LightSR: Lightweight and Efficient Image Super-Resolution

Varun Sankar Moparthi
Electrical and Computer Engineering
A69030810

Chandrakant Indudhar Viraktamath
Electrical and Computer Engineering
A69036779

Project Page
Project Video

Abstract

Single Image Super-Resolution (SISR) aims to reconstruct a high-resolution (HR) image from a given low-resolution (LR) input, a task critical in applications such as mobile photography, medical imaging, video surveillance, and satellite observation. These applications often operate under limited computational resources, especially on edge devices like smartphones, drones, and embedded systems. Therefore, there is a growing need for super-resolution models that are both lightweight and capable of producing high-quality reconstructions.

Key challenges include achieving a balance between model compactness and reconstruction accuracy, minimizing memory and computational overhead, and effectively capturing both local textures and global semantic information. An efficient SISR model must reduce artifacts and noise while preserving structural fidelity, generalize across diverse image domains and scaling factors, and operate in real-time on resource-constrained platforms.

Recent advancements address these challenges by integrating CNNs and Transformers. Notably, SRConvNet a Transformer-style convolutional network incorporating Fourier-modulated attention and dynamic convolutions achieves state-of-the-art performance with low complexity. In this project, we implement and evaluate SRConvNet, training it on the DIV2K and RealSR dataset, a real-world benchmark, to assess its effectiveness in producing high-quality, efficient super-resolution suitable for deployment on edge devices.

1 Introduction

Single Image Super-Resolution (SISR) is the task of reconstructing a high-resolution (HR) image from a single low-resolution (LR) input. As an ill-posed inverse problem, SISR suffers from the inherent ambiguity caused by the loss of high-frequency information during downsampling. Despite this challenge, SISR has become increasingly important across various domains where high-quality visual content is essential.

Applications of SISR span a wide range of real-world scenarios. In **mobile photography**, it enhances image detail without relying on large, power-intensive sensors. In **medical diagnostics**, it improves visual clarity while remaining within safety and exposure limitations. In **video surveillance**, it helps recover crucial facial or object-level details from noisy or low-resolution footage. **Remote sensing** also benefits from SISR, as it enhances spatial resolution in satellite imagery constrained by sensor resolution or bandwidth.

While state-of-the-art deep learning methods have significantly advanced the quality of SISR, they typically rely on deep convolutional or transformer-based models that are computationally and

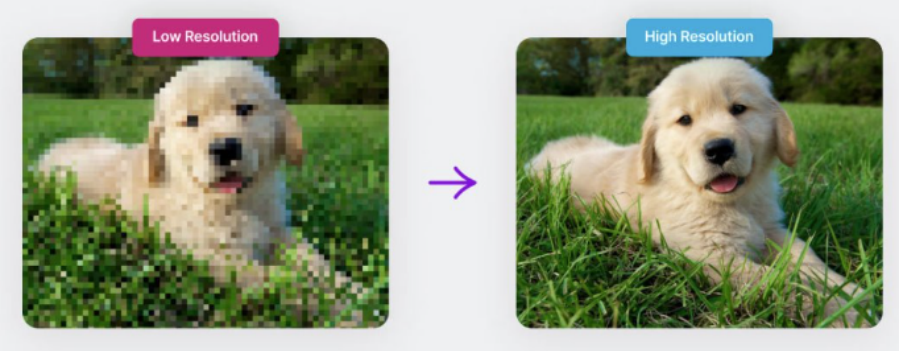

Figure 1: An example of super-resolution based upscaled image

memory intensive. This limits their practicality for deployment on edge devices such as smartphones, drones, and embedded platforms that operate under stringent resource constraints.

2 Motivation

The growing demand for intelligent visual enhancement on edge devices calls for super-resolution models that are not only accurate but also lightweight and efficient. These devices often have limited computing power, memory, and energy availability, making it critical to design models that meet these constraints without compromising visual fidelity.

