,y) = \begin{cases} 1 & \text{if } M(x,y) > T_h, \\ 1 & \text{if } T_l < M(x,y) \le T_h, \\ & \text{and is connected to a strong edge,} \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

This preserves strong edges while also including weak edges that are part of a continuous line, effectively eliminating noise-related weak edges. The edge features extracted by either the Sobel or Canny operator are then passed to the subsequent feature-processing module. This integration enriches the model's ability to perceive and retain structural details, ultimately improving the super-resolution process and resulting in higher recognition accuracy for the downstream gesture recognition task.

3.2.2 Multi-Path Feature Processing and Fusion: The extracted edge map I_{elr} is then fed to the multi-path module to process the edge features at different scales and contexts. This module consists of two parallel branches:

1) ***Feature extraction branch***: This branch processes the features at an upsampled scale to learn a rich representation. Let the function for this branch be denoted by H_{ext}. The edge map is first upsampled using nearest-neighbor interpolation $\mathcal{U}_{nn}$, then passed through a convolutional layer with a kernel size of 3 and 64 channels (W_{ext}, b_{ext}), a Scaled Exponential Linear Unit (SELU) activation function [43] instead of the usual ReLU activation function, and a Batch Normalization ($\mathcal{BN}$) layer. We specifically choose SELU to avoid the "dying ReLU" problem and enable the learning of more complex features within a shallow block [44]. Since the feature-extraction component consists of a single convolutional layer, avoiding dying nodes is essential to realize the model's potential fully. In this case, the SELU function maps all input nodes, enabling the network to learn more complex features that are typically achieved with a deeper network. The SELU function is mathematically described as follows

$$\text{SELU}(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha(e^x - 1) & \text{if } x \leq 0 \end{cases}. \quad (7)$$

where λ and α are both constant with the approximate value of 1.6733 and 1.0507, respectively. As described above, the SELU function has a soft exponential tail for negative x values, which helps neurons continue learning even with negative activations rather than dying, unlike the ReLU. Finally, the proposed model incorporates a batch normalization layer to stabilize and accelerate training. The output of this branch, F_e, is given by

$$F_{ext} = \mathcal{BN}(SELU(W_{ext} * \mathcal{U}_{nn}(I_{elr}) + b_{ext})). \quad (8)$$

2) ***Feature augmentation***: This branch provides multi-scale contextual information. Let its function be H_{aug}. The edge map is passed through a convolutional layer (W_{aug}, b_{aug}) and then a bicubic interpolation layer $(\mathcal{U}_{bc})$ to upsample the features while preserving smoothness. This provides structural context to guide the final reconstruction. The output F_{aug}, can be written as

$$F_{aug} = \mathcal{U}_{bc}(W_{aug} * I_{elr} + b_{aug}). \quad (9)$$

Finally, the outputs of these two branches are fused. First, they are combined element-wise via addition. This combined feature map is then passed through a final fusion block, H_{fuse}, which consists of another sequence of convolution, SELU activation, and batch normalization, to learn the optimal combination of the features. The entire fusion process yields the final edge feature map, F_e

$$\begin{aligned} F_{fused} &= F_{ext} + F_{aug}, \\ F_e &= H_{fuse}(F_{fused}). \end{aligned} \quad (10)$$

The complete operation of the EFE block, which transforms the input LR image into a high-level edge feature map, can be summarized by the function $A(\cdot)$ as

$$F_e = A(I_{lr}). \quad (11)$$

3.3 Overall EFE Super Resolution Network Architecture

The architecture of the proposed Edge-Enhanced EDSR (E2DSR) model is designed to reconstruct an HR image I^*_{hr} with enhanced visual quality from a given low-resolution input I_{lr}, as shown in Figure 4. Our approach builds upon the deep residual learning framework originally introduced by EDSR, but with key modifications to explicitly integrate high-frequency edge information and improve computational efficiency.

The backbone of our E2DSR model is a streamlined deep residual network. To balance reconstruction quality and computational cost, we employ a 16-layer residual architecture. This provides sufficient depth to learn complex mappings between low- and high-resolution images while maintaining a manageable number of parameters suitable for practical applications. Similar to the original EDSR, our model leverages a residual learning framework where skip connections are used to ensure stable information flow across layers. This design mitigates the vanishing gradient problem commonly encountered in deep neural networks, thereby enabling effective training.

