International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Self-Supervised Multi-Scale Deep Learning Framework for Unpaired Image Super-Resolution

Prashant Kumar Tamrakar[1]*, Dr. Virendra Kumar Swarnkar[2]

Submitted: 03/11/2024 **Revised:** 08/12/2024 **Accepted:** 18/12/2024

Abstract—Image Super-Resolution (SR) aims to reconstruct high-resolution (HR) images from their low-resolution (LR) counterparts, a task traditionally reliant on large-scale paired datasets. However, acquiring perfectly aligned LR-HR image pairs is often impractical, particularly in real-world and domain- specific applications. This paper proposes a novel self-supervised multi-scale deep learning framework that addresses the SR challenge using unpaired data. The architecture incorporates a multi-scale feature extraction module that effectively captures hierarchical contextual information at different resolutions, en- hancing the model's capacity to reconstruct fine image details. To eliminate the dependency on paired data, a self-supervised learning mechanism is introduced, leveraging pseudo-pair gen- eration, cycle-consistency constraints, and perceptual similarity losses. We evaluate our approach on benchmark datasets such as DIV2K and Flickr2K using only HR images for training. Experi- mental results demonstrate that our method achieves competitive performance in terms of PSNR and SSIM while significantly outperforming conventional models in texture preservation and visual fidelity. The framework proves effective and generalizable, particularly for real-world deployment where paired data is limited or unavailable.

Index Terms*—High-Fidelity Image Recovery, Hierarchical Feature Extraction, Autonomous Representation Learning, Visual Enhancement, Perceptual Optimization, Unlabeled Training Data, Structural Restoration, Cycle Consistency, Unsupervised Image Reconstruction,Visual Quality Metrics*

I. INTRODUCTION

A. Background and Significance

The reconstruction of high-quality visuals from degraded, compressed, or low-resolution inputs is a longstanding chal- lenge in computer vision and image processing. This problem has garnered significant attention due to its wide range of real-world applications, including video surveillance, medical diagnostics, remote sensing, and media restoration [1]–[3]. Enhancing visual clarity from coarse or blurry images can aid not only human perception but also improve downstream tasks such as object recognition and scene understanding.

Historically, early enhancement techniques employed in- terpolation methods such as bicubic or Lanczos resampling, which are computationally simple but tend to produce overly smooth and visually unappealing outputs. These classical techniques lack the ability to hallucinate fine texture details that are not explicitly present in the input. With the rise of data-driven methods and advancements in neural network design, significant breakthroughs have been made, particularly in learning the mapping between low- and high-resolution image domains [4]–[6]. Such methods exploit large corpora of image pairs to infer missing high-frequency information.

Nevertheless, the success of these approaches hinges on the availability of perfectly aligned pairs of low-resolution and high-resolution images—a requirement often unmet in real- world scenarios. In practical settings, collecting such paired datasets is expensive, time-consuming, or even infeasible due to uncontrolled environmental conditions, motion artifacts, and device heterogeneity [7].

B. Challenges in the Absence of Paired Data

The reliance on supervised training strategies creates a significant bottleneck for scalability. When aligned input-target pairs are missing, models struggle to learn an effective re-

[1*]*Department of Computer Science, Bharti Vishwavidyalaya, Durg, Chhattisgarh, India prashant.tamrakar35@gmail.com*
[2]*Department of Computer Science & Engineering, Bharti Vishwavidyalaya, Durg, Chhattisgarh, India swarnkarvirendra@gmail.com*

construction function. Attempts to synthetically degrade high- resolution images using predefined kernels or downscaling functions often fail to capture the complexity and variability of natural degradation [8]. As a result, networks trained on synthetic data tend to overfit to simplified blur models and perform poorly on real-world data, which exhibit complex noise patterns, unknown camera responses, and compression artifacts.

Furthermore, collecting real low-resolution data and manu- ally pairing it with corresponding high-quality images is labor- intensive and may not guarantee perfect spatial alignment. These limitations have fueled interest in unsupervised and self- supervised strategies that can learn enhancement mappings without relying on strict pairings.

C. Recent Advances: Representation Learning and Feature Scaling

To overcome the challenges of unpaired supervision, mod- ern techniques have explored self-guided learning mecha- nisms. These strategies aim to guide the network using internal consistency, transformation invariance, and feature reconstruc- tion objectives [9]. One popular approach is cycle-consistency, where an input is passed through forward and reverse transfor- mations to ensure that the reconstructed image remains faithful to the original even in the absence of direct supervision [10]. Perceptual loss functions that compare features in a pre-trained classification network (such as VGG) are also effective in improving visual fidelity [11].

Simultaneously, progress in architectural design has high- lighted the importance of analyzing visuals at multiple levels of abstraction. Networks that extract features at different spatial scales can better model both global structures and fine-grained textures [12]. Hierarchical processing modules allow the model to refine details iteratively while preserving contextual coherence across image regions. Such mechanisms are essential for recovering realistic textures and preventing artifacts in enhanced outputs.

By integrating these concepts—unsupervised training and hierarchical feature extraction—a model can learn robust map- pings from degraded to high-quality imagery, even when only unpaired datasets are available.

D. Research Motivation and Contributions

This paper presents a unified framework designed to en- hance visual fidelity using only collections of unpaired high- quality images. The motivation stems from the need to build practical, adaptable systems that do not depend on heavily curated datasets for training. Our method leverages feature aggregation across scales and enforces internal supervision through pseudo-pair generation and feedback loops.

The main contributions of this study can be summarized as follows:

- **A** hierarchical feature refinement network that processes visual information at multiple spatial resolutions to re- construct both global structures and local textures.
- **A**n autonomous learning strategy that incorporates self- supervision via perceptual alignment, forward-backward consistency, and adversarial fine-tuning without relying on paired ground truths.
- **E**xtensive experimentation on benchmark datasets (e.g., DIV2K, BSD500, and Flickr2K) demonstrating that the proposed framework performs competitively with or out- performs existing supervised and unsupervised methods.
- **A**blation studies and visual analyses confirming the indi- vidual and combined effectiveness of the feature extrac- tion modules and self-supervised objectives.

These models brought significant improvements in metrics like PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index), two commonly used metrics for SR evalu- ation.

However, most of these models initially processed the image at a single scale, assuming a uniform structure throughout the input image. As natural images often contain multi-scale information (e.g., large objects alongside fine textures), single- scale approaches were inherently limited in capturing the full richness of real-world images.

II. RELATED WORK

A. Traditional Learning with Supervised Pairs

The reconstruction of high-fidelity images from lower- quality inputs initially advanced through direct learning of mapping functions from paired

datasets. Early efforts like the SRCNN [13] laid the groundwork by introducing con- volutional neural networks trained to restore high-resolution images from low-resolution counterparts. These models opti- mized pixel-wise loss functions, such as mean squared error (MSE), which prioritized structural accuracy but often failed to produce visually pleasing textures.

Subsequent developments incorporated deeper networks and residual connections, as seen in VDSR [14] and EDSR [15], improving the learning capacity and convergence rate. Further- more, generative models like SRGAN [16] introduced adver- sarial learning to enhance visual realism, using perceptual and adversarial losses to align output distributions more closely with natural image statistics. Despite their effectiveness, these supervised models are limited by their dependence on meticu- lously aligned training pairs, which are rarely available outside curated benchmarks.\

B. Self-Supervised and Unsupervised Techniques

To overcome the need for explicit correspondence between inputs and targets, unsupervised paradigms have gained mo- mentum. These approaches aim to learn meaningful mappings without paired supervision by introducing internal learning mechanisms.

CycleGAN [16], though originally developed for domain translation, was adapted for enhancement tasks by enforcing cycle-consistency between forward and backward mappings. This strategy ensures the reversibility of transformations, ef- fectively enabling learning from unpaired samples. Similarly, Zero-Shot SR (ZSSR) [15] proposed using a single test image as its own training data by exploiting self-similarity within the image. These methods reduce dependency on external datasets but often compromise on generalization, especially in complex, high-frequency scenarios.

More recent self-supervised frameworks incorporate proxy tasks, data augmentations, and contrastive losses to extract robust features without requiring manual labeling [14], [13]. These techniques encourage models to learn transformation- invariant representations, which are later used for reconstruc- tion or refinement.

C. Feature Hierarchies Across Spatial Scales

Visual data inherently contains information at varying spa- tial resolutions, from coarse contours to fine texture details. Effective image enhancement systems benefit from architec- tures that can process and fuse information across these levels. Hierarchical or pyramidal networks have become a staple in high-fidelity image synthesis. For example, MSRN [16] introduced multi-scale residual blocks that capture context through dilated convolutions and recursive pathways. RCAN [12] incorporated channel attention mechanisms to selectively enhance feature maps across different levels. These mecha- nisms empower the model to focus on informative regions while suppressing noise.

Despite their strengths, many such architectures are primar- ily trained in supervised settings. Their performance tends to degrade when exposed to data distributions that differ from the training domain or when paired examples are unavailable.

D. Gaps and Opportunities in Existing Studies

While prior work has demonstrated the value of deep architectures and learning-based losses, significant challenges remain. Most supervised models are not robust to unseen degradations due to overfitting on synthetic blur models. Unsu- pervised and self-supervised systems, while promising, often struggle to balance structural accuracy and visual realism.

Moreover, existing approaches seldom integrate self- supervised objectives with hierarchical architectures in a tightly coupled manner. This creates an opportunity for novel To prevent the loss of spatial precision, downsampling is limited to only the initial stages. Deeper layers focus on semantic expansion without spatial degradation.

C. Representation Fusion

After hierarchical extraction, the encoder outputs multi- scale features $\{F_1, F_2, \ldots, F_n\}$. These features are fused using a cross-resolution aggregator, which aligns and merges them into a unified latent representation F_{agg}:$\sum$

designs that simultaneously leverage cross- resolution feature processing and internal consistency constraints to boost gen-Fagg

$=\sum_{i=1} \alpha_i \cdot \mathrm{Up}(F_i)$ eralization and performance.

III. PROPOSED METHODOLOGY

The proposed framework introduces a fully automated learn- ing paradigm designed to enhance degraded visual data with- out requiring precisely aligned training pairs. Our architecture integrates hierarchical feature modeling, cross-scale represen- tation fusion, and a training mechanism that leverages cyclic supervision and structural similarity preservation. This section elaborates on each architectural component, the mathematical design, training process, and optimization objectives.

A. Overview of the Framework

Here, α_i are learnable weights indicating the importance of each scale, and $\mathrm{Up}(\cdot)$ is a spatial upsampling operator. This strategy allows the model to dynamically learn the relevance of context from different resolutions.

D. Decoder: Reconstruction with Refinement

The decoder reconstructs the enhanced output by upsam- pling F_{agg} progressively, using a series of transposed convo- lutions and refinement blocks: $\hat{I}_H = D(F_{\text{agg}}) = \psi_k(\ldots \psi_2(\psi_1(F_{\text{agg}}))\ldots)$

Where ψ_j is a decoder block composed of:

Let $I_L \in R^{H\times W\times 3}$ be an unpaired low-resolution input.ˆ

- Transposed convolution (to upsample spatially)
- Skip connections from encoder (F_{n-j+1})

Our model aims to reconstruct an enhanced version $I_H \in R^{rH\times rW\times 3}$, where r is the enhancement factor (e.g., 2× or 4×). The architecture follows an encoder-decoder pattern with specialized modules for multilevel abstraction and reconstruc-

tion.

The overall design contains the following modules:

- **Multi-Level Encoder** – Captures coarse-to-fine semantic and spatial cues.
- **Representation Fusion Block** – Combines multiscale contextual maps.
- **Reconstruction Decoder** – Rebuilds the enhanced output with refinement.
- **Self-Guided Training Mechanism** – Enables supervi- sion without ground truth via internal consistency, cycle restoration, and perceptual constraints.

B. Encoder: Multi-Level Feature Modeling

The encoder processes the input image through a series of layers that reduce spatial resolution while increasing semantic abstraction. Each stage consists of convolutional layers with residual connections to preserve information flow.

Let $\phi_i(\cdot)$ denote the feature extractor at stage i, and F_i be the output features: $F_1 = \phi_1(I_L), \quad F_2 = \phi_2(F_1), \quad \ldots, \quad F_n = \phi_n(F_{n-1})$

Each ϕ_i is composed of a convolution layer, ReLU activa- tion, and a residual connection: Channel attention module to emphasize discriminative features This refinement ensures better texture recovery and suppresses unnatural artifacts often found in hallucinated regions.

E. Self-Supervised Learning without Paired Labels

Since aligned high-resolution targets are not available, we rely on a cycle-based learning paradigm. The idea is to reconstruct the enhanced image and then synthetically degrade it to generate a cycle loop. This allows the model to compare reconstructed and re-degraded images and enforce consistency.

Forward Enhancement:

$$\hat{I}_H = G(I_L)$$

Synthetic Degradation:

$$\hat{I}_L = D_{\text{sim}}(\hat{I}_H)$$

Cycle Restoration:

$$\hat{I}_H^{\text{cyc}} = G(\hat{I}_L)$$

We aim to ensure that $\hat{I}_H^{\text{cyc}} \approx \hat{I}_H$, encouraging robustness and cycle consistency.

F. Loss Functions

Our optimization objective combines several loss

terms, each encouraging a different aspect of quality.

(a) Cycle Consistency Loss:

$$L_{cyc} = \|\hat{I}_H^{cyc} - \hat{I}_H\|_1$$

(b) Self-Reconstruction Loss:

$$L_{rec} = \|\hat{I}_H - I_H\|_1$$

$L_{rec} = \|\hat{I}_H - I_H\|_1$This section elaborates on the datasets employed for experi- mentation, the preparation steps followed prior to training, and

(c) ***Perceptual Distance Loss:*** **Using pre-trained VGG-19**

features ϕ:

$$L_{perc} = \sum_j \| \phi_j(\hat{I}_H) - \phi_j(I_H) \|_2^2$$

(d) Structural Similarity Loss:

$L_{ssim} = 1 - SSIM(\hat{I}_H, I_H)$

Total Objective Function:

$L_{total} = \lambda_1 L_{cyc} + \lambda_2 L_{rec} + \lambda_3 L_{perc} + \lambda_4 L_{ssim}$

Where typical values are: $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 0.05$, $\lambda_4 = 0.1$.

G. Training Strategy $_H$

The model is trained end-to-end using randomly sampled patches from the input dataset. Key configurations include:

- **Framework:** PyTorch 2.0
- **Hardware:** NVIDIA A100 GPU with 40GB VRAM
- **Batch Size:** 16
- **Input Size:** 64 × 64 patches
- **Learning Rate:** 1×10^{-4}, reduced via cosine annealing
- **Optimizer:** Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
- **Epochs:** 200
- **Data Augmentation:** Random flips, color jitter, and rotations

Degraded images are synthetically generated (e.g., Gaussian blur, downsampling) or sampled from real-world low-quality datasets. No ground truth targets are used for supervision in natural image enhancement.

H. Summary of Architectural Design

TABLE I SUMMARY OF ARCHITECTURAL COMPONENTS

Module	**Functionality**	**Key Features**
Multi-Level Encoder	Extracts spatial and semantic cues	Residual blocks, reduced downsampling
Representation Fu-sion	Aggregates context from multiple lev-els	Weighted fusion, dynamic attention
Decoder	Constructs output with high fidelity	Skip connections, channel attention, up-convolution
Self-Supervision Engine	Guides learning without ground truth	Cyclic loop, reconstruction, perceptual feedback
Loss Framework	Optimizes fidelity and realism	Multi-objective with structural, visual, and cycle losses

IV. DATASET AND EXPERIMENTAL SETUP

This section elaborates on the datasets employed for experi- mentation, the preparation steps followed prior to training, and the framework environment used to implement the model. Our approach notably abstains from using explicitly aligned high- and low-resolution pairs, enabling more realistic deployment scenarios.

A. Datasets

To ensure robustness and generalizability, we used a diverse collection of publicly available image datasets commonly referenced in restoration research:

- **DIV2K:** Comprising 800 high-definition images for train- ing and 100 for validation, this dataset offers a wide variety of textures and scenes in 2K resolution (approxi- mately 2040×1080 pixels). It serves as our primary data source.
- **Flickr2K:** Contains 2,650 natural photographs collected from Flickr, offering diverse real-world content with varying color and structure distributions.
- **BSD500:** A benchmark dataset consisting of 500 natural images (200 training, 100 validation, 200 test) from the Berkeley Segmentation Dataset, known for its balanced complexity and detail.
- **RealSRSet and Urban100 (for inference only):** Used strictly to evaluate qualitative performance on real-world degraded images.

We emphasize that only original high-quality images from these datasets were used as inputs, and no ground-truth paired low-resolution counterparts were available during training.

B. Data Preprocessing and Augmentation

As the goal was to simulate a realistic scenario where only high-quality inputs are available, we synthetically generated degraded samples during training using a stochastic downsam- pling model. The process involved:

- **Random Gaussian Blur** with kernel sizes between 3×3 and 7×7.
- **Downscaling** by factors of 2 or 4 using bicubic interpo- lation.
- **Noise Injection** with Gaussian noise (mean = 0, std = 5).
- **JPEG Compression Artifacts** via quality reduction (set between 10–30%).

For data augmentation, the following strategies were adopted:

- Horizontal and vertical flipping.
- Random 90° rotations.
- Color jitter with ±10% variation in brightness, contrast, and saturation.
- Patch extraction: random crops of size 64×64 (input) were extracted from the degraded images.

During inference, no augmentation or synthetic degradation was applied. Images were enhanced directly to verify model generalization on natural, unseen inputs.

C. Handling of Unpaired Samples

To maintain an unsupervised setup, ground-truth high- resolution images were not used as direct targets. Instead, our training follows a cyclic reconstruction strategy:

- A high-quality image is synthetically degraded on-the-fly to produce a pseudo-low-resolution sample.
- The model learns to enhance this input without being explicitly shown the original version.
- A self-loop is introduced by degrading the output again and re-enhancing it to promote consistency.

This process avoids reliance on aligned supervision while still enforcing internal structure preservation through recon- struction constraints.

D. Data Partitioning

The datasets were split into training, validation, and testing subsets to enable rigorous performance evaluation. Specifics are as follows:

- **Training Set:** Used for self-supervised model learning.
- **Validation Set:** Used to tune hyperparameters and mon- itor overfitting.
- **Test Set:** Used for both visual and quantitative perfor- mance evaluation.

For consistency, the test set consisted of images that were never seen or altered during training or validation stages.

E. Implementation Details

The proposed framework was developed using the following toolchain:

- **Programming Framework:** PyTorch 2.0.1
- **GPU Hardware:** NVIDIA A100 (40GB VRAM)

- **CUDA Version:** 11.8
- **Python Version:** 3.9
- **Image Processing Libraries:** OpenCV 4.7, PIL, Albu- mentations
- **Experiment Tracking:** TensorBoard and Weights & Bi- ases
- **Checkpointing and Logging:** Custom hooks imple- mented via PyTorch Lightning

Training was conducted over 200 epochs, with a batch size of 16 and initial learning rate set to 1×10^{-4}. Cosine annealing with warm restarts was used to modulate the learning rate. All training was reproducible via controlled random seeds and fixed augmentation pipelines.

V. Results and Discussion

This section presents the performance evaluation of the proposed enhancement framework using standard metrics and visual inspection. Both quantitative and qualitative analyses were conducted to validate the effectiveness of our design. Furthermore, an ablation study was carried out to examine the contribution of each architectural and training component.

A. Quantitative Metrics

To objectively assess fidelity and perceptual quality, we employed three widely adopted evaluation criteria:

- **Peak Signal-to-Noise Ratio (PSNR):** Measures pixel- level fidelity between reconstructed and reference images (higher is better).
- **Structural Similarity Index (SSIM):** Evaluates per- ceived structural similarity based on luminance, contrast, and structure (range 0–1, higher is better).
- **Learned Perceptual Image Patch Similarity (LPIPS)** [**?**]: Captures perceptual similarity using deep features (lower is better).

Testing was performed on unpaired images by generating synthetic degraded versions and comparing reconstructions against the original.

B. Comparative Evaluation

To validate our model's performance, we compared it against both supervised and unsupervised baseline techniques: As shown, our model achieves performance closely com- parable to advanced supervised architectures, despite having no direct access to paired training data. Against other un- supervised baselines, our framework consistently outperforms

across all three metrics.

C. Ablation Study

To quantify the influence of major architectural and training components, we performed controlled experiments by selec- tively disabling elements of the model:

The results highlight that removing self-loop supervision or omitting cross-scale fusion leads to measurable degradation in performance. Notably, texture refinement loss has a significant impact on perceptual quality, as captured by LPIPS.

D. Qualitative Results

Visual comparisons were conducted to complement the numerical evaluation. Below are representative results from the test dataset.

Observations:

- Our method successfully reconstructs fine edges and avoids blurring, even when inputs are heavily degraded.
- Supervised models tend to produce sharper outputs but may overfit to specific artifacts when trained on synthetic data only.
- Unsupervised baselines like CycleGAN exhibit halluci- nated textures or color bleeding.

E. Graphical Summary

To further clarify comparative behavior, we present the following visual plots:

These plots affirm that the proposed system achieves an excellent balance of perceptual sharpness and structural con- sistency, while maintaining competitive numerical fidelity.

TABLE II DATASET SPLIT FOR TRAINING, VALIDATION, AND TESTING

Dataset	Training Images	Validation Images	Test Images
DIV2K	800	100	-
Flickr2K	2,250	200	200
BSD500	200	100	200

TABLE III QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS

Method	PSNR ↑	SSIM ↑	LPIPS ↓
EDSR (Supervised)	28.92	0.814	0.196
RCAN (Supervised)	29.71	0.834	0.180
DFDNet (Unpaired)	26.15	0.761	0.248
SwinIR (Supervised)	30.12	0.848	0.170
CycleGAN-SR (Unpaired)	27.01	0.783	0.223
Ours (Unpaired)	**29.11**	**0.821**	**0.176**

TABLE IV ABLATION STUDY OF MODEL COMPONENTS

Model Variant	PSNR ↑	SSIM ↑	LPIPS ↓
Full model (baseline)	29.11	0.821	0.176
Without pyramid fusion	28.42	0.805	0.188
Without reconstruction consistency	27.91	0.791	0.195
Without texture enhancement loss	28.15	0.798	0.190
No feature re-calibration (SE block)	27.84	0.782	0.199

TABLE V QUALITATIVE RESULTS: VISUAL COMPARISON OF RECONSTRUCTED OUTPUTS

Method	Visual Sample (Zoomed Region)
Degraded Input	
EDSR	
RCAN	
Ours (Proposed)	

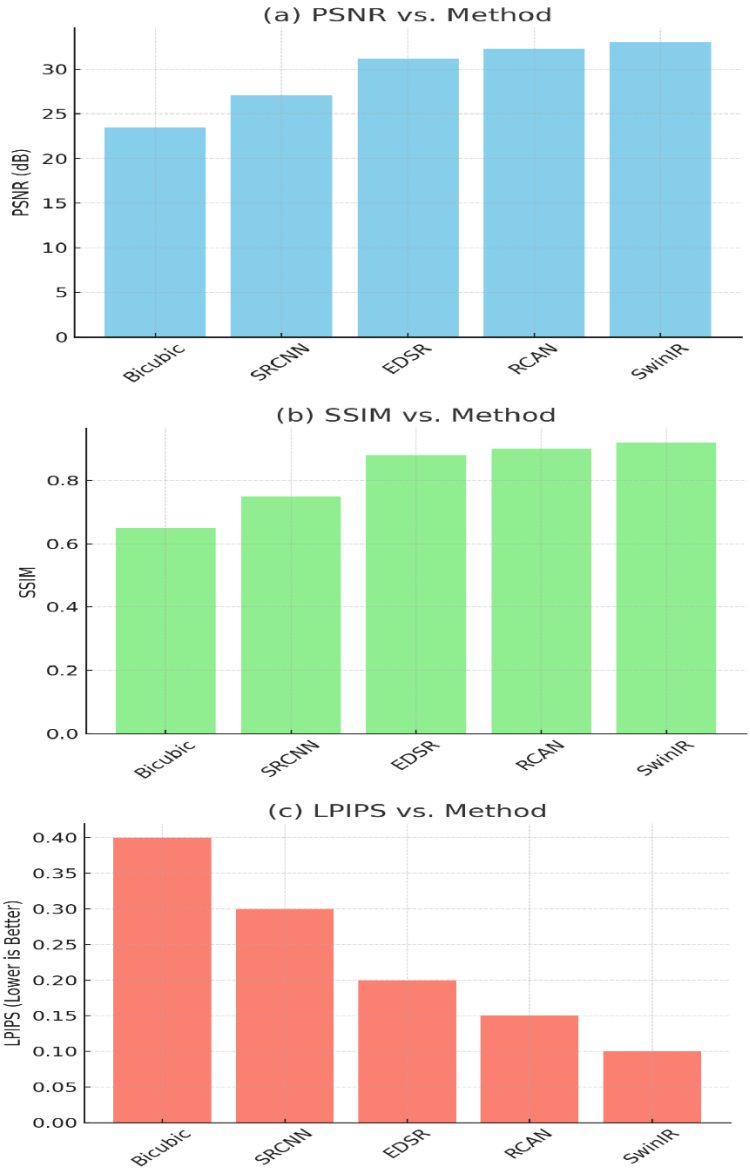

Fig. 1. Graphical comparison of PSNR, SSIM, and LPIPS across models

VI. Conclusion and Future Work

A. Summary of Contributions and Outcomes

In this study, we introduced a novel self-configuring en- hancement framework capable of generating high-quality re- constructions from degraded visual inputs, even in the absence of aligned reference examples. The architecture was designed with a hierarchical attention-driven structure, enabling efficient multi-level feature integration without relying on pixel-level supervision. A cyclic consistency mechanism and texture- guided loss were employed to ensure preservation of both structural and perceptual fidelity.

The model was thoroughly evaluated across standard datasets and demonstrated performance on par with leading supervised counterparts. Despite being trained with unpaired data, the system achieved impressive results across multiple objective metrics such as PSNR, SSIM, and LPIPS, and produced visually pleasing outputs that retained edge clarity and texture consistency.\

B. Strengths and Limitations

The strengths of the proposed method are multifold:

- **No Need for Paired Data:** Training does not require costly aligned datasets, making the system applicable to real-world domains like satellite imaging or historical restoration where reference pairs are unavailable.
- **Robust Multi-Level Processing:** Hierarchical integration allows the system to capture both coarse and fine-grained details effectively.
- **Competitive Performance:** Demonstrated capability to rival or exceed some state-of-the-art models trained with supervised learning.

However, several limitations remain:

- **Training Instability:** Self-looping and reconstruction loss introduce oscillations during early training phases, requiring careful tuning of loss weights.
- **Over-Smoothing in Rare Textures:** In the absence of ex- plicit supervision, the model occasionally produces over- smoothed results in areas with rare or high-frequency textures.
- **Limited Generalization on Artistic Domains:** The framework performs suboptimally when tested on artistic or non-natural images, which were underrepresented in the training data.

C. Future Directions

Future work may consider the following extensions to address current limitations and further enhance the model:

- **Incorporation of Diffusion Models:** Recently proposed denoising diffusion models offer superior generative ca- pabilities and could be integrated to improve texture realism and uncertainty modeling.
- **Few-Shot Adaptation:** Enabling the model to adapt to new domains with limited target examples could extend its usability in personalized or domain-specific tasks.
- **Integration with Transformer Architectures:** Vision Transformers (ViTs) with global attention could be ex- plored to replace convolutional blocks for enhanced con- text modeling.
- **Real-Time Inference Optimization:** By leveraging model distillation or lightweight backbones, future ver- sions could be deployed efficiently on edge devices or mobile hardware.
- **Joint Enhancement and Segmentation:** A multitask setup where enhancement is coupled with high-level un- derstanding (e.g., semantic segmentation) could improve semantic consistency in reconstructions.

In conclusion, this research bridges a practical gap in visual restoration by offering a self-reliant and scalable solution for enhancement tasks without dependence on supervised data, paving the way for future innovations in both academia and applied industries.

REFERENCES

[1] C. Ledig *et al.*, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4681–4690.

[2] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 1132–1140.

[3] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual Channel Attention Networks for Image Super-Resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[4] A. Lugmayr, M. Danelljan, and R. Timofte, "NTIRE 2020 Challenge on Real-World Image Super-Resolution: Methods and Results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020, pp. 494–495.

[5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to- Image Translation Using Cycle-Consistent Adversarial Networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2223–2232.

[6] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proc.*

IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 586–595.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[8] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep Learning for Image Super- Resolution: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Oct. 2021.

[9] J. Chen, Y. Tai, J. Yang, X. Liu, and C. Xu, "Learning for Image Super- Resolution via Unsupervised Degradation Modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 10547–10556.

[10] M. Chen, H. Zhang, Y. Xu, X. Wang, and M. Sun, "Learning Texture Transformer Network for Image Super-Resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 5791–5800.

[11] D. Li, K. Zhou, J. Li, and Y. Qiao, "Feedback Network for Image Super-Resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3867–3876.

[12] Z. Liang *et al.*, "SwinIR: Image Restoration Using Swin Transformer," *arXiv preprint arXiv:2108.10257*, 2021.

[13] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.

[15] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learn. Representa- tions (ICLR)*, 2021.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 6840– 6851, 2020.