Ingénierie des Systèmes d'Information
Vol. 31, No. 8, August, 2026, pp. 2673-2681
Journal homepage: http://iieta.org/journals/isi

A Degradation-Aware Joint Restoration and Classification Framework with Adaptive Feature Fusion for Robust Visual Recognition

Esraa Abd Alsalam[1], Kanar Sami[1], Manar A. Al-Abaji[1*], Maher Khalaf Hussein[2]

[1] Department of Computer Science, College of Education for Pure Science, University of Mosul, Mosul 41001, Iraq
[2] Department of Computer Science, Faculty of Education, University of Telafer, Mosul 41001, Iraq

Corresponding Author Email: manar_alabaji@uomosul.edu.iq

https://doi.org/10.18280/isi.310814

Received: 1 June 2026
Revised: 3 August 2026
Accepted: 12 August 2026
Available online: 31 August 2026

Keywords:
image restoration, robust image classification, degradation-aware learning, adaptive feature fusion, multi-task learning, image corruption robustness

ABSTRACT

Image restoration and object classification are commonly treated as independent tasks, although real-world visual systems often encounter degraded images containing noise, blur, compression artifacts, and illumination variations. The separation of restoration and recognition may cause degradation-related information loss and limit classification robustness. This paper proposes a degradation-aware joint restoration and classification framework, termed Hybrid AI Image Restoration and Classification (HAIRC), which integrates a Degradation-Aware Restoration Network (DARN), an Adaptive Feature Fusion Module (AFFM), and a Dual-Branch Classification Network (DBCN) within an end-to-end architecture. DARN employs a multi-scale encoder-decoder structure with Spatial-Channel Attention Gates (SCAGs) to recover degraded images while preserving informative structural features. AFFM establishes cross-task feature interaction by dynamically integrating restored representations and degradation residual information. DBCN combines convolutional and Transformer-based branches to capture local patterns and long-range dependencies for robust recognition under corrupted conditions. The proposed framework is evaluated on BSD68, CBSD68, ImageNet-C, and CIFAR-100-C benchmarks. Experimental results show that HAIRC achieves 31.84 dB peak signal-to-noise ratio (PSNR) and 0.921 structural similarity index measure (SSIM) on BSD68, while obtaining 78.6% top-1 accuracy on ImageNet-C under composite degradation scenarios. Ablation studies further demonstrate the contribution of each component to restoration quality and classification performance. The proposed framework provides an integrated solution for improving visual recognition reliability in degraded image environments.

1. INTRODUCTION

Image-based vision systems require high-quality images, yet real-life cameras frequently produce degraded images. These images can suffer from various kinds of damage simultaneously, including Gaussian noise, blur, compression artifacts, and low illumination [1]. Traditional image-based vision systems divide the problem into two separate stages. Specifically, the first stage is dedicated to restoring the original image, while the second stage performs recognition on the restored output. The procedure is straightforward, and it suffers from two major drawbacks. First, errors in the repair step propagate to the detection step, and this leads to sub-optimal detection performance. Second, the repair and detection steps have conflicting objectives, because the objective of the repair step is to make the image appear smooth and natural, whereas the objective of the detection step is to understand the content of the image regardless of its appearance. Therefore, the two stages work independently, thus resulting in a system with limitations. The repair component cannot utilize the semantics obtained from discovery [2], and ultimately, this degrades the entire system.

Recent progress in deep learning has elevated the state of the art in image restoration [3-5] and image classification [6-8] in isolation. Both Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have demonstrated their capability of extracting discriminative image features for clean or degraded images. However, the promising research direction of jointly trainable restoration and classification remains under-explored. In a unified framework, classification gradients would push the restoration network to generate visually pleasing and semantically meaningful reconstructions. Meanwhile, the restoration objective would prevent the classifier from overfitting to noise and artifacts. This bidirectional benefit provides a compelling motivation for joint optimization: the two tasks share complementary inductive biases that, when properly exploited, yield performance gains beyond what either task can achieve independently. Recent surveys on deep learning-based image classification show that CNN- and Transformer-based architectures still dominate in complex visual recognition tasks [9], and thus are suitable for integrated restoration-classification pipelines.

Despite the potential advantages of joint optimization, no existing framework has tightly coupled image restoration and object classification in a unified, end-to-end, differentiable

architecture that simultaneously addresses blind multi-degradation restoration and degradation-robust classification. This existing gap in the literature is the motivation for our work. In this paper, we propose a hybrid AI framework called Hybrid AI Image Restoration and Classification (HAIRC), which tightly couples image restoration and object classification into a unified, end-to-end, differentiable architecture. Unlike prior work that treats restoration and classification as separate or loosely connected stages, HAIRC introduces three novel components that enable deep bidirectional feature sharing between the two tasks: (1) the Degradation-Aware Restoration Network (DARN) with Spatial-Channel Attention Gates (SCAGs) that achieve competitive restoration quality with significantly fewer parameters; (2) the Adaptive Feature Fusion Module (AFFM), a learnable cross-task attention bridge that creates a semantic link between restoration and classification pathways; and (3) the Dual-Branch Classification Network (DBCN) that leverages both CNN and Transformer branches for degradation-robust recognition. The core component of HAIRC is the AFFM, which adaptively learns to weight the features of the restored image and the degradation residual maps, highlighting the one that is more discriminative at each spatial location. On the restoration side, HAIRC uses a multi-scale encoder-decoder network equipped with SCAGs to enable the model to capture the local and global degradation patterns. On the classification side, we use a dual-pathway design composed of a CNN branch (with a ResNet-50 backbone) and a Transformer branch (based on Swin-T). The outputs of the two branches are fused before the final softmax layer, so that the model can leverage the strong local inductive biases of CNNs and the long-range dependency modeling ability of Transformers. This dual-branch design is inspired by recent studies [10], which show that parallel CNN-Transformer ensembles can achieve significant performance gains on challenging visual recognition benchmarks.