The primary architectural innovation of E2DSR is the integration of the Edge Feature Enhancement (EFE) Block, as detailed in the previous section. Unlike the standard EDSR [16] architecture, which implicitly learns features in the pixel domain, our model incorporates a parallel path that explicitly captures and refines edge information. The feature map generated by the EFE Block, which is rich in high-frequency structural details, is fused with the output of the main residual backbone. This fusion provides direct, explicit guidance to the network, thereby enabling it to better preserve and reconstruct sharp edges and fine textures that are critical for both visual sharpness and downstream recognition tasks.

Finally, the fused feature map, enriched with both deep contextual features and explicit edge information, is passed to an upsampling module. Following modern efficient designs, upsampling is performed at the end of the network using a pixel-shuffle layer to generate the final high-resolution output image, I^*_{hr}.

3.4 E2DSR-based Gesture Recognition Application

To evaluate the effectiveness of the proposed E2DSR model in a practical machine-vision context, we integrate it into a gesture-recognition pipeline. This downstream task is an ideal test case, as accurate classification of hand gestures depends heavily on the clarity of fine-grained details, such as finger position and contour, which are often lost in low-resolution imagery.

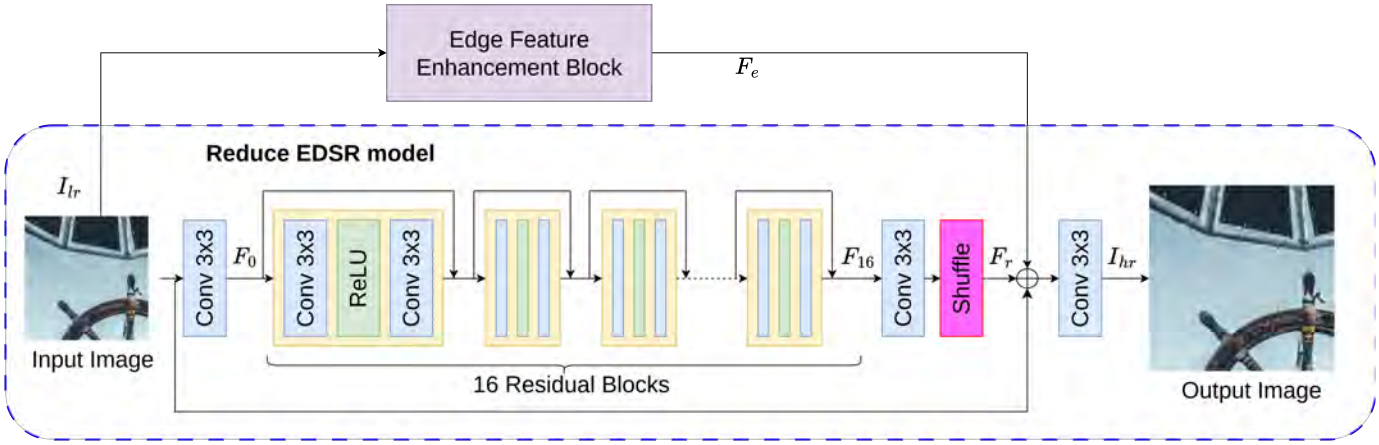


Figure 4. Architecture of the proposed Edge-Enhanced Deep Super-Resolution (E2DSR) network. The model consists of a streamlined residual backbone and a parallel Edge Feature Enhancement (EFE) block. The EFE block extracts and refines edge features from the low-resolution input, which are then fused with deep features from the backbone to explicitly guide the reconstruction of the final high-resolution image via a pixel-shuffle upsampler.

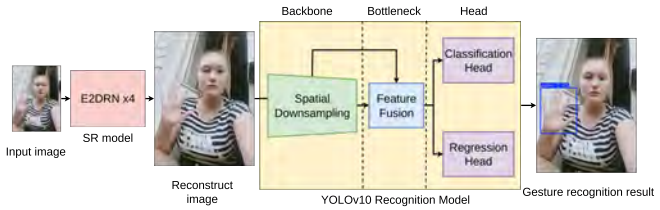


Figure 5. Pipeline for machine gesture recognition task with the use of E2DSR.

The proposed pipeline, illustrated in Figure 5, first enhances the quality of the input image using E2DSR and then passes the super-resolved output to a state-of-the-art object detector for classification. For the recognition component, we employ YOLOv10 [41], a real-time, end-to-end object detection model that represents the latest advancement in the YOLO family. We selected YOLOv10 for its excellent balance of speed and precision, making it highly suitable for real-time applications on both cloud systems and edge devices.

As illustrated in the pipeline, the process begins by improving the input image quality using the E2DSR model with a super-resolution scale factor of 4×. The enhanced images are then passed to the YOLOv10 model for gesture recognition. The detector is trained to classify four specific gesture classes from the HaGRID dataset [42]): palm, two up, two up inverted, and stop. This experimental setup enables us to directly and quantitatively assess how the architectural improvements of E2DSR affect the performance of a high-level machine-vision task.

4 Experimental Results

4.1 Dataset and Metrics

To evaluate general image enhancement capabilities of our model, we used five standard super-resolution benchmark datasets: Set 5 [45], Set 14 [46], Urban 100 [47], BSD 100 [48], and DIV2K [49]. For the downstream task, we created a custom subset from the HaGRID dataset [42] for gesture recognition, containing 7,138 images across four gesture classes: palm, two up, two up inverted, and stop, which is shown in Table I.

Table I
Distribution of images across the four gesture classes in our custom HaGRID subset used for the recognition task

Gesture	Number
Palm	1770
Two up	1855
Two up inverted	1765
Stop	1748
Total	**7138**

We assessed image reconstruction quality using the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM). For the downstream gesture recognition task, we evaluate the detector's performance using mean Average Precision (mAP). PSNR is used to measure the pixel-wise reconstruction quality. It is defined based on the Mean Squared Error (MSE) between the ground-truth HR image (I_{hr}) and the super-resolved image (I_{sr})

$$\text{MSE} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I_{HR}(i,j) - I_{SR}(i,j)]^2. \quad (12)$$

PSNR is then calculated as

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}_I^2}{\text{MSE}}\right), \quad (13)$$

where MAX_I is the maximum possible pixel value of the image. A higher PSNR value indicates a better quality of reconstruction. Structural Similarity Index Measure (SSIM) evaluates the perceptual similarity between two images by considering luminance, contrast, and structure. The SSIM index is calculated as

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \quad (14)$$

where μ_x and μ_y are the local mean, σ_x and σ_y are the standard deviations, and σ_{xy} is the cross-covariance for image windows x and y. The constants c_1 and c_2 are included to stabilize the division. The SSIM value ranges from -1 to 1, where 1 indicates perfect structural similarity. Mean Average Precision (mAP) is the primary metric for evaluating object detection model performance in the gesture recognition task. It is based on the concepts of Precision and Recall

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \tag{15}$$

where TP, FP, and FN are the counts of true positives, false positives, and false negatives, respectively. The Average Precision (AP) for a single class is calculated as the area under the precision-recall curve. The mAP is then the mean of the AP values across all gesture classes

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i. \tag{16}$$

For training, we use a dataset S consist of N pair of low-resolution image I_{lr} and high-resolution image I_{hr} with $S = (I_{lr}|I_{hr})_{i=1}^{m}$ During training, we used a batch size of 16 with the Adam optimizer. The learning rate was initialized at 10^{-4} and have every 10^{-5} batch updates. We used the MSE as the loss function for all models to ensure that performance differences are attributable to architectural changes rather than the training objective

$$\begin{aligned} \mathcal{L} &= \arg\min_{\theta} \text{MSE}(I^*_{hr}, I_{hr}) \\ &= \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \|f_\theta(I^i_{lr}) - I^i_{hr}\|^2, \end{aligned} \tag{17}$$

where $I^*_{hr} = f_\theta(I_{lr})$ denotes the super-resolved output.

4.2 Quantitative and Qualitative Comparison

To evaluate the quality improvements achieved by the proposed method, we compared it against five widely-used SR techniques: Bicubic interpolation [6], VDSR [13], SRCNN [15], SRResNet [28], EDSR [16], and SWinIR [50]. All models were trained on the same dataset to ensure a fair comparison.

As shown in Table II, the E2DSR models (E2DSR_S using Sobel and E2DSR_C using Canny) achieve competitive PSNR and SSIM scores across all benchmark datasets. In particular, compared with bicubic interpolation, E2DSR yields an approximately 2.7 dB improvement in PSNR. Average PSNR for both E2DSR_S and E2DSR_C (28.72) is the highest among all models, indicating strong overall performance. Compared to the baseline EDSR, our method provides a consistent, albeit modest, improvement in these distortion-based metrics, with an average PSNR gain of 0.09 dB. This improvement is small but consistent, showing that E2DSR methods provide further refinement on a strong baseline. These findings highlight the effectiveness of incorporating high-frequency information, which significantly enhances the quality of szuper-resolved images. We also compare our method with SwinIR, a recent transformer-based state-of-the-art super-resolution method that leverages self-attention mechanisms to capture long-range dependencies. As expected, SwinIR achieves higher PSNR and SSIM scores across most benchmark datasets, consistent with its design objective of maximizing reconstruction quality through global attention mechanisms. However, this performance advantage comes at a significant computational cost.

To assess the computational complexity of the proposed method, we measure inference time on an NVIDIA Tesla T4 GPU, along with the number of parameters and multiply–accumulate operations (MACs), as reported in Table III. The addition of the EFE block introduces a moderate computational overhead compared to the streamlined EDSR baseline, including an increase in parameters and MACs. SwinIR requires 37.99 ms for inference, which is approximately 4.5× slower than our E2DSR_S (8.41 ms) and 6.4× slower than the baseline EDSR (5.98 ms). While our E2DSR models introduce a moderate computational overhead compared to EDSR due to the addition of the EFE block, they remain significantly more efficient than transformer-based approaches. This trade-off between reconstruction quality and computational efficiency is central to our design philosophy. The primary objective of E2DSR is not to maximize PSNR/SSIM metrics, but rather to enhance performance on downstream machine vision tasks while maintaining practical inference speeds.

The visual comparisons in Figure 6 further underscore the benefits of our approach. In challenging images with fine textures and repeating geometric patterns (*e.g.*, img_008 from Urban100), our E2DSR model reconstructs visibly sharper edges and more coherent structures compared to the baseline EDSR and other methods. These results confirm that the EFE block enables the model to better preserve the high-frequency details that are critical for perceptual quality.

4.3 Gesture Recognition Assessment

The primary goal of our task-aware model is to improve performance on downstream machine vision tasks. Table IV presents the assessment of gesture recognition in the HaGRID data set. As shown, the E2DSR model yields a substantial improvement in recognition accuracy, increasing from 0.336 (LR input) to 0.822. For the "Palm" gesture, E2DSR_S achieves a notable mAP of 0.838, which is significantly higher than Bicubic (0.743) and LR (0.400), showing its strong capability to enhance the features of fine gestures. The most challenging gesture, "Two up inverted", also benefits from E2DSR_S (0.798 vs. Bicubic's 0.770), indicating the method's robustness even in more complex or less common cases. This significant gain is a key finding of our work. It demonstrates that the architectural focus on enhancing edge and structural fidelity directly yields a more effective feature representation for the YOLOv10 recognition model, thereby validating our task-aware design philosophy.

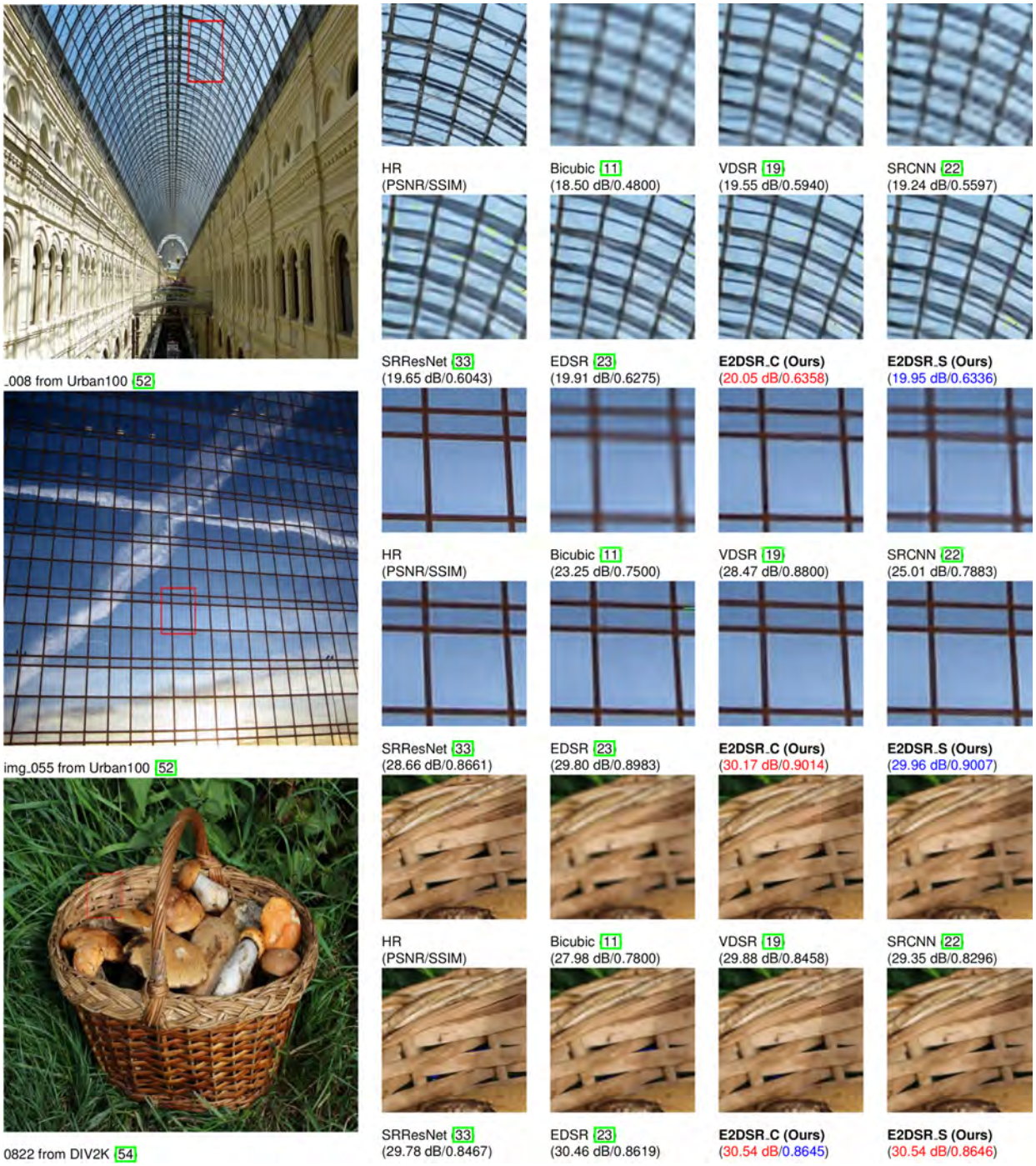


Figure 6. Qualitative comparison of our models with other works on ×4 super-resolution. Red indicates the best performance, and blue indicates the second-best.

5 Conclusion

In this paper, we introduce E2DSR, an edge-enhanced deep residual network for image super-resolution, designed to bridge the gap between models optimized for perceptual quality and the requirements of downstream machine-vision tasks. By integrating an explicit Edge Feature Enhancement (EFE) block into a streamlined EDSR backbone, our model effectively combines both pixel-level and high-frequency edge cues. This task-aware approach enables an improved reconstruction of the critical structural features that are often lost in standard super-resolution methods. Experimental results confirmed that E2DSR outperforms the baseline EDSR and other SR models in terms of both visual quality and task-specific accuracy. Most notably, integrating E2DSR into a gesture recognition pipeline significantly improved accuracy, increasing mean Average Precision (mAP) from 0.776 to 0.822. These findings validate that our architectural focus on explicit edge guidance provides a direct and substantial benefit to machine perception tasks.