Several key challenges must be addressed in this context:

- **Model compactness vs. reconstruction quality:** Designing architectures that offer a strong trade-off between performance and efficiency.
- **Capturing both local and global context:** Local textures and edges contribute to detail sharpness, while global context ensures semantic consistency.
- **Minimizing artifacts:** The upscaling process must avoid introducing visual distortions or noise.
- **Generalization:** Models should perform well across different image domains, degradations, and upscaling factors.
- **Real-time inference:** The ability to run efficiently on resource-limited hardware is essential for practical use.

To this end, we explore lightweight super-resolution architectures, specifically focusing on transformer-inspired convolutional networks such as *SRConvNet*, which incorporate Fourier-modulated attention and dynamic convolution layers. Our aim is to implement and evaluate such models on real-world datasets like *RealSR* and *DIV2K*, assessing their viability for deployment in edge environments.

3 Previous Works

Recent advances in image super-resolution have focused on reducing model complexity while maintaining high reconstruction quality, particularly for deployment on edge devices. Fang et al. introduced the Hybrid Network of CNN and Transformer (HNCT), which combines convolutional layers for efficient feature extraction with Swin Transformers and Enhanced Spatial Attention (ESA) blocks to capture both local and global dependencies. This approach achieves a strong balance between accuracy and computational efficiency, making it suitable for resource-constrained platforms.

Zou et al. proposed the Self-Calibrated Efficient Transformer (SCET), a fully lightweight model that decouples spatial and channel modeling. SCET leverages channel-wise attention to reduce computational complexity while maintaining strong super-resolution performance, making it highly optimized for deployment on constrained hardware.

Garas et al. developed a simple Transformer-style network (STSN) for lightweight image super-resolution. Their model uses convolutional modulation blocks (Conv2Former) to replace traditional self-attention mechanisms, achieving competitive results with reduced computational demands, suitable for devices with limited computing power.

Li et al. proposed SRConvNet, a convolutional network that mimics Transformer architectures by integrating Fourier and dynamic layers. SRConvNet delivers high-quality super-resolution results while maintaining fast and efficient inference, making it a promising approach for lightweight image super-resolution.

Our project involves implementing SRConvNet and training it on the RealSR dataset, a real-world image super-resolution benchmark introduced at CVPR 2020, which focuses on realistic degradations and noise

4 Methods

4.1 Model Architecture: SRConvNet

SRConvNet is a lightweight and efficient architecture for Single Image Super-Resolution (SISR) that blends convolutional efficiency with transformer-inspired attention mechanisms. The design focuses on maintaining high visual fidelity while enabling real-time deployment on resource-constrained platforms.

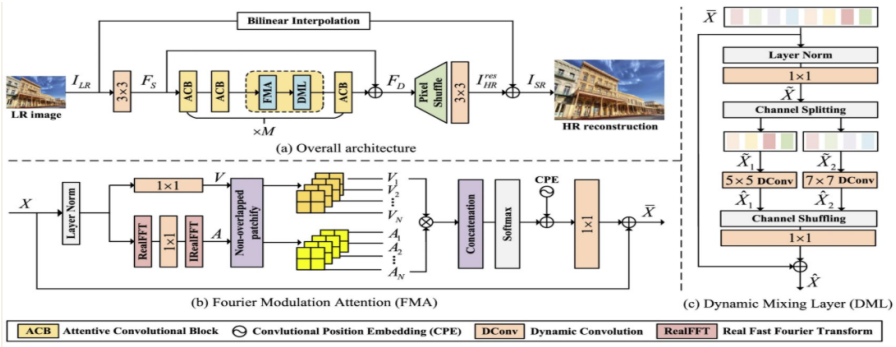

Figure 2: SuperConvNet Model Architecture

4.1.1 Architecture Overview

- **Initial Feature Extraction:** The input low-resolution (LR) image is passed through a standard convolutional layer followed by a GELU activation function to extract shallow features.
- **Context-Aware Refinement:** A stack of *Attentive Convolutional Blocks (ACBs)* refines the features by capturing both local and non-local dependencies efficiently.
- **Upsampling:** The refined feature maps are upsampled using either Pixel Shuffle or Bilinear Interpolation, depending on the deployment setting.
- **Residual Learning:** A residual connection is established by interpolating the original LR input to the target resolution and adding it to the network output, promoting stability and learning efficiency.