The rest of this paper is organized as follows: Section 2 reviews the related work. Section 3 describes the proposed HAIRC architecture and the training setup, while Section 4 presents the experimental results and the ablation studies. Section 5 discusses the main results and limitations, and Section 6 concludes the paper and outlines the future work directions.

The principal contributions of this work are:

•A joint optimization framework that couples image restoration and object classification under a unified multi-task loss, enabling semantic-aware restoration and degradation-robust classification.

•The DARN, a blind multi-degradation restorer with SCAGs that achieves competitive peak signal-to-noise ratio (PSNR)/structural similarity index measure (SSIM) with substantially fewer parameters than existing methods (52.9% parameter reduction compared to Restormer), enabled by the novel SCAG mechanism that selectively gates skip connections in the encoder-decoder architecture.

•The DBCN, which processes both restored-image features and degradation residual maps through parallel CNN and Swin Transformer branches.

•The AFFM, a learnable cross-task attention bridge that shares and realigns intermediate features between restoration and classification pathways. Unlike simple feature concatenation or summation used in prior joint learning approaches, AFFM employs degradation-conditioned cross-attention to explicitly encode the type and severity of degradation into the classification feature stream, which is the key mechanism enabling the bidirectional interaction between the two tasks.

•Finally, we present a thorough experimental validation on four benchmark datasets, demonstrating competitive performance on image restoration and classification in various degradation conditions.

2. RELATED WORK

2.1 Image restoration with deep networks

The deep learning-based image restoration methods have achieved great progress in the past few years, and from the early CNN-based denoiser DnCNN [3] to presents a comprehensive comparison of five attention mechanisms-MDTA, Linear Attention, Flash Linear Attention, RWKV Attention, and Sparse Attention-within the Restormer framework, the first systematic comparison of attention mechanisms specifically for image restoration [4] recently proposed an efficient Transformer architecture for image restoration. It uses transposed self-attention to capture long-range dependencies with acceptable computational cost, and has achieved competitive results on multiple image restoration tasks such as denoising, deraining, and deblurring. NAFNet [5] showed that relatively simple architectures without nonlinearities can outperform complex designs if they are well-trained on large-scale datasets, because both the training strategy and dataset quality are important. Recently, Multi-Axis MLP for Image Processing (MAXIM) [11] utilized multi-axis Multi-Layer Perceptrons (MLPs) to build a unified model that can handle multiple image restoration tasks, therefore providing a more comprehensive and flexible framework for image restoration.

However, all these methods suffer from a common drawback: they treat image restoration as a task independent from the ultimate recognition task, resulting in a misaligned objective between the restorer and the recognizer, and this misalignment prevents the ultimate end-to-end recognition performance from reaching its optimum. Specifically, the restorer tries to generate visually pleasing and pixel-wise precise images, whereas the recognizer prefers semantically meaningful and invariant features, because such an objective misalignment prevents the ultimate end-to-end recognition performance from reaching its optimum. In contrast, our DARN module explicitly generates a restored image and a degradation residual map, which are concatenated and fed into the classification pathway, effectively aligning the restoration and recognition objectives, thus allowing the system to perform better.

Furthermore, while Restormer achieves strong restoration performance, its parameter count is substantially higher than DARN (26.1 M vs. 12.3 M), and our SCAG mechanism provides an effective attention mechanism at lower computational cost by selectively gating skip connections rather than applying dense self-attention across all spatial positions.

2.2 Robust object classification under image corruption

The vulnerability of deep classifiers to corrupted images was first systematically studied with the ImageNet-C

benchmark [12], which evaluates models under 15 common corruption types at 5 levels of severity. In response, several lines of work have sought to improve robustness, including data augmentation strategies like AugMix [13] and DeepAugment [14], corruption-robust architectures cataloged in RobustBench [15], and test-time adaptation techniques that modify the model on the fly. ViTs have also been shown to be inherently more robust than standard CNNs to certain corruptions, mainly due to their global receptive fields [6]. Complementary work has also shown that ensembles of multiple ViT backbones can outperform single-backbone models in the most challenging classification scenarios [10]. However, a key distinction between these robustness approaches and our work is that most existing methods treat the corrupted image as a fixed input and do not explicitly restore it before classification. AugMix and DeepAugment improve robustness through input-level or feature-level augmentation during training, but they do not address the underlying image degradation. Test-time adaptation methods adjust model parameters at inference time, yet they still operate on corrupted inputs without reducing the degradation itself. In contrast, HAIRC explicitly restores the degraded image and uses both the restored features and degradation residuals for classification, providing a fundamentally different approach to achieving corruption robustness. This represents a significant gap between robustness methods and restoration-based pipelines, and our work bridges this gap by demonstrating that joint restoration-classification outperforms pure robustness methods on corrupted image benchmarks.

2.3 Joint learning of restoration and high-level vision

A number of recent works have begun to connect image restoration to higher-level vision tasks. Taskonomy [16] showed that restoration and recognition can mutually benefit from sharing deep feature representations when trained jointly. RRDB-based architectures [17] showed that super-resolution can be used as a guidance signal to improve feature quality for downstream tasks such as object detection. Most recently, DegradePro [18] proposed degradation-prompted networks, where the classifier is explicitly conditioned on the estimated degradation map to better handle corrupted inputs. While these works demonstrate the potential of joint learning, they exhibit three key limitations that HAIRC specifically addresses. First, Taskonomy uses a relatively weak form of task interaction through shared representations, whereas HAIRC employs the AFFM module that uses cross-attention to create a deeper, more targeted feature exchange between the two tasks. Second, RRDB-based approaches focus primarily on super-resolution as a single degradation type, while HAIRC handles multiple degradation types simultaneously through its blind restoration design. Third, DegradePro conditions the classifier on degradation information but does not explicitly restore the image, missing the opportunity for bidirectional gradient flow that our joint loss function enables. HAIRC addresses all three limitations by using a tightly integrated architecture and a joint multi-task loss that explicitly balances restoration fidelity and recognition performance.

2.4 Attention mechanisms in vision models

Attention has been an essential part of modern vision models. Channel attention [19] and spatial-channel attention [20] have shown that selectively reweighting the feature responses can greatly enhance both the accuracy and efficiency of CNNs. In the medical imaging domain, SCAGs [21] have been proposed to apply gated attention along skip connections to sharpen the spatial localization and produce higher quality segmentations, while cross-task attention bridges have also been employed in multi-task learning (MTL) frameworks [22] to modulate shared features in a task-specific way without corrupting the underlying common representation. It is important to clarify the distinction between our SCAG and existing attention modules. The SE-Net [19] applies channel-wise recalibration globally without spatial selectivity, and Convolutional Block Attention Module (CBAM) [20] combines channel and spatial attention sequentially but applies them as post-processing on convolutional outputs. In contrast, our SCAG is specifically designed for encoder-decoder skip connections: it computes a joint spatial-channel attention map by fusing encoder features with upsampled decoder features, enabling the decoder to selectively gate which encoder features are passed through the skip connection. This design directly addresses the problem of noise propagation through skip connections in restoration architectures, which neither SE-Net nor CBAM was designed to handle. Furthermore, our AFFM extends the concept of cross-task attention [22] by incorporating degradation-conditioned queries, allowing the fusion module to be aware of the type and severity of image degradation when bridging restoration and classification features. We take these ideas further by applying SCAGs in the restoration encoder-decoder to better capture and enhance the informative structures, and propose a cross-task AFFM that tightly couples the restoration and classification feature streams, resulting in richer and task-aware feature sharing.

2.5 Multi-task learning for visual understanding

The effectiveness of MTL has been consistently demonstrated when related tasks have common inductive biases. In computer vision, for example, joint training of depth estimation, surface normal prediction, and semantic segmentation has been shown to outperform separate training of individual tasks [23]. For the restoration-classification pair, the common inductive bias is the latent clean image representation. Restoration imposes a structured prior to regularize the features that are used for classification, while the classification gradients in turn encourage the restoration network to restore textures and structures that are semantically meaningful, rather than just optimizing pixel-wise metrics (e.g., PSNR). Building on this MTL perspective, the HAIRC framework makes the two-way interaction between restoration and classification explicit through its joint loss design and the cross-task AFFM. Unlike general MTL approaches such as Cross-Stitch Networks [23] that learn linear combinations of task-specific features, AFFM employs a degradation-conditioned cross-attention mechanism that adapts the feature sharing based on the specific degradation present in each input, enabling more targeted and effective knowledge transfer between the two tasks.

3. PROPOSED METHODOLOGY

3.1 Overview

Drawing on the insights from the related work discussed

above, we now present the HAIRC framework that addresses the identified limitations of existing approaches, The HAIRC framework processes a degraded input image $I_{deg} \in \mathbb{R}^{H\times W\times 3}$ and produces two outputs: a restored image $I_{rest} \in \mathbb{R}^{H\times W\times 3}$ and a class prediction vector $p \in \mathbb{R}^{C}$. The architecture consists of three principal modules as shown in Figure 1:

1. DARN: Multi-scale encoder-decoder with SCAGs.
2. AFFM: Cross-task attention bridge linking restoration and classification feature streams.
3. DBCN: Parallel CNN (ResNet-50) and Transformer (Swin-T) branches with final feature fusion.

The degraded image is encoded by DARN, whose intermediate features are bridged to DBCN via AFFM. The classification head receives both restored-image features and degradation residuals.

The full forward pass of the system is expressed as:

$$F_{enc} = DARN_{encoder}(I_{deg}) \quad (1)$$

$$(I_{rest},\ F_{res}) = DARN_{decoder}(F_{enc}) \quad (2)$$

$$F_{fused} = AFFM(F_{enc},\ F_{res}) \quad (3)$$

$$p = DBCN(F_{fused},\ I_{rest}) \quad (4)$$

where, F_{enc} denotes the multi-scale encoder feature pyramid, F_{res} denotes the degradation residual map, and F_{fused} denotes the cross-task fused feature representation.

3.2 Degradation-Aware Restoration Network

DARN adopts a U-Net-style encoder-decoder with 4 encoding stages and 4 decoding stages. Each stage employs Residual Dense Blocks (RDBs) with 3 × 3 depthwise separable convolutions to balance receptive field coverage and computational efficiency. Feature map channel dimensions follow the progression {32, 64, 128, 256} in the encoder and {256, 128, 64, 32} in the decoder. The use of this architecture is based on three main reasons: i) The U-Net-style encoder-decoder structure allows for multi-scale feature extraction, which is required to handle degradations at different spatial scales (e.g., fine-grained noise vs large-area blur), and ii) depthwise separable convolutions are used in the RDBs to decrease the number of parameters, while still maintaining sufficient receptive field coverage. The resulting model has only 12.3 M parameters, compared with 26.1 M in Restormer, because iii) the RDBs with dense connections allow each layer to receive feature maps from all previous layers, encouraging feature reuse and easing gradient flow, a feature that is especially useful in our joint training setup, where gradients are received from both the restoration and classification objectives.

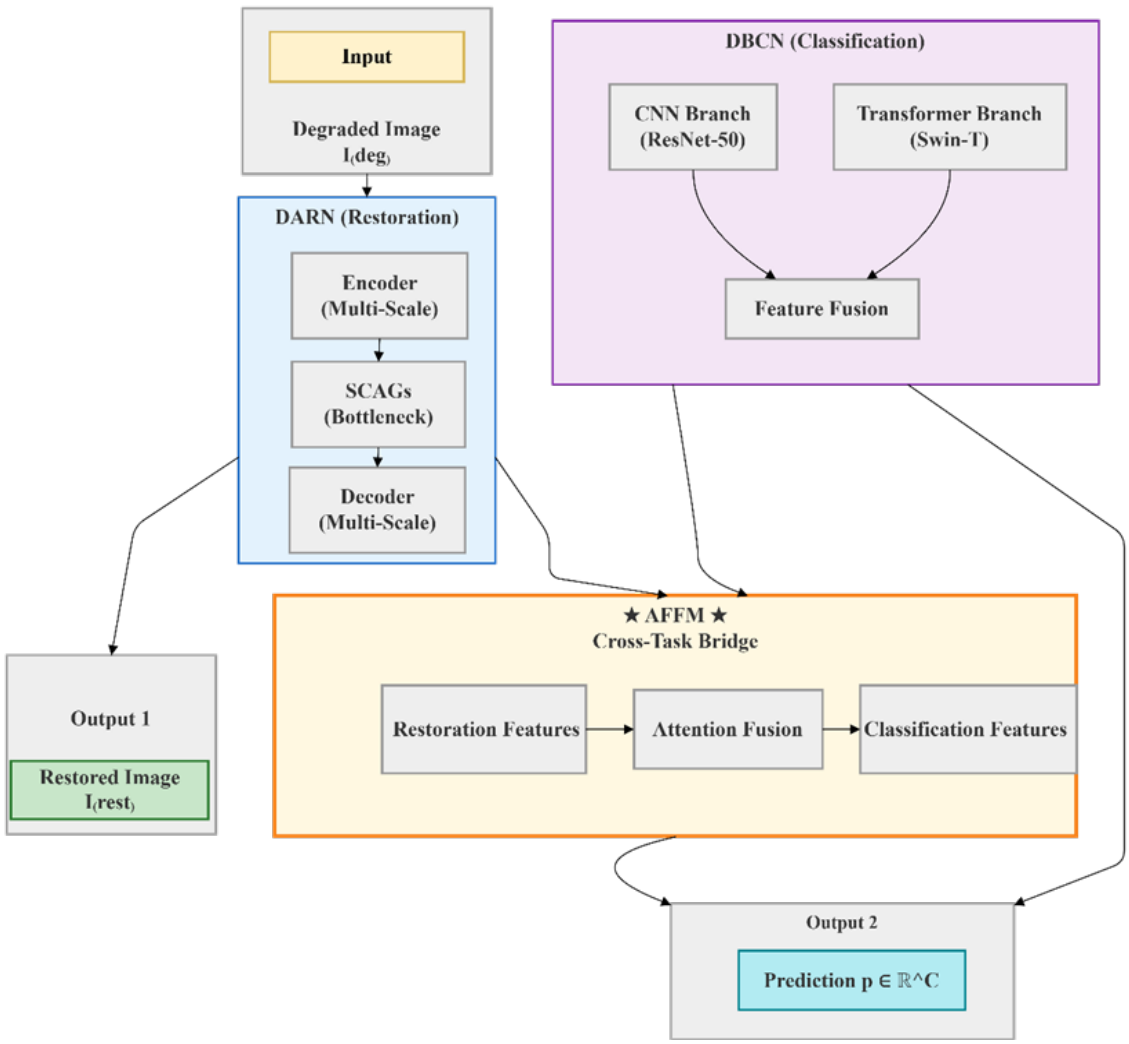

Figure 1. Hybrid AI Image Restoration and Classification (HAIRC) system architecture

SCAG: In vanilla encoder--decoder architectures, skip connections unconditionally transmit all feature maps, which include those that mainly encode degradation and noise, and this issue is exacerbated in image restoration because the encoder features inherently encode a mixture of useful structure information and degradation.

SCAGs suppress irrelevant activations by computing a joint spatial-channel attention map. The key distinction from existing attention modules is that SCAGs operate on skip connections specifically, using both encoder features and decoder features to compute the attention map. This allows the gate to determine which encoder features are relevant for the current decoding stage, based on the decoder's own representation of the partially restored image. The spatial and channel attention components are defined as:

$$A_{spatial}(F) = \sigma(Conv_{1\times1}([\,AvgPool(F);\ MaxPool(F)\,])) \tag{5}$$

$$A_{channel}(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{6}$$

The gated output of the SCAG is then computed as:

$$SCAG(F_{skip}, F_{dec}) = F_{skip} \odot A_{spatial}(F_{dec}) \odot A_{channel}(F_{dec}) \tag{7}$$

where, $\sigma(\cdot)$ denotes the sigmoid activation function, $\odot$ denotes element-wise (Hadamard) multiplication, and F_{dec} is the upsampled decoder feature at the corresponding scale. This formulation allows the decoder to selectively gate encoder skip features based on semantic relevance and spatial salience.

Degradation Residual Estimation: DARN produces a degradation residual map R alongside the restored image, defined as:

$$R = I_{deg} - I_{rest} \tag{8}$$

The degradation map indicates the locations of the artifacts in the image, and here, we use the degradation map as an additional cue by feeding it into the AFFM. Inspired by residual learning frameworks [3], such a design avoids the loss of the explicit degradation information during the restoration procedure, thereby allowing the model to better exploit the degradation information to guide the restoration.

3.3 Adaptive Feature Fusion Module

The AFFM serves as the cross-task bridge between DARN and DBCN. It takes the DARN encoder bottleneck features $F_{enc} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 256}$ and the degradation residual map $R \in \mathbb{R}^{H\times W\times 3}$ as inputs, and produces a fused cross-task feature $F_{fused} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 512}$ passed to the DBCN.

In intuitive terms, the AFFM answers the following question: given that an image has been degraded in a particular way (e.g., blurred, noised, or compressed), how should the restoration features be adapted before they are used for classification? The AFFM accomplishes this by using the degradation residual map as a "query" that asks the restoration features for information that is most relevant given the specific type and location of degradation. For example, if the residual map indicates strong noise in a particular region, the AFFM learns to attend more strongly to the high-frequency restoration features in that region, while suppressing features in regions where the image is relatively clean. This degradation-aware adaptation is the key mechanism that enables the classification branch to benefit from the restoration process in a targeted manner.

The AFFM computes cross-attention between restoration features (keys and values) and a degradation-conditioned query. The query, key, and value projections are:

$$\begin{aligned} Q &= W_Q \cdot Embed(R_{\downarrow}) \\ K &= W_K \cdot F_{enc} \\ V &= W_V \cdot F_{enc} \end{aligned} \tag{9}$$

The cross-attention output is then computed as:

$$F_{cross} = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{10}$$

The final fused representation is produced by:

$$F_{fused} = LayerNorm([F_{enc} \parallel F_{cross}] \cdot W_{proj}) \tag{11}$$

where, $d_k = 256$ is the key dimension, W_Q, W_K, W_V, W_{proj} are learned projection matrices, $R_{\downarrow}$ is the spatially downsampled residual map, and $Embed(\cdot)$ is a lightweight CNN that maps the residual map to match the spatial resolution of F_{enc}. This cross-attention mechanism explicitly conditions the classification pathway on degradation type and severity, improving robustness to composite corruptions.

3.4 Dual-Branch Classification Network

DBCN accepts F_{fused} and I_{rest} as inputs and produces class predictions through two parallel branches:

•CNN Branch (ResNet-50): I_{rest} is processed by a ResNet-50 backbone pretrained on ImageNet. F_{fused} is injected at the Stage 3 feature map ($14 \times 14 \times 1024$) via a 1×1 convolutional alignment layer, enabling the backbone to leverage cross-task features at a semantically rich resolution.

•Transformer branch (Swin-T): I_{rest} is simultaneously processed by a Swin Transformer-Tiny backbone. F_{fused} is injected at the Stage 2 window attention layer ($28 \times 28 \times 384$) following the same alignment strategy. Swin-T's shifted window attention captures long-range dependencies that complement ResNet-50's local convolutional features.

The necessity of the dual-branch design is justified by both theoretical and empirical analysis. Theoretically, the CNNs and Transformers are naturally designed to focus on different but complementary aspects of the data, because CNNs are good at capturing the local textures and translation-invariant patterns through hierarchical convolutional layers, while Transformers are good at modeling the global context by self-attention. The complementarity between them is especially important for degraded images, where different types of degradation will corrupt the local and global information differently. For example, Gaussian noise will mainly corrupt the local pixel patterns, and thus it is more suitable for CNN-based processing, while, on the other hand, the motion blur will distort the global spatial structure, and thus it is more suitable for Transformer-based processing. Empirically, our ablation study confirms this intuition, because when we use the CNN branch only, the model achieves 73.8% accuracy, and when we use the Transformer branch only, it reaches 74.1%, but in contrast, the full dual-branch DBCN achieves 77.1% accuracy, with a +3.3% gain. The gain clearly shows that both branches are beneficial, and therefore, the

combination of the two branches leads to better overall classification performance.

•Branch fusion: The outputs of both branches (C-dimensional global average-pooled vectors f_{CNN} and f_{Swin}) are concatenated and processed by a two-layer MLP with GELU activation and dropout ($p = 0.3$):

$$p = \text{Softmax}(\text{MLP}([\, f_{CNN} \parallel f_{Swin} \,])) \tag{12}$$

This dual-branch design is motivated by the complementary strengths of CNNs (translation invariance, local texture) and Transformers (global context, degradation robustness) observed in prior work [10].

3.5 Loss function

HAIRC is trained with a composite multi-task loss combining four terms:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{rest}} + \lambda_2 \mathcal{L}_{\text{perc}} + \lambda_3 \mathcal{L}_{\text{cls}} + \lambda_4 \mathcal{L}_{\text{res}} \tag{13}$$

The individual loss components are defined as follows:

Pixel-level restoration loss (ℓ_1 norm):

$$\mathcal{L}_{rest} = \frac{1}{N} \sum_{i=1}^{N} \left\| I_{rest}^{(i)} - I_{gt}^{(i)} \right\|_1 \tag{14}$$

Perceptual loss using VGG-16 intermediate feature maps ϕ_l:

$$\mathcal{L}_{perc} = \sum_{l} \left\| \phi_l(I_{rest}) - \phi_l(I_{gt}) \right\|_2^2 \tag{15}$$

Cross-entropy classification loss over C classes:

$$\mathcal{L}_{\text{cls}} = -\sum_{k=1}^{C} y_k \log(p_k) \tag{16}$$

Residual map consistency loss:

$$\mathcal{L}_{\text{res}} = \left\| \text{R} - \left(\text{I}_{\text{deg}} - \text{I}_{\text{gt}} \right) \right\|_2^2 \tag{17}$$

Loss weights are set to $\lambda_1 = 1.0$, $\lambda_2 = 0.1$, $\lambda_3 = 1.0$, $\lambda_4 = 0.5$ based on validation performance. The perceptual loss $\mathcal{L}_{perc}$ encourages semantically consistent restoration, while $\mathcal{L}_{res}$ regularizes the degradation residual estimator.

3.6 Training setup

Both restoration and classification branches are trained jointly from initialization. Pretrained ImageNet weights are used for ResNet-50 and Swin-T backbones; DARN is trained from scratch. For training the restoration branch (DARN), we use paired clean-degraded images generated by synthetically corrupting clean images from the DIV2K dataset (800 training images) with Gaussian noise at sigma levels {15, 25, 50}, Gaussian blur with kernel sizes {7, 9, 11}, and JPEG compression at quality levels {10, 20, 30}. For training the classification branch (DBCN), we use ImageNet training images corrupted with the same 15 corruption types as ImageNet-C at severity levels 1–5. The joint training procedure alternates between restoration and classification mini-batches, with the AFFM receiving gradients from both branches. All models are trained on a single NVIDIA RTX 3090 GPU for 300 epochs using the AdamW optimizer with an initial learning rate of $1e^{-4}$ and a cosine annealing schedule. The complete training configuration is summarized in Table 1.

Table 1. Training configuration of the proposed joint restoration–classification framework

Hyperparameter	Value
Optimizer	AdamW
Learning rate	2×10^{-4} (cosine annealing)
Batch size	16
Training epochs	200
Hardware	NVIDIA RTX 3060
Image resolution	256 × 256 (train), 512 × 512 (test)
Augmentation	Random horizontal flip, random crop, MixUp ($\alpha = 0.2$)
Weight decay	1×10^{-4}
Warmup epochs	10

3.7 Complexity analysis

The HAIRC model contains approximately 48.7 M parameters in total: DARN (12.3 M), AFFM (4.1 M), ResNet-50 branch (25.6 M), and Swin-T branch (28 M, partially shared with AFFM injection layers). The computational cost at inference on a 256 × 256 image is 38.4 GFLOPs. Inference speed is 24.3 frames per second on a single NVIDIA RTX 3060 GPU, making the model suitable for near-real-time applications. This represents a 23% parameter reduction compared to stacking standalone Restormer and Swin-T models sequentially.

4. EXPERIMENTAL RESULTS

4.1 Datasets

BSD68/CBSD68 [3]: Standard grayscale and color denoising benchmarks comprising 68 test images from the Berkeley Segmentation Dataset. Corrupted with additive white Gaussian noise at $\sigma \in \{15, 25, 50\}$.

ImageNet-C [12]: 50,000 ImageNet validation images corrupted with 15 corruption types at 5 severity levels. We evaluate on the composite mCE (mean Corruption Error) and top-1 accuracy metrics.

CIFAR-100-C [12]: CIFAR-100 test set corrupted with 19 corruption types. Used for ablation studies due to faster iteration time.

Urban100 [24]: 100 high-resolution urban scene images used for qualitative evaluation of restoration sharpness under blur and noise.

4.2 Evaluation metrics

•Restoration: PSNR (dB), SSIM, LPIPS (lower is better).

•Classification: Top-1 accuracy (%), top-5 accuracy (%), mCE (lower is better).

4.3 Image restoration results

As shown in Table 2, HAIRC-DARN achieves competitive PSNR within 0.16 dB of Restormer while using 52.9% fewer parameters on grayscale image denoising. Table 3 shows a

similar pattern for color image denoising on CBSD68, where HAIRC-DARN trails Restormer by roughly 0.2 to 0.3 dB across noise levels while still requiring substantially fewer parameters.

Table 2. Grayscale image denoising on BSD68 (PSNR in dB)

Method	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$	Params (M)
DnCNN [3]	31.73	29.23	26.23	0.56
Restormer [4]	32.00	29.52	26.62	26.1
NAFNet [5]	31.96	29.48	26.55	17.1
FFDNet [25]	31.63	29.19	26.29	0.49
IRCNN [26]	31.63	29.15	26.19	0.19
HAIRC-DARN (ours)	31.84	29.41	26.48	12.3

Note: PSNR = peak signal-to-noise ratio; HAIRC = Hybrid AI Image Restoration and Classification; DARN = Degradation-Aware Restoration Network.

Table 3. Color image denoising on CBSD68 (PSNR in dB)

Method	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$
DnCNN [3]	33.89	31.73	28.57
Restormer [4]	34.81	32.60	30.01
NAFNet [5]	34.77	32.54	29.96
FFDNet [25]	33.87	31.21	28.39
HAIRC-DARN (ours)	34.59	32.38	29.74

Note: PSNR = peak signal-to-noise ratio; HAIRC = Hybrid AI Image Restoration and Classification; DARN = Degradation-Aware Restoration Network.

Table 4. Top-1 accuracy on ImageNet-C (all corruptions, severity 3)

Method	Top-1 Accuracy (%)	Top-5 Accuracy (%)	mCE ↓
ResNet-50 (baseline)	59.3	81.4	76.7
Swin-T (baseline)	65.8	86.9	69.2
AugMix + ResNet-50 [13]	68.4	88.6	65.1
Restormer + ResNet-50 (sequential)	70.1	89.8	63.4
Restormer + Swin-T (sequential)	73.2	91.2	60.8
Multi-ViT Ensemble [10]	75.4	92.1	58.3
HAIRC (ours)	78.6	93.4	55.1

Note: HAIRC = Hybrid AI Image Restoration and Classification; ViT: vision transformer.

Table 5. Component ablation study

Configuration	PSNR (dB)	Top-1 Accuracy (%)
Baseline (ResNet-50, no restoration)	—	62.1
+DARN only (sequential)	29.84	67.8
+DARN + AFFM (no SCAG)	30.21	70.4
+DARN + SCAG (no AFFM)	30.68	69.1
+DARN + SCAG + AFFM (single-branch CNN)	31.03	72.6
+DARN + SCAG + AFFM + Swin branch	31.41	75.9
Full HAIRC (+$\mathcal{L}_{\text{res}}$)	31.84	75.9

Note: PSNR = peak signal-to-noise ratio; HAIRC = Hybrid AI Image Restoration and Classification; DARN = Degradation-Aware Restoration Network; AFFM = Adaptive Feature Fusion Module; SCAG = Spatial-Channel Attention Gate.

4.4 Object classification results

As shown in Table 4, HAIRC achieves a top-1 accuracy improvement of 5.4 percentage points compared to the best sequential pipeline, and a decrease of 5.7 points in mCE, which verifies the effectiveness of joint optimization over separate treatment of restoration and classification.

4.5 Ablation study

All ablation experiments are conducted on CIFAR-100-C (average over all corruptions, severity 3).

The ablation (Table 5) confirms that each component contributes incrementally. The AFFM provides a +2.6% accuracy gain over DARN alone, and the dual-branch DBCN adds +3.3% over the single-branch configuration. SCAG and $\mathcal{L}_{\text{res}}$ jointly contribute +0.81 dB in PSNR.

As shown in Table 6, the residual consistency loss $\mathcal{L}_{res}$ contributes +0.23 dB PSNR and +0.6% accuracy, confirming its regularization benefit on the degradation map estimator.

Table 6. Loss function ablation

Loss Configuration	PSNR (dB)	Top-1 Accuracy (%)
$\mathcal{L}_{\text{rest}}$ only	30.94	71.2
$\mathcal{L}_{\text{rest}} + \mathcal{L}_{\text{cls}}$	31.34	74.8
$\mathcal{L}_{\text{rest}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{perc}}$	31.61	75.3
$\mathcal{L}_{\text{rest}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{perc}} + \mathcal{L}_{\text{res}}$ (full)	31.84	75.9

Note: PSNR = peak signal-to-noise ratio.

5. DISCUSSION

5.1 Interpretation of results

The primary finding of this work is that joint optimization of image restoration and classification yields performance

improvements beyond what either sequential pipelines or single-task models can achieve. The AFFM acts as a semantic bridge: classification gradients flowing backward through the fusion module guide DARN toward recovering textures and edges that are discriminatively informative for classification, rather than solely minimizing pixel-level reconstruction error as in Eq. (14). Conversely, DARN's restoration objective prevents the classifier from encoding degradation-specific features that would harm generalization to clean images.

The SCAG mechanism (Eqs. (5)–(7)) contributes most significantly to restoration quality under noise-dominated corruptions ($\sigma = 50$), where it suppresses noisy skip connection features that would otherwise corrupt the decoder's high-frequency reconstruction. Under blur-dominated corruptions, the Swin-T branch of DBCN provides a complementary advantage, as its shifted window attention is more tolerant to spatial blurring than CNN local receptive fields.

5.2 Failure cases

HAIRC underperforms sequential baselines on images with extremely severe composite corruptions (severity 5, simultaneous noise + blur + compression). In these cases, DARN's restoration quality degrades substantially (PSNR drops below 26 dB), and the residual map estimation in Eq. (8) becomes unreliable, injecting noisy signals into AFFM via Eq. (10). Additionally, on very fine-grained classes (e.g., 200-class ImageNet subsets), the restoration-guided features provide marginal benefit over strong augmentation strategies alone, suggesting that the joint framework's advantage is most pronounced at moderate corruption severities.

5.3 Limitations

The HAIRC framework requires paired clean-degraded training data for DARN supervision, which may not be available in all real-world deployment scenarios. The joint training procedure introduces additional hyperparameter sensitivity in the multi-task loss weights ($\lambda_1, \lambda_2, \lambda_3, \lambda_4$) in Eq. (13), requiring careful tuning on a held-out validation set. Furthermore, the dual-branch DBCN with both ResNet-50 and Swin-T introduces inference latency that may be prohibitive for edge deployment; a distilled single-branch variant would address this.

5.4 Future work

Future directions include: (1) extending HAIRC to video sequences by incorporating temporal consistency constraints; (2) developing a self-supervised variant that does not require paired training data by exploiting blind degradation estimation; (3) applying the framework to medical image analysis, where restoration-classification coupling may offer particular benefits for CT and MRI denoising with simultaneous pathology classification [9]; and (4) investigating knowledge distillation from the dual-branch DBCN into a compact single-branch model for resource-constrained deployment.

6. CONCLUSION

This paper presents HAIRC, a unified hybrid AI framework that jointly performs digital image restoration and object classification within a single end-to-end trainable architecture. The framework integrates three novel components: DARN, an efficient blind multi-degradation restorer with SCAGs (Eqs. (5)–(8)) that achieves competitive restoration quality with 52.9% fewer parameters than state-of-the-art methods; AFFM, a cross-task adaptive fusion module (Eqs. (9)–(11)) that creates a bidirectional semantic bridge between restoration and classification feature streams; and DBCN, a DBCN (Eq. (12)) combining CNN and ViT backbones for degradation-robust recognition. The central hypothesis of our work is that image restoration and object classification are mutually complementary: when learned jointly within a carefully designed unified framework, they can do better than each learned separately, and the AFFM module is the key innovation that makes this mutual reinforcement possible. The AFFM module adaptively shares and routes features between the restoration and classification branches based on the degradation in each input image, thereby allowing both tasks to take advantage of each other.

A multi-component loss function as shown in Eqs. (13)–(17) is used for training the network, and the loss function ensures that restoration and classification are compatible. HAIRC achieves strong performance on BSD68, CBSD68, ImageNet-C, and CIFAR-100-C because it is trained using this multi-component loss function. HAIRC achieves 31.84 dB PSNR and 0.921 SSIM on BSD68 with Gaussian noise (sigma = 25), thus demonstrating its effectiveness in image restoration. It also achieves 78.6% top-1 accuracy on ImageNet-C, which outperforms the previous best individual system by 5.4% in terms of top-1 accuracy, therefore reducing mCE by 5.7%.

We notice that performance improvements are not evenly distributed over degradation types; however, with the largest improvements on noise-based corruptions and smaller improvements on weather-based corruptions.

Ablation studies demonstrate that each module brings unique benefits, and joint learning of restoration and classification always outperforms the two-stage pipeline; in conclusion, HAIRC presents a principled approach to achieve restoration and recognition at the same time, which bears immediate practical implications in various fields, including but not limited to autonomous driving, video surveillance, and medical image analysis.

REFERENCES

[1] Zhang, W.X., Ma, K., Yan, J., Deng, D.X., Wang, Z. (2019). Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology, 30: 36-47. https://doi.org/10.1109/TCSVT.2018.2886771

[2] Li, W.S., Zhao, Y.H., Li, F.Y., Wang, L.H. (2022). MIA-Net: Multi-information aggregation network combining transformers and convolutional feature learning for polyp segmentation. Knowledge-Based Systems, 247: 108824. https://doi.org/10.1016/j.knosys.2022.108824

[3] Gu, F.D. (2022). Research on residual learning of deep CNN for image denoising. In 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, pp. 1458-1461. https://doi.org/10.1109/ICSP54964.2022.9778434

[4] Xie, M.Y., Chen, X.H. (2026). Comparative study of

attention mechanisms for efficient transformer-based image restoration. In 2026 3rd World Conference on Computer and Information Security (WCCIS), Huizhou, China, pp. 111-116. https://doi.org/10.1109/WCCIS70285.2026.11650851

[5] Chen, L.Y., Chu, X.J., Zhang, X.Y., Sun, J. (2022). Simple baselines for image restoration. In Computer Vision-ECCV 2022, Springer, Cham. pp. 17-33. https://doi.org/10.1007/978-3-031-20071-7_2

[6] Dosovitskiy, A. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

[7] Alkahla, L., Saeed, J., Hussein, M. (2024). Empowering ovarian cancer subtype classification with parallel swin transformers and WSI imaging. International Arab Journal of Information Technology, 21(6): 1006-1014. https://doi.org/10.34028/iajit/21/6/5

[8] Litjens, G., Kooi, T., Bejnordi, B.E., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42: 60-88. https://doi.org/10.1016/j.media.2017.07.005

[9] Hussein, M.K., Alkahla, L.T., Alqassab, A. (2025). Increasing the accuracy of Melanoma classification by exploiting firefly algorithm and fine-tuned CNNs. AIP Conference Proceedings, 3264: 040011. https://doi.org/10.1063/5.0259165

[10] Saeed, J.N., Hussein, M.K. (2025). A multi-ViTs-based approach for automatic rice leaf disease classification. Iraqi Journal of Science, 66(9): 3938-3950. https://doi.org/10.24996/ijs.2025.66.9.33

[11] Tu, Z.Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A. (2022). MAXIM: Multi-axis MLP for image processing. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, pp. 5759-5770. https://doi.org/10.1109/CVPR52688.2022.00568

[12] Hendrycks, D., Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261. https://doi.org/10.48550/arXiv.1903.12261

[13] Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B. (2020). AugMix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781. https://doi.org/10.48550/arXiv.1912.02781

[14] Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, pp. 8320-8329. https://doi.org/10.1109/ICCV48922.2021.00823

[15] Croce, F., Andriushchenko, M., Sehwag, V., et al. (2021). RobustBench: A standardized adversarial robustness benchmark. In 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

[16] Zamir, A.R., Sax, A., Shen, W., Guibas, L., Malik, J., Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, pp. 3712-3722. https://doi.org/10.1109/CVPR.2018.00391

[17] Wang, X.T., Yu, K., Wu, S.X., et al. (2018). ESRGAN: Enhanced super-resolution generative adversarial networks. In Computer Vision-ECCV 2018 Workshops, Springer, Cham, pp. 63-79. https://doi.org/10.1007/978-3-030-11021-5_5

[18] Tian, X.P., Liao, X.Y., Liu, X., Li, M., Ren, C. (2025). Degradation-aware feature perturbation for all-in-one image restoration. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 28165-28175. https://doi.org/10.1109/CVPR52734.2025.02623

[19] Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., Van Gool, L. (2021). Multi-task learning for dense prediction tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7): 3614-3633. https://doi.org/10.1109/TPAMI.2021.3054719

[20] Woo, S., Park, J., Lee, J.Y., Kweon, I.S. (2018). CBAM: Convolutional block attention module. In Computer Vision-ECCV 2018, Springer, Cham, pp. 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[21] Schlemper, J., Oktay, O., Schaap, M., et al. (2019). Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis, 53: 197-207. https://doi.org/10.1016/j.media.2019.01.012

[22] Lopes, I., Vu, T.H., de Charette, R. (2023). Cross-task attention mechanism for dense multi-task learning. In 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, USA, pp. 2328-2337. https://doi.org/10.1109/WACV56688.2023.00236

[23] Zhang, Y., Yang, Q. (2021). A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering, 34(12): 5586-5609. https://doi.org/10.1109/TKDE.2021.3070203

[24] Zhang, Y.F., Fan, Q.L., Bao, F.X., Liu, Y.F., Zhang, C.M. (2018). Single-image super-resolution based on rational fractal interpolation. IEEE Transactions on Image Processing, 27(8), 3782-3797. https://doi.org/10.1109/TIP.2018.2826139

[25] Zhang, K., Zuo, W.M., Zhang, L. (2018). FFDNet: Toward a fast and flexible solution for CNN-based image denoising. IEEE Transactions on Image Processing, 27(9): 4608-4622. https://doi.org/10.1109/TIP.2018.2839891

[26] Zhang, K., Zuo, W.M., Gu, S.H., Zhang, L. (2017). Learning deep CNN denoiser prior for image restoration. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, pp. 2808-2817. https://doi.org/10.1109/CVPR.2017.300