The primary limitation of the current approach is the moderate increase in computational complexity from the EFE block. Future work will focus on reducing this overhead through model optimization techniques,

Table II
Objective quality assessment with various SR models (PSNR/SSIM)

Dataset	Bicubic [6]	VDSR [13]	SRCNN [15]	SRResNet [28]	EDSR [16]	SwinIR [50]	E2DSR_S	E2DSR_C
Set 5 [45]	25.86/ 0.7483	29.19/ 0.8435	28.55/ 0.8255	29.47/ 0.8397	29.70/ 0.8522	**29.88**/ **0.8556**	29.79/ 0.8550	29.80/ 0.8550
Set 14 [46]	23.79/ 0.6427	26.00/ 0.7289	25.59/ 0.7144	26.26/ 0.7352	26.30/ 0.7375	**26.43**/ **0.7143**	26.37/ 0.7406	26.36/ 0.7407
Urban 100 [47]	21.25/ 0.6098	23.66/ 0.7279	22.93/ 0.6942	23.81/ 0.7344	23.97/ 0.7436	**24.18**/ **0.7523**	24.15/ 0.7507	24.13/ 0.7498
BSD 100 [48]	24.87/ 0.6465	26.74/ 0.7275	26.35/ 0.7151	26.92/ 0.7334	26.95/ 0.7352	**27.05**/ **0.7390**	27.01/ 0.7374	27.00/ 0.7372
DIV2K [49]	26.88/ 0.7365	27.15/ 0.8259	26.49/ 0.8041	27.62/ 0.8366	27.62/ 0.8367	**27.81**/ **0.8418**	27.76/ 0.8408	27.79/ **0.8418**
HaGRID [42]	33.37/ 0.8988	36.93/ 0.9308	35.57/ 0.9177	36.03/ 0.9197	37.22/ 0.9327	**37.26**/ 0.9332	37.23/ **0.9334**	37.25/ **0.9334**
Average	26.00/ 0.7138	28.28/ 0.7973	27.58/ 0.7785	28.35/ 0.7999	28.63/ 0.8070	**28.77**/ 0.8060	28.72/ **0.8096**	28.72/ **0.8096**

Note: Bold indicates the best performance, and underlining indicates the second-best.

Table III
Comparison of computational complexity

Models	**Time Process (GPU)**	**Params**	**MACs**
SRCNN [15]	1.43 ms	69 K	1.56 G
VDSR [13]	4.12 ms	667 K	15.06 G
SRResNet [28]	10.60 ms	1547 K	50.47 G
EDSR [16]	5.98 ms	1776 K	29.47 G
SwinIR [50]	37.99 ms	897 K	33.49 G
E2DSR_S	8.41 ms	1862 K	39.62 G
E2DSR_C	8.88 ms	1862 K	39.62 G

Table IV
Gesture recognition assessment

Gesture	LR image	Bicubic [6]	E2DSR_S
Palm	0.400	0.743	**0.838**
Two up	0.346	0.785	**0.826**
Two up inverted	0.320	0.770	**0.798**
Stop	0.398	0.808	**0.826**
Average	0.366	0.776	**0.822**

such as network pruning and quantization, to develop a more lightweight architecture suitable for real-time deployment on edge devices. Furthermore, we plan to explore the application of the E2DSR framework to other machine-vision domains in which high-fidelity edge reconstruction is critical.

Acknowledgment

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number NCUD.02-2024.09.

References

[1] H. Xiao, Z. Yang, T. Liu, S. Liu, X. Huang, and J. Dai, "Deep learning for medical imaging super-resolution: A comprehensive review," *Neurocomputing*, p. 129667, 2025.

[2] S. Sharma, P. Sinha, R. K. Prajapati, N. Chauhan, and S. K. Yadav, "Enhanced Image Reconstruction and Contextual Intelligence For Cyber-Forensics," in *Proceedings of the 2025 International Conference on Next Generation Information System Engineering (NGISE)*, vol. 1. IEEE, 2025, pp. 1–6.

[3] D. Berardini, L. Migliorelli, A. Galdelli, and M. J. Marín-Jiménez, "Edge artificial intelligence and super-resolution for enhanced weapon detection in video surveillance," *Engineering Applications of Artificial Intelligence*, vol. 140, p. 109684, 2025.

[4] W. Tan, B. Yan, and B. Bare, "Feature super-resolution: Make machine see more clearly," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3994–4002.

[5] X. HoangVan, D. B. Dinh, T. N. Canh, and V.-T. Nguyen, "ESRPCB: An edge guided super-Resolution model and ensemble learning for tiny Printed Circuit Board defect detection," *Engineering Applications of Artificial Intelligence*, vol. 159, p. 111547, 2025.