4.1.2 Fourier Modulated Attention (FMA)

To enhance global feature modeling in a computationally efficient manner, SRConvNet integrates a Fourier Modulated Attention (FMA) module.

*Design and Functionality

- The FMA module projects intermediate features into the frequency domain using a Fourier transform.

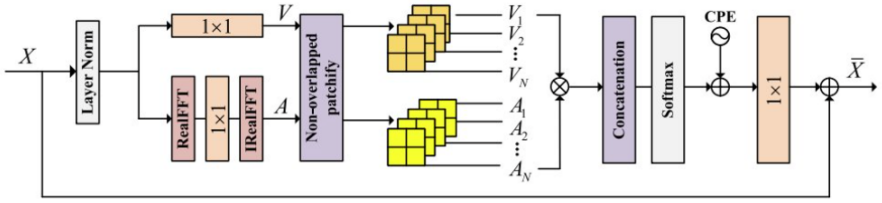

Figure 3: SuperConvNet - Fourier Modulated Attention Module

- It applies modulation based on attention maps computed in this frequency space.
- The modulated features are then inverse-transformed back into the spatial domain to obtain enriched global representations.

*Advantages

- Enables long-range dependency modeling at a reduced computational cost compared to traditional self-attention mechanisms.
- Preserves global semantic structure, improving overall super-resolution quality.

4.1.3 Attentive Convolutional Blocks (ACBs)

Each ACB is designed to be both context-aware and lightweight, offering an efficient mechanism to extract fine-grained local features.

- Utilizes spatial chunking to divide input features into patches.
- Applies multi-head convolutional attention with diverse kernel sizes (e.g., 5×5 and 7×7), enabling effective capture of both small-scale textures and broader contextual features.

4.1.4 Dynamic Multi-Scale Learning (DML)

To handle variations in scale and structure across different images, the architecture incorporates a DML mechanism:

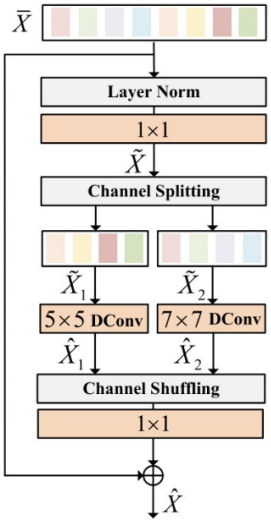

Figure 4: SuperConvNet - Dynamic Multi-Scale Learning Module

- Learns adaptive representations across multiple resolutions through dynamic reshaping.

- Facilitates information exchange between scales to improve contextual aggregation and robustness to input variation.

4.1.5 Combined Impact

The integration of FMA, ACB, and DML mechanisms allows SRConvNet to effectively capture both local texture details and global semantic coherence. These modules work in tandem to ensure high reconstruction quality while keeping the model lightweight and suitable for deployment on edge devices. The residual learning path further enhances convergence and stabilizes the reconstruction process.

4.2 Dataset Details

To train and evaluate SRConvNet across varying levels of degradation and resolution, we employ both real-world and synthetic datasets. These datasets support $2\times$, $3\times$, and $4\times$ upscaling factors and offer a diverse set of image characteristics, ensuring model robustness and generalization.

4.2.1 Training Datasets

RealSR:

- A real-world dataset consisting of approximately 595 low-resolution (LR) and high-resolution (HR) image pairs.
- Images are captured using DSLR cameras across multiple focal lengths and devices.
- Contains real degradations caused by optical blur, sensor noise, and misalignment, unlike synthetic datasets.
- Supports $2\times$, $3\times$, and $4\times$ super-resolution tasks.
- Image resolutions vary and simulate practical super-resolution scenarios on real camera hardware.

