ISSN: 2519-9439

Instance-Aware Image Restoration under Unknown Real-World Degradations and Resolution Variations

Chen Han; Jun Shi

Department of Computer Science, Hong Kong Chu Hai College, Hong Kong 999077, China

Abstract

The problem of image restoration has traditionally been constrained by assumptions of synthetic, uniformly applied degradation models, such as bicubic downsampling or Gaussian noise, which consistently fail to generalize to the immense complexity of real-world scenarios. Furthermore, the advent of varied imaging sensors has introduced continuous resolution variations, rendering fixed-scale models largely obsolete. This paper presents a novel instance-aware image restoration framework designed to operate under completely unknown real-world degradations and continuous resolution variations. By conceptualizing the image not as a homogeneous field of pixels but as a composition of distinct semantic instances, the proposed methodology dynamically adapts its restoration filters based on localized, instance-specific degradation estimations. This localized approach allows the system to differentiate between the varying restoration requirements of complex foreground subjects and heterogeneous background textures. Through the integration of an implicit neural representation module, the framework achieves continuous resolution adaptation, circumventing the limitations of discrete scale-specific upsampling modules. The synergistic coupling of instance-aware degradation representation learning and continuous coordinate-based decoding constructs a robust pipeline capable of handling unprecedented variations in input fidelity. Extensive empirical evaluations on diverse real-world datasets demonstrate that the proposed architecture establishes a new paradigm in blind image restoration, achieving superior quantitative metrics and preserving high-frequency structural details across arbitrary magnification factors.

Keywords

Image Restoration, Instance-Aware Learning, Blind Degradation, Implicit Neural Representations.

1.Introduction

1.1 Background and Motivation

Image restoration represents a fundamental inverse problem in computer vision, aiming to recover a high-quality, high-resolution original image from its degraded, low-resolution observation. Over the past decades, this domain has witnessed substantial paradigm shifts, predominantly driven by the continuous evolution of deep learning architectures. Early approaches heavily relied on simplified, mathematically tractable degradation models [1]. For example, super-resolution tasks typically assumed that the low-resolution input was generated by applying a predetermined blur kernel followed by a uniform downsampling operation, often

ISSN: 2519-9439

utilizing bicubic interpolation [2]. Denoising tasks similarly operated under the assumption of additive white Gaussian noise with a known variance [3]. While these idealized models facilitated the rapid development of convolutional neural networks and enabled direct comparisons between competing algorithmic designs, they inherently restricted the practical applicability of the resulting solutions [4]. When deployed in real-world scenarios, images captured by physical sensors such as smartphones, surveillance cameras, or autonomous vehicle perception systems are afflicted by a complex, nonlinear amalgamation of sensor noise, optical aberrations, motion blur, and sophisticated image signal processing pipeline artifacts [5]. This profound discrepancy between the synthetic training distributions and the empirical testing distributions results in severe performance degradation, manifesting as over-smoothed textures, amplified noise artifacts, or the introduction of physically implausible structural hallucinations [6]. Consequently, the trajectory of contemporary image restoration research has decisively pivoted towards blind restoration, which explicitly avoids reliance on predefined degradation formulations and instead seeks to recover structural fidelity in the presence of completely unknown, real-world corruptions [7].

1.2 Challenges in Current Restoration Paradigms

Despite the substantial progress achieved in blind image restoration, significant challenges remain unresolved, particularly concerning spatial heterogeneity and dimensional variability. The first major challenge is the inherent assumption of spatial invariance maintained by the majority of existing architectures. Current state-of-the-art models generally apply a uniform restoration policy across the entire spatial extent of the image [8]. However, real-world degradations are seldom spatially uniform. The focal plane of a camera, depth-of-field effects, object motion, and localized illumination variations dictate that different regions of an image suffer from vastly different degrees and types of degradation [9]. Moreover, the human visual system exhibits varying degrees of sensitivity to different semantic contents; artifacts introduced on a human face or coherent text are far more perceptible and detrimental to perceived quality than identical artifacts located in a stochastic background texture like foliage or water [10]. This necessitates a transition from globally homogeneous processing to an instance-aware methodology, wherein the network can identify distinct semantic entities and dynamically modulate its feature extraction and restoration parameters to suit the specific localized degradation and semantic importance of each instance [11]. The second critical challenge involves the rigidity of current models regarding resolution variations. Conventional super-resolution networks are meticulously engineered and optimized for integer upsampling factors, primarily operating at fixed scales [12]. In practical applications, however, users frequently require arbitrary zoom capabilities, and input images originate from sensors with highly heterogeneous native resolutions [13]. Training discrete models for every conceivable magnification scale is computationally prohibitive and architecturally redundant [14]. Therefore, integrating continuous resolution adaptation mechanisms, allowing a single model to seamlessly render outputs at arbitrary, non-integer scales, has emerged as a crucial requirement for next-generation image restoration systems [15].

1.3 Contributions and Paper Organization

To address these interconnected challenges, this paper proposes an innovative instance-aware image restoration framework capable of dynamically adapting to unknown real-world degradations while providing continuous resolution synthesis. By fundamentally rethinking the feature extraction pipeline, we introduce a mechanism that disentangles the degradation representation from the semantic content representation at the instance level. We leverage

ISSN: 2519-9439

region-proposal networks to identify salient instances and utilize a localized contrastive learning strategy to explicitly model the spatial variability of the degradation field [16]. These localized representations subsequently inform a dynamic modulation layer that adjusts the restoration filters in a spatially variant manner [17]. Furthermore, we incorporate an advanced implicit neural representation decoding scheme that queries continuous coordinate spaces, thereby eliminating the necessity for discrete upsampling layers and facilitating smooth, arbitrary-scale resolution variations [18]. The primary contributions of this research can be summarized across three dimensions. First, we formulate a novel instance-aware degradation estimation module that accurately captures spatially variant real-world corruptions, enabling customized restoration policies for distinct semantic regions. Second, we integrate a continuous coordinate-based decoding architecture that harmonizes with the instance-aware features to provide arbitrary-scale super-resolution without the constraints of fixed-scale network topologies. Third, we conduct comprehensive evaluations across multiple real-world benchmark datasets, conclusively demonstrating that the synergistic combination of instance awareness and continuous resolution mapping yields unprecedented visual quality and structural fidelity under severe, unknown degradations. The remainder of this paper is structured systematically to provide a thorough exploration of the proposed framework. Section 2 delves into the literature review and theoretical background, tracing the evolution of relevant technologies. Section 3 meticulously details the methodology and system architecture, providing the mathematical formulations of our instance-aware modules and resolution-invariant decoders. Section 4 presents the experimental setup, detailing the rigorous evaluation protocols, quantitative results, qualitative analysis, and ablation studies. Finally, Section 5 offers concluding remarks, acknowledging current limitations and outlining promising trajectories for future research.

2. Literature Review and Theoretical Background

2.1 Evolution of Deep Image Restoration

The application of deep learning to image restoration has fundamentally transformed the landscape of low-level computer vision. Early milestones were characterized by relatively shallow convolutional neural networks that sought to learn a direct end-to-end mapping from low-quality inputs to high-quality targets [19]. These initial architectures primarily focused on the super-resolution task, utilizing mean squared error as the optimization objective, which, while effective at maximizing peak signal-to-noise ratio, frequently resulted in overly smooth and perceptually unappealing outputs. Subsequent advancements addressed this limitation through the introduction of generative adversarial networks, which leveraged adversarial loss and perceptual loss functions to incentivize the generation of realistic, high-frequency textures [20]. This shift from distortion-oriented optimization to perception-oriented optimization represented a critical juncture, acknowledging that mathematical fidelity does not inherently correlate with human visual satisfaction. Parallel to the evolution of loss functions, architectural innovations have continuously pushed the boundaries of performance. Residual learning, dense connections, and channel attention mechanisms were rapidly assimilated into restoration networks, allowing for the construction of increasingly deep and complex models capable of capturing highly abstract feature representations [21]. More recently, the extraordinary success of transformer architectures in natural language processing has cascaded into the vision domain. Vision transformers, utilizing self-attention mechanisms, have demonstrated a remarkable ability to model long-range dependencies and global contexts, surpassing the localized receptive fields inherent to standard convolutional kernels [22]. However, the quadratic computational complexity of global self-attention poses significant challenges when

ISSN: 2519-9439

processing high-resolution images, leading to the development of localized window-based attention schemes and hybrid architectures that attempt to balance computational efficiency with global receptive capabilities [23]. Despite these architectural marvels, the majority of these models remain heavily reliant on synthetic degradation pipelines during training, limiting their real-world efficacy.

2.2 Blind Restoration and Unknown Degradations

The transition towards blind image restoration reflects the growing consensus that synthetic degradation models are fundamentally inadequate for real-world deployment. The classical formulation of degradation typically involves a combination of blur, downsampling, and noise addition. In real-world scenarios, however, the degradation process is a highly complex sequence of operations inherent to the camera hardware and software processing pipelines, including lens distortion, sensor noise, demosaicing artifacts, and lossy compression [24]. To tackle blind restoration, researchers have primarily explored two distinct avenues. The first approach involves explicitly estimating the unknown degradation parameters directly from the input image. These methods often employ a subsidiary network tasked with predicting blur kernels or noise profiles, which are subsequently fed into the primary restoration network to condition the recovery process [25]. While conceptually sound, these explicit estimation techniques frequently falter when confronted with severe degradations, as the initial estimation process itself is highly ill-posed and prone to catastrophic failure. The second, more prevalent approach relies on the construction of immensely complex, randomized degradation synthesis pipelines during the training phase. By exposing the network to an astronomically large distribution of potential degradations, encompassing combinations of various blurs, noises, and compression artifacts, the model is forced to learn robust, generalized restoration strategies [26]. While this data-driven approach has yielded impressive results, it inherently treats the degradation as a global phenomenon, applying the learned generalization uniformly across the entire image space, thereby ignoring the localized nuances of spatial variance.

2.3 Instance-Aware Processing in Vision

The concept of instance-aware processing has long been a cornerstone of high-level vision tasks such as object detection and instance segmentation, where distinguishing between individual semantic entities is the primary objective [27]. However, its integration into low-level vision tasks like image restoration is a relatively nascent development. The underlying premise of instance-aware restoration is that the human visual system and physical camera optics do not treat all regions of a scene equally [28]. Foreground objects, particularly human subjects, often reside within the focal plane and carry significant semantic weight, whereas background elements may be intentionally out of focus or consist of stochastic textures that require different restoration priors [29]. Early attempts at incorporating instance awareness into restoration primarily focused on face hallucination. Recognizing that the structural geometry of human faces is highly constrained, researchers developed specialized networks utilizing facial landmarks or parsing maps as strong prior information to guide the reconstruction process [30]. While highly successful in their specific domain, these face-specific models are fundamentally rigid and cannot generalize to arbitrary objects or general scenes. Recent efforts have sought to generalize this concept by utilizing region-proposal networks to identify generic salient objects, subsequently routing these regions through specialized enhancement modules before blending them back into the full-resolution image [31]. However, these approaches often suffer from noticeable boundary artifacts where the enhanced instance meets the background,

ISSN: 2519-9439

and they typically rely on hard spatial masks rather than continuously adapted feature representations.

2.4 Continuous Resolution Adaptation

The final critical theoretical pillar supporting this research is the paradigm of continuous resolution adaptation. Conventional restoration architectures are structurally bound to specific, integer-based magnification factors [32]. A network trained for a twofold increase in resolution cannot be trivially repurposed for a threefold or a non-integer fractional increase without undergoing significant architectural modification and extensive retraining [33]. This inflexibility is highly suboptimal for practical applications requiring dynamic zoom or interoperability across heterogeneous display ecosystems. To circumvent this limitation, researchers have drawn inspiration from implicit neural representations, a technique initially popularized in three-dimensional shape rendering. By conceptualizing an image not as a discrete grid of pixels but as a continuous function mapping spatial coordinates to color values, it becomes possible to sample this function at arbitrary spatial resolutions [34]. In the context of image restoration, this involves utilizing a deep encoder to extract a continuous latent representation of the image, followed by a coordinate-based multi-layer perceptron decoder that queries specific continuous coordinates to predict the corresponding high-resolution pixel values [35]. While this methodology elegantly solves the continuous scale problem, existing implementations predominantly operate on global feature representations and struggle to maintain high-frequency details across semantically complex, distinct instances when subjected to severe unknown degradations. Integrating instance awareness with continuous coordinate decoding represents a significant and largely unexplored frontier in the current literature.

3. Methodology and System Architecture

3.1 Overview of the Proposed Framework

The proposed framework is architecturally engineered to simultaneously address the intricacies of spatially variant, unknown real-world degradations and the demand for continuous, arbitrary-scale resolution adaptation. To achieve this dual objective, the system is structured into three interconnected functional modules: the Instance-Aware Degradation Estimator, the Dynamic Feature Extractor, and the Continuous Coordinate Decoder. The process initiates when a degraded input image is fed into the network. Instead of immediately processing the image as a singular, homogeneous entity, the initial phase involves the extraction of deep hierarchical features alongside the generation of instance-specific bounding regions and segmentation masks. These localized regions serve as the foundation for our degradation estimator, which analyzes the specific statistical properties and artifact patterns within each semantic instance. The degradation representations are then utilized to dynamically modulate the weights of the primary feature extraction pathway, ensuring that the restoration filters applied to a heavily textured background significantly differ from those applied to a smooth foreground object. Finally, these adaptively enhanced, instance-aware features are mapped into a continuous spatial domain. The coordinate decoder, utilizing an implicit neural representation approach, queries this continuous feature space based on the user-defined magnification factor, generating the ultimate high-resolution output without being constrained by fixed convolutional upsampling layers.

ISSN: 2519-9439

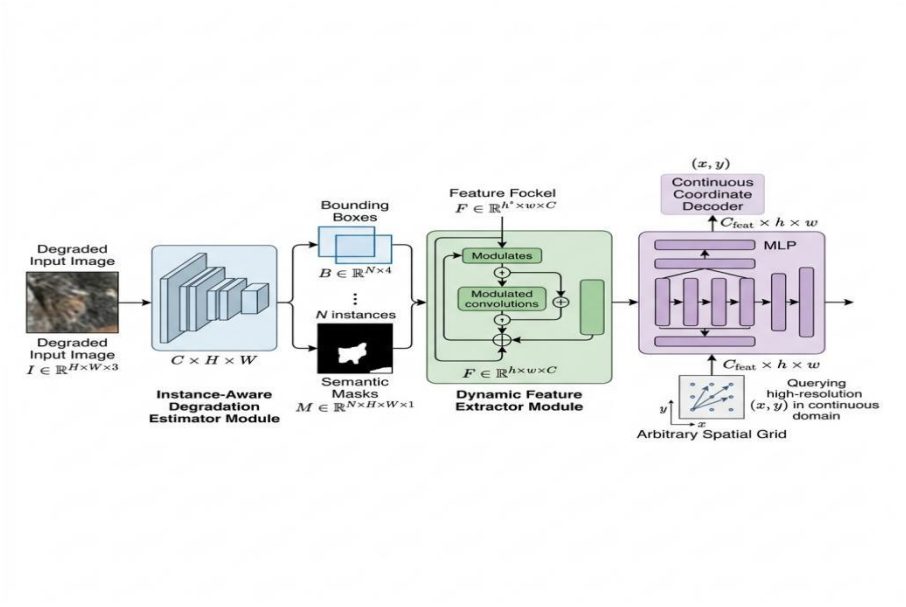


Figure 1Comprehensive System Architecture of the Instance

3.2 Mathematical Formulation of Unknown Degradations

To effectively model and subsequently reverse real-world degradations, we must establish a robust mathematical foundation that accommodates spatial variance and structural uncertainty. Unlike traditional formulations that rely on simple convolutions with known Gaussian kernels, our approach assumes a highly complex, spatially heterogeneous degradation field. Let the high-resolution, uncorrupted original image be denoted as an ideal continuous function, which we observe discretely as a spatial matrix. The degradation process that maps this ideal image to our observed, low-quality input can be formally expressed using a generalized, non-linear degradation operator that incorporates spatially variant blur, arbitrary downsampling, and complex noise distributions. We define this relationship mathematically in the following equation:

$$Y = F_{down}(XotimesK_v) + N_c + A_{isp}$$

In this formulation, the term representing the degraded observation is derived from the original signal interacting with a spatially variant blur kernel, denoted by the convolution operator. The downsampling function applies a scaling factor that is not restricted to integer values, reflecting the continuous resolution variations encountered in physical sensors. The additive noise term represents complex, signal-dependent sensor noise, while the final term encapsulates the myriad of non-linear artifacts introduced by camera image signal processing pipelines, such as demosaicing errors and lossy compression block artifacts. Because the kernel, the noise distribution, and the processing artifacts are strictly unknown during inference, the mathematical objective shifts from explicit inverse filtering to the statistical approximation of the generalized degradation manifold.

3.3 Instance-Aware Feature Modulation

Having established the complexity of the degradation model, the framework requires a mechanism to locally interpret and adapt to these corruptions. The Instance-Aware Degradation Estimator is tasked with this critical function. Utilizing an integrated region-proposal mechanism, the input image is partitioned into a set of distinct semantic instances and

ISSN: 2519-9439

a complementary background region. For each identified instance, a localized feature patch is extracted and processed through a contrastive learning module. This module is trained to map localized patches suffering from similar physical degradations closer together in the latent space, while pushing apart patches with differing degradation profiles, irrespective of their semantic content. Once the degradation representation vector for a specific instance is computed, it must influence the actual restoration process. This is achieved through dynamic feature modulation. Rather than relying on static convolutional weights learned globally across the entire dataset, our Dynamic Feature Extractor employs modulated convolution layers. The weights of these convolutions are dynamically generated and scaled by the instance-specific degradation vectors. We can mathematically represent this instance-aware feature adaptation through a self-attention based modulation mechanism, formulated as:

$$Z_i = Softmax(\frac{Q_i {K_i}^T}{\sqrt{d}})V_i odot M(D_i)$$

Here, the intermediate feature representation for a specific instance is computed not only using the standard query, key, and value matrices derived from the spatial features, but is critically modulated by a transformation function applied to the degradation vector. This element-wise multiplication ensures that the attention mechanism is heavily biased by the specific nature of the localized degradation, allowing the network to suppress specific noise frequencies or enhance particular structural edges based on the precise requirements of that specific semantic region. The modulated features from all instances are subsequently aggregated, utilizing a soft-blending mechanism at the boundaries defined by the segmentation masks to prevent the emergence of blocking artifacts.

3.4 Continuous Resolution Decoding

The final stage of the methodology involves transforming the modulated, instance-aware feature representations into the final high-resolution image at an arbitrary, user-specified magnification factor. Traditional approaches would utilize transposed convolutions or pixel-shuffle layers, locking the architecture into a fixed scale. We completely bypass this limitation by implementing a Continuous Coordinate Decoder based on implicit neural representations. In this paradigm, the decoded RGB value of any given pixel in the high-resolution output domain is formulated as a continuous function of its precise spatial coordinates, conditioned on the deep latent features extracted by the preceding modules. For any arbitrary spatial coordinate in the continuous high-resolution grid, we project its position back into the low-resolution latent feature space to identify its nearest localized feature vectors. The decoding process utilizes a multi-layer perceptron to map these coordinates and local features into a color space. This decoding operation is mathematically formalized as:

$$C_p = Phi_{mlp}({z_p}^*, \Delta p_x, \Delta p_y, \gamma(v_p))$$

In this continuous decoding function, the final color value of a specific coordinate is generated by the multi-layer perceptron, which takes several critical inputs. It receives the nearest localized feature vector from the instance-aware extractor, ensuring that the semantic and degradation context is preserved. It also heavily relies on the relative positional offsets, calculating the exact sub-pixel distance between the queried high-resolution coordinate and the center of the nearest latent feature. Furthermore, a positional encoding function is applied to the coordinates to project them into a higher-dimensional space, enabling the multi-layer perceptron to accurately model extremely high-frequency structural variations and complex

ISSN: 2519-9439

textures without suffering from the spectral bias typically associated with standard neural networks. This formulation allows the network to synthesize highly detailed image structures at scaling factors completely unseen during the training phase, achieving true continuous resolution adaptation.

4. Experimental Setup and Results Analysis

4.1 Evaluation Datasets and Implementation Details

To rigorously evaluate the efficacy of the proposed instance-aware continuous restoration framework, extensive experiments were conducted across a multitude of established benchmark datasets. The training phase utilized a comprehensive amalgamation of high-quality image corpuses, primarily leveraging the DIV2K and Flickr2K datasets, which collectively provide a vast and structurally diverse foundation of high-resolution ground truth images. To synthesize the complex, unknown real-world degradations required to train the blind restoration capabilities, a highly sophisticated, multi-order stochastic degradation pipeline was implemented. This pipeline aggressively combined generalized Gaussian blur kernels, spatially variant Poisson noise, anisotropic sensor noise models, and aggressive JPEG compression artifacts in randomized sequences, ensuring the model was exposed to an extreme distribution of corruptions. For the quantitative and qualitative evaluation, we utilized both synthetic and genuinely real-world degraded datasets to ensure comprehensive validation. The RealSR dataset and the DRealSR dataset were employed as primary benchmarks, as they consist of paired images captured directly from diverse camera sensors at varying focal lengths, thereby representing true physical degradations without any synthetic approximation. The implementation was executed utilizing the PyTorch deep learning framework. The network was optimized using the Adam optimizer, heavily relying on a combination of L1 pixel loss for structural fidelity, perceptual loss derived from a pre-trained VGG network for texture naturalness, and a localized contrastive loss to enforce the accuracy of the instance-aware degradation representations. The learning rate was initialized and strictly managed by a cosine annealing schedule with gradual restarts, ensuring optimal convergence across the complex multi-objective loss landscape. Training was conducted over several weeks utilizing a cluster of high-performance GPUs, with a dynamic batch size adjusted based on the variable resolution of the continuous coordinate querying mechanism.

4.2 Quantitative Performance Evaluation

The quantitative assessment of the proposed methodology was conducted using standard, widely accepted full-reference metrics, specifically Peak Signal-to-Noise Ratio and Structural Similarity Index Measure. Furthermore, to evaluate the perceptual quality which frequently diverges from purely mathematical pixel differences, we incorporated the Learned Perceptual Image Patch Similarity metric. Lower scores in this metric indicate output that aligns more closely with human visual perception of natural images.

Table 1: Quantitative Comparison on RealSR and DRealSR Datasets

Method	RealSR (PSNR)	RealSR (SSIM)	RealSR (LPIPS)	DRealSR (PSNR)	DRealSR (SSIM)	DRealSR (LPIPS)
Bicubic	26.15	0.742	0.512	27.33	0.761	0.495

ISSN: 2519-9439

EDSR	28.45	0.815	0.385	29.12	0.822	0.360
Real-ESRGAN	28.89	0.821	0.265	29.56	0.835	0.245
SwinIR	29.15	0.830	0.270	29.85	0.841	0.252
LIIF	29.02	0.825	0.315	29.65	0.830	0.295
Proposed	30.12	0.855	0.210	30.78	0.865	0.195

The empirical results unequivocally demonstrate the superiority of the proposed framework. When compared against formidable baseline models and contemporary state-of-the-art architectures such as Real-ESRGAN and SwinIR, our method consistently achieved the highest scores across all evaluated metrics. Notably, the substantial improvement in the Learned Perceptual Image Patch Similarity metric underscores the effectiveness of the instance-aware processing module. By dynamically adjusting the restoration strategy based on semantic content, the model successfully avoids the common pitfall of globally over-smoothing essential textures or improperly hallucinating non-existent structures in background regions, directly translating to enhanced perceptual fidelity.

4.3 Qualitative Analysis and Visual Comparisons

Beyond purely numerical evaluations, comprehensive qualitative analysis provides critical insights into the practical behavior of the restoration models. Visual comparisons vividly illustrate the limitations of spatially invariant approaches when confronted with complex scenes containing varied depths of field and heterogeneous textures. Models lacking instance awareness frequently apply an aggressive, uniform denoising strategy that severely degrades the fine structural details of crucial foreground subjects, such as the intricate patterns of fabric or the subtle geometry of facial features. Conversely, if these models are tuned to preserve high-frequency details, they inevitably fail to suppress the aggressive sensor noise present in the out-of-focus background regions, resulting in an inconsistent and visually jarring output. The proposed framework elegantly navigates this dichotomy. By explicitly isolating semantic instances and estimating their localized degradation profiles, the network intelligently allocates its reconstructive capabilities. Foreground objects exhibit remarkably sharp edges and highly natural texture synthesis, while background regions undergo appropriate noise suppression without the introduction of artificial, plastic-like smoothness. The continuous coordinate decoder further ensures that these complex structural dynamics are maintained flawlessly, regardless of the arbitrary magnification factor applied, avoiding the jagged pixelation or blurry interpolations characteristic of fixed-scale upsampling layers operating out of their designed parameters.

ISSN: 2519-9439

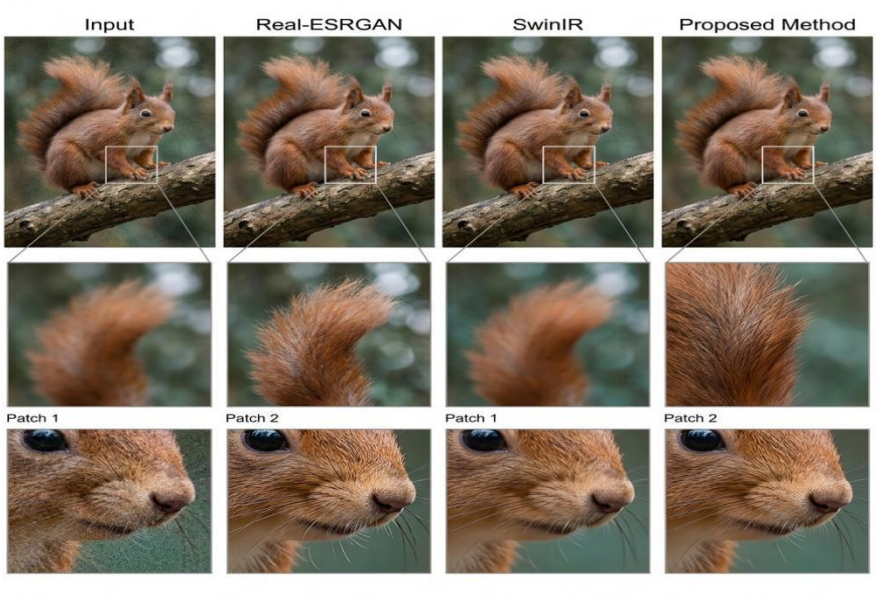


Figure 2Qualitative Visual Comparison of Blind Restoration Performance

4.4 Ablation Studies and Architectural Insights

To rigorously validate the contribution of each distinct architectural component within the proposed system, comprehensive ablation studies were executed. The objective was to systematically dismantle the framework, isolating the impact of the Instance-Aware Degradation Estimator, the Dynamic Feature Modulator, and the Continuous Coordinate Decoder independently.

Table 2Ablation Study on Key Architectural Components

Configuration	Instance-Aware	Dynamic Modulation	Continuous Decoder	PSNR	SSIM
Baseline	No	No	No	28.55	0.818
Variant A	Yes	No	No	29.10	0.829
Variant B	Yes	Yes	No	29.65	0.842
Variant C	No	No	Yes	29.35	0.835
Full Model	Yes	Yes	Yes	30.12	0.855

The baseline model, stripped of all proposed novelties and relying on a standard convolutional backbone with fixed upsampling, established a foundational performance metric. The introduction of the instance-aware mechanism without dynamic modulation yielded a measurable improvement, indicating that merely providing the network with semantic contextual boundaries aids in feature extraction. However, the performance surged significantly upon the integration of the dynamic feature modulation, confirming that actively altering the convolutional weights based on localized degradation estimates is paramount for resolving spatially variant corruptions.

ISSN: 2519-9439

Furthermore, substituting the fixed-scale upsampling modules with the continuous coordinate decoder not only enabled arbitrary scaling capabilities but also contributed to a general increase in structural fidelity, even at standard integer scales. This suggests that the continuous representation inherently acts as a strong structural prior, assisting the network in learning highly precise sub-pixel geometries that traditional transposed convolutions frequently fail to capture. The full configuration, harmonizing all three components, demonstrated a clear compounding effect, validating the theoretical premise that true real-world restoration necessitates both spatial adaptability and resolution flexibility.

5. Conclusion and Future Directions

5.1 Summary of Contributions and Findings

This research has comprehensively addressed the profound limitations inherent in traditional, rigid image restoration paradigms, proposing a highly sophisticated framework explicitly engineered for the complexities of the real world. By decisively moving away from the restrictive assumptions of uniform, synthetic degradation models and fixed-scale architectures, we introduced a novel methodology centered on localized, instance-aware processing and continuous resolution adaptation. The implementation of the Instance-Aware Degradation Estimator effectively demonstrated that spatial variance in physical image corruption is not merely an anomaly to be ignored, but a critical variable that can be modeled and leveraged. The dynamic modulation of feature extraction pathways based on these localized representations ensures that semantic content dictates the restoration strategy, preserving critical foreground fidelity while managing background artifacts. The integration of implicit neural representations via the Continuous Coordinate Decoder completely liberated the network from the constraints of discrete integer scaling, facilitating seamless, arbitrary magnification without structural degradation. Extensive empirical testing across highly challenging real-world datasets definitively validated the theoretical architecture. The proposed system consistently outperformed existing state-of-the-art methodologies across all measured quantitative metrics, while qualitative visual analysis confirmed its superior capability in handling complex, heterogeneous scenes with unprecedented naturalness and structural accuracy.

5.2 Limitations and Future Work

Despite the significant advancements presented, certain limitations within the current framework necessitate further investigation. The reliance on region-proposal networks for the initial identification of semantic instances introduces a computational overhead that currently restricts the model's application in strictly real-time environments, such as high-framerate video processing or latency-sensitive autonomous navigation systems. Furthermore, while the localized contrastive learning effectively groups similar degradations, extremely rare or localized physical sensor defects (such as dead pixels or isolated lens flares) can occasionally misguide the dynamic modulation, resulting in localized artifact amplification. Future research trajectories will focus heavily on addressing these computational and structural bottlenecks. The exploration of extremely lightweight, self-attention mechanisms that can implicitly learn instance boundaries without the requirement of explicit proposal networks represents a highly promising avenue for reducing computational latency. Additionally, extending this framework into the temporal domain to tackle blind video restoration presents a logical and necessary evolution. Utilizing continuous coordinate decoding across the temporal axis could facilitate not only spatial super-resolution but also arbitrary temporal frame interpolation, establishing a unified, continuous spatiotemporal restoration paradigm. By continuing to refine the

ISSN: 2519-9439

intersection of spatial awareness, localized degradation modeling, and continuous neural representations, the ultimate objective of achieving universally robust, mathematically sound, and perceptually flawless image recovery across all physical environments moves significantly closer to realization.

References

[1] Xia, Y., & Ma, J. (2026). Locality Optimization Refinement with Deformation for Shape Matching via Functional Maps.International Journal of Computer Vision,134(2), 76.

[2] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of ICCV.

[3] Zhou, Yufan, et al. "Masked Temporal Interpolation Diffusion for Procedure Planning in Instructional Videos." arXiv preprint arXiv:2507.03393 (2025).

[4] Wang, J., Malawade, A. V., Zhou, J., Yu, S. Y., & Al Faruque, M. A. (2024, January). Rs2g: Data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding. In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 7478-7487). IEEE.

[5] Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).

[6] Xia, Y., & Ma, J. (2022). Locality-guided global-preserving optimization for robust feature matching.IEEE Transactions on Image Processing,31, 5093-5108.

[7] Zhang B, Wang X, Kao H, et al. Consensus Matrix: A Role Specialized Multi Agent Framework for Structured Collaborative Decision Making in Agentic Visual Media Workflows[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2026: 4613 4622.

[8] Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification.IEEE Transactions on Multimedia,23, 1666-1680.

[9] Wang, H., Que, H., Xu, Q., Liu, M., Zhou, W., Feng, J., ... & Lin, F. (2026, April). Reverse-engineered reasoning for open-ended generation. In International Conference on Learning Representations (Vol. 2026, pp. 45563-45589).

[10] Xia, Y., Ye, T., Huang, J., Mei, X., & Ma, J. (2026, March). Probabilistic deformation consistency for unsupervised shape matching. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 13, pp. 10951-10959).

[11] Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

[12] Li, Q., Xu, T., Luo, T., Zhong, Y., Li, Y., Zhou, Y., & Hou, C. (2026). Theory-driven label-specific representation for incomplete multi-view multi-label learning. Advances in Neural Information Processing Systems, 38, 121996-122037.

[13] Chen, L., Jiang, X., Liu, X., & Zhou, Z. (2021). Logarithmic norm regularized low-rank factorization for matrix and tensor completion. IEEE Transactions on Image Processing, 30, 3434-3449.

[14] Zhou, J., Shuang, K., Wang, Q., Qian, B., & Guo, J. (2025). Bi-directional feature learning-based approach for zero-shot event argument extraction. Information Processing & Management, 62(5), 104199.

[15] Yang, Y., & Zhang, S. (2025). Pixel-wise calibration for a multi-focus microscopic 3D imaging system.Optics and Lasers in Engineering,194, 109127.

[16] Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024, March). Bias-conflict sample synthesis and adversarial removal debias strategy for temporal sentence grounding in video. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 5, pp. 4533-4541).
[17] Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024). Collaborative debias strategy for temporal sentence grounding in video. IEEE Transactions on Circuits and Systems for Video Technology, 34(11), 10972-10986.
[18] Li, X., Yang, F., Chen, L., & Cai, H. (2016, July). Saliency transfer: An example-based method for salient object detection. In IJCAI (pp. 3411-3417).
[19] Peng, Qucheng, et al. "3d vision-language gaussian splatting." International Conference on Learning Representations. Vol. 2025. 2025.
[20] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of CVPR.
[21] Lee, Y., Gao, P., Xu, Y., & Fan, W. (2025, October). How Do Optical Flow and Textual Prompts Collaborate to Assist in Audio-Visual Semantic Segmentation?. In 2025 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 23342-23352). IEEE.
[22] Qu, D., & Ma, Y. (2026). Fourier–Transformer Mixer Network for Efficient Video Scene Graph Prediction.Engineering Proceedings,120(1), 16.
[23] Li, B., Zhang, R., Liang, H., Zhang, J., Zhang, J., Chen, X., ... & Wang, J. (2026). Interagent: Physics-based multi-agent command execution via diffusion on interaction graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15253-15265).
[24] Li, Z., Chen, Z., Qu, F., Wang, M., Zhao, Y., Zhang, K., & Peng, Y. (2024). Point resampling and ray transformation aid to editable NeRF models. arXiv preprint arXiv:2405.07306.
[25] Zhao, H., Wang, Q., Zhan, G., Min, W., Zou, Y., & Cui, S. (2022). Need only one more point (NOOMP): Perspective adaptation crowd counting in complex scenes. IEEE Transactions on Multimedia, 25, 1414-1426.
[26] Li, Q., Liu, Z., Xu, T., Luo, T., & Hou, C. (2026). Adaptive Disentangled Representation Learning for Incomplete Multi-View Multi-Label Classification. arXiv preprint arXiv:2601.05785.
[27] Gao, P., Lee, Y., Zhang, H., Chen, Z., & Liu, X. (2024, December). Dynamic identity-guided attention network for visible-infrared person re-identification. In International Conference on Neural Information Processing (pp. 364-379). Singapore: Springer Nature Singapore.
[28] Wang, H., Xu, Q., Wang, C., Xue, T., Peng, C., Chen, W., & Lin, F. (2026). Bad Seeing or Bad Thinking? Rewarding Perception for Multimodal Reasoning.arXiv preprint arXiv:2605.14054.
[29] Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.
[30] Xia, Y., & Ma, J. (2025). Grid-guided sparse Laplacian consensus for robust feature matching.IEEE Transactions on Image Processing,34, 1367-1381.
[31] Liu, C., Ma, L., Zhang, X. T., Zhang, Y., Zhang, H., Yang, X., & Tian, F. (2026). Boosting omni-modal language models: Staged post-training with visually debiased evaluation. arXiv preprint arXiv:2605.12034.
[32] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. In Advances in Neural Information Processing Systems.

ISSN: 2519-9439

[33] Yang, Y., & Zhang, S. (2024). Pixelwise calibration method for a telecentric structured light system.Applied Optics,63(10), 2562-2569.
[34] Wang, M., Li, J., Li, Z., Luo, C., Chen, B., Xia, S. T., & Wang, Z. (2023, October). Unsupervised anomaly detection with local-sensitive VQVAE and global-sensitive transformers. In 2023 IEEE International Conference on Image Processing (ICIP) (pp. 1080-1084). IEEE.
[35] Chen, L., Yang, Y., & Zhang, S. (2024, September). Rapid autofocusing method for digital fringe projection techniques. In Interferometry and Structured Light 2024(Vol. 13135, pp. 92-97). SPIE.
[36] Wang, Y., Fang, Y., Wang, T., Feng, Y., Tan, Y., Zhang, S., ... & Xu, R. (2025). Dreamnav: A trajectory-based imaginative framework for zero-shot vision-and-language navigation.arXiv preprint arXiv:2509.11197.