[6] J. Liu, Z. Gan, and X. Zhu, "Directional bicubic interpolation—A new method of image super-resolution," in *Proceedings of the 3rd International Conference on Multimedia Technology (ICMT-13)*. Atlantis Press, 2013, pp. 463–470.

[7] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the nonlocal-means to super-resolution reconstruction," *IEEE Transactions on image processing*, vol. 18, no. 1, pp. 36–51, 2008.

[8] X. Wang, "Interpolation and sharpening for image upsampling," in *Proceedings of the 2022 2nd International Conference on Computer Graphics, Image and Virtualization (ICCGIV)*. IEEE, 2022, pp. 73–77.

[9] K. A. Kiani and T. Drummond, "Solving robust regularization problems using iteratively re-weighted least squares," in *Proceedings of the 2017 IEEE Winter Conference*

on Applications of Computer Vision (WACV). IEEE, 2017, pp. 483–492.
[10] M. M. Khattab, A. M. Zeki, A. A. Alwan, and A. S. Badawy, "Regularization-based multi-frame super-resolution: a systematic review," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 7, pp. 755–762, 2020.
[11] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.
[12] J. He, L. Yu, Z. Liu, and W. Yang, "Image super-resolution by learning weighted convolutional sparse coding," *Signal, Image and Video Processing*, vol. 15, no. 5, pp. 967–975, 2021.
[13] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.
[14] C. Tian, Y. Zhang, W. Zuo, C.-W. Lin, D. Zhang, and Y. Yuan, "A heterogeneous group CNN for image super-resolution," *IEEE transactions on neural networks and learning systems*, vol. 35, no. 5, pp. 6507–6519, 2022.
[15] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.
[16] B. Lim, S. Son, H. Kim, S. Nah, and 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.
[17] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 457–466.
[18] S. Wang, T. Zhou, Y. Lu, and H. Di, "Detail-preserving transformer for light field image super-resolution," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, 2022, pp. 2522–2530.
[19] D. Marr and E. Hildreth, "Theory of edge detection," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 207, no. 1167, pp. 187–217, 1980.
[20] I. Biederman, "Recognition-by-components: a theory of human image understanding," *Psychological review*, vol. 94, no. 2, p. 115, 1987.
[21] K. Nazeri, H. Thasarathan, and M. Ebrahimi, "Edge-informed single image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
[22] M. Wang, J. Wang, Y. Li, and H. Lu, "Edge computing with complementary capsule networks for mental state detection in underground mining industry," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 7, pp. 8508–8517, 2022.
[23] Z. Jiang, H. Yue, Y.-K. Lai, J. Yang, Y. Hou, and C. Hou, "Deep edge map guided depth super resolution," *Signal Processing: Image Communication*, vol. 90, p. 116040, 2021.
[24] J. Kim, G. Li, I. Yun, C. Jung, and J. Kim, "Edge and identity preserving network for face super-resolution," *Neurocomputing*, vol. 446, pp. 11–22, 2021.
[25] K. Kim and S. Y. Chun, "Sredgenet: Edge enhanced single image super resolution using dense edge detection network and feature merge network," *arXiv preprint arXiv:1812.07174*, 2018.
[26] P. N. Michelini, Y. Lu, and X. Jiang, "Edge-SR: super-resolution for the masses," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 1078–1087.
[27] F. Fang, J. Li, and T. Zeng, "Soft-edge assisted network for single image super-resolution," *IEEE Transactions on Image Processing*, vol. 29, pp. 4656–4668, 2020.
[28] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.
[29] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.
[30] X. Ye, X. Duan, and H. Li, "Depth super-resolution with deep edge-inference network and edge-guided depth filling," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1398–1402.
[31] X. Cheng, X. Li, J. Yang, and Y. Tai, "SESR: Single image super resolution with recursive squeeze and excitation networks," in *Proceedings of the 2018 24th International conference on pattern recognition (ICPR)*. IEEE, 2018, pp. 147–152.
[32] X. Zhang, H. Zeng, and L. Zhang, "Edge-oriented convolution block for real-time super resolution on mobile devices," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 4034–4043.
[33] Y. Wang, "Edge-enhanced feature distillation network for efficient super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 777–785.
[34] L. Hu, L. Hu, and M. Chen, "Edge-enhanced infrared image super-resolution reconstruction model under transformer," *Scientific Reports*, vol. 14, no. 1, p. 15585, 2024.
[35] E. Bamani, E. Nissinman, I. Meir, L. Koenigsberg, and A. Sintov, "Ultra-range gesture recognition using a web-camera in human–robot interaction," *Engineering Applications of Artificial Intelligence*, vol. 132, p. 108443, 2024.
[36] L. Chen, Q. Sun, Z. Xu, Y. Liao, and Z. D. Chen, "A low-resolution infrared gesture recognition method combining weak information reconstruction and joint training strategy," *Digital Signal Processing*, vol. 158, p. 104922, 2025.
[37] Y. Xi, J. Zheng, W. Jia, X. He, H. Li, Z. Ren, and K.-M. Lam, "See clearly in the distance: Representation learning GAN for low resolution object recognition," *IEEE access*, vol. 8, pp. 53 203–53 214, 2020.
[38] R. Li and X. Zhao, "LSwinSR: UAV Imagery Super-Resolution based on Linear Swin Transformer," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
[39] Y. Li, G. Dong, P. Huang, Z. Ma, and X. Wang, "A Gesture Recognition Framework Based on Multi-frame Super-resolution Image Sequence," in *Proceedings of the 2020 Chinese Automation Congress (CAC)*. IEEE, 2020, pp. 4519–4524.
[40] R. Kushwaha, M. Kumar, and D. Kumar, "VRFNet-ASLiT: Fused Deep CNN and Adaptive Super Resolution Transform Based Hand Gesture Recognition," *IEEE Sensors Journal*, 2024.
[41] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.
[42] A. Kapitanov, K. Kvanchiani, A. Nagaev, R. Kraynov, and A. Makhliarchuk, "HaGRID–HAnd Gesture Recognition Image Dataset," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4572–4581.
[43] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," *Advances in neural information processing systems*, vol. 30, 2017.
[44] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis, "Dying relu and initialization: Theory and numerical examples," *arXiv preprint arXiv:1903.06733*, 2019.
[45] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," *British Machine Vision Conference*, 2012.
[46] R. Zeyde, M. Elad, and M. Protter, "On single im-

age scale-up using sparse-representations," in *Proceedings of the International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[47] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5197–5206.

[48] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings of the 8th IEEE international conference on computer vision. ICCV 2001*, vol. 2. IEEE, 2001, pp. 416–423.

[49] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 126–135.

[50] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.

Xiem HoangVan is an Associate Professor at the Faculty of Electronics and Telecommunications, VNU-University of Engineering and Technology, Vietnam. He received his Ph.D. degree from Lisbon University, Portugal, in 2015, his M.Sc. degree from Sungkyunkwan University, South Korea, in 2011, all in Electrical and Computer Engineering. His research interests are machine learning, image, and video communications. Prof. Xiem has published about 100 papers on robotics, image, and video processing and regularly reviews for many renowned IEEE, IET, and EURASIP journals and serves as a technical committee member for international conferences and funding agencies worldwide. He has received several technical awards for his contributions to image and video coding, including 5 Best Paper awards, i.e., at Picture Coding Symposium 2015 (Australia), the International Workshop on Advanced Image Technology 2018 (Thailand), REC-ECIT 2022, IEEE-RIVE 2023, and ATC 2024. He is a recipient of the Fraunhofer Portugal award 2015, the Golden Globe Award for Young Scientists (under 35 years old) in Science and Technology 2019, and the VNU Top Young Scientist Award 2019.

Long Luong Ha is currently an undergraduate student Robotics Engineering at the University of Engineering and Technology, Viet Nam National University (VNU). His current research interests include Machine Vision, Deep Learning, Robotics, and Computer Vision.

Thanh Nguyen Canh (*Graduate Student Member, IEEE*) received his Engineering Degree in Robotics Engineering at the University of Engineering and Technology, Viet Nam National University (VNU) in 2022, and his M.Sc. degree from Japan Advanced Institute of Science and Technology (JAIST), Japan in 2024. Currently, he is pursuing a Ph.D. degree at JAIST. His research interests include Robotics, SLAM, AI, and robot vision. Further info on his homepage: https://thanhnguyencanh.github.io/