DIV2K:

- A high-quality dataset comprising 800 training and 100 validation images at 2K resolution (typically 1920×1080 or higher).
- LR images are generated using bicubic downsampling at $2\times$, $3\times$, and $4\times$ scales.
- Serves as a clean, standardized benchmark for training and validating super-resolution models in a controlled setting.

4.2.2 Evaluation Dataset

We evaluate the trained models on standard SISR benchmarks hosted on FigShare:

- **Set5**, **Set14**, **BSD100**, and **Urban100** – widely adopted datasets in the SR community.
- Available for $2\times$, $3\times$, and $4\times$ upscaling evaluations.
- These benchmarks provide diverse content including smooth textures, sharp edges, complex structures, and urban scenes.
- Enable standardized comparison of model performance in terms of PSNR, SSIM, and visual quality.

5 Experiments

5.1 Training Setup

To ensure effective and efficient training of SRConvNet across multiple upscaling factors, we adopt a modular training pipeline leveraging mixed precision and distributed strategies.

5.1.1 Training Configuration

- **Scale-specific Models:** Separate models are trained independently for $2\times$, $3\times$, and $4\times$ super-resolution tasks.
- **Precision:** Mixed precision training is employed using PyTorch's Automatic Mixed Precision (AMP) to accelerate training and reduce memory usage.
- **Optimizer:** Adam optimizer with a learning rate of 1×10^{-4} and $\epsilon = 1 \times 10^{-8}$ is used for stable convergence.
- **Scheduler:** A MultiStepLR scheduler reduces the learning rate by a factor of 0.5 at epochs 50 and 100.
- **Batching and Distribution:** Batch size is set to 2 images per GPU, and distributed training is implemented using PyTorch's DistributedDataParallel (DDP).
- **Loss Function:** The network is trained using L1 loss computed on the RGB channels between predicted and ground-truth high-resolution images.
- **Patch-based Training:** Training is performed on image patches of size 96×96, extracted randomly with standard data augmentations (e.g., flipping, rotation).
- **Transfer Initialization:** For improved convergence, pretrained weights from the $2\times$ model are used to initialize $3\times$ and $4\times$ models.

5.1.2 Validation and Checkpoints

- **Validation Metrics:** Every 5 epochs, the model is evaluated on a held-out validation set using PSNR and SSIM computed on the luminance (Y) channel in the YCbCr color space.
- **Model Saving:** Checkpoints and corresponding evaluation metrics are saved every 5 epochs for tracking training progress and enabling model selection.

6 Results and Analysis

The SRConvNet models were trained and evaluated for single image super-resolution at scaling factors of 2×, 3×, and 4×. The results demonstrate the effectiveness and stability of the proposed architecture across all scales.

6.1 Training and Validation Progress for DIV2K

To assess the learning behavior and generalization capability of SRConvNet, we monitored the training and validation loss across epochs for each upscaling factor ($2\times$, $3\times$, and $4\times$). The loss curves are depicted in Figure 5, where both training and validation losses decrease steadily and converge to low values, indicating stable and effective training. The close alignment between training and validation losses suggests minimal overfitting and strong generalization to unseen data.

For all scaling factors, the loss curves demonstrate rapid initial convergence followed by gradual stabilization. Notably:

- **$2\times$ Upscaling:** Both training and validation losses quickly decrease and remain closely matched throughout training, reflecting efficient learning and strong generalization.
- **$3\times$ Upscaling:** The model maintains low and stable losses, with only minor fluctuations, confirming robust convergence.
- **$4\times$ Upscaling:** Although the loss values are slightly higher due to the increased difficulty of the task, the training and validation curves remain closely aligned, indicating that the model effectively handles more challenging super-resolution scenarios.

The consistent convergence and minimal gap between training and validation losses across all scales validate the design choices in SRConvNet, including the use of Fourier Modulated Attention and Dynamic Multi-Scale Learning. These results highlight the model's capability to balance efficiency and accuracy, making it well-suited for real-world deployment on edge devices

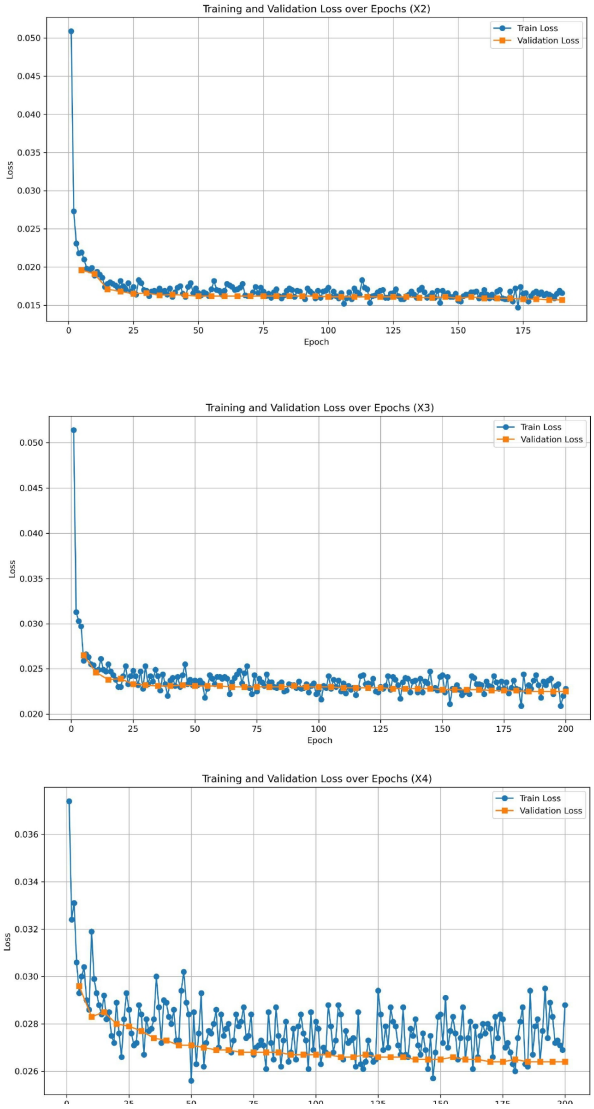

Figure 5: Training and validation loss curves for SRConvNet at $2\times$, $3\times$, and $4\times$ upscaling factors.

6.2 Results on DIV2K dataset

The SRConvNet model was evaluated on four widely used benchmark datasets: BSD100, Set5, Set14, and Urban100. Performance was measured using the Peak Signal-to-Noise Ratio (PSNR) metric at three upscaling factors: $\times 2$, $\times 3$, and $\times 4$. The PSNR trends for each dataset and scale are illustrated in Figure 6, and detailed quantitative results are summarized in Table 1.

The validation results reveal several key trends:

- **Consistent Convergence:** PSNR improves rapidly in the initial epochs and stabilizes, indicating effective learning and convergence.
- **Scaling Factor Impact:** As the upscaling factor increases from $\times 2$ to $\times 4$, PSNR values decrease for all datasets, reflecting the increased challenge of reconstructing fine details.
- **Dataset Complexity:** Set5 achieves the highest PSNR values, likely due to its simpler image content. BSD100 and Set14 show moderate performance, while Urban100, with complex urban scenes, yields the lowest PSNR, especially at higher scales.

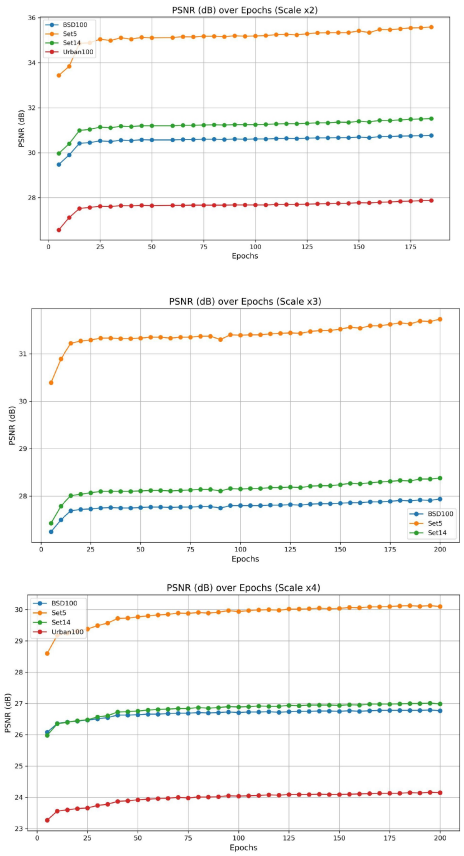

Figure 6: PSNR (dB) over epochs for BSD100, Set5, Set14, and Urban100 validation datasets at $\times 2$, $\times 3$, and $\times 4$ upscaling factors.

DIV2K Dataset								
	PSNR							
	BSD100		Set5		Set14		Urban100	
Model	Max	Final	Max	Final	Max	Final	Max	Final
X2	30.77	30.77	35.59	35.59	31.52	31.52	27.88	27.88
X3	27.94	27.94	31.73	31.73	28.38	28.38	N/A	N/A
X4	26.79	26.77	30.13	30.1	27.01	26.99	24.16	24.15

Table 1: Maximum and final PSNR values on validation datasets for each scaling factor.

- **Stability:** The difference between maximum and final PSNR values is minimal, suggesting the model maintains peak performance without overfitting.
- **Generalization:** The model demonstrates robust generalization across diverse datasets, as evidenced by stable and competitive PSNR scores.

The SSIM results provide further insight into the perceptual quality and structural fidelity of the super-resolved images:

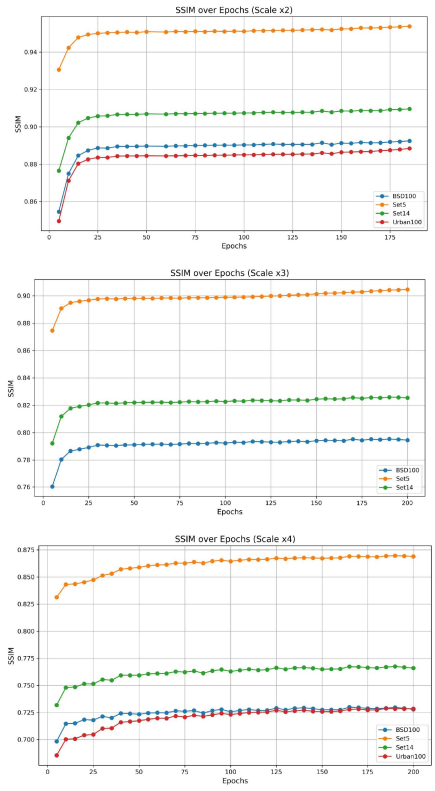

Figure 7: SSIM over epochs for BSD100, Set5, Set14, and Urban100 validation datasets at $\times 2$, $\times 3$, and $\times 4$ upscaling factors.

DIV2K Dataset								
	SSIM							
	BSD100		Set5		Set14		Urban100	
Model	Max	Final	Max	Final	Max	Final	Max	Final
X2	0.8925	0.8925	0.9538	0.9538	0.9096	0.9096	0.8884	0.8884
X3	0.7945	0.7945	0.9047	0.9047	0.8254	0.8254	N/A	N/A
X4	0.7289	0.7278	0.8693	0.8689	0.7667	0.766	0.7284	0.7285

Table 2: Maximum and final SSIM values on validation datasets for each scaling factor.

- **SSIM Convergence:** SSIM values for all datasets increase rapidly during the initial epochs and then plateau, indicating that the model quickly learns to preserve structural similarity in the reconstructed images.

- **Scaling Factor Impact:** As with PSNR, higher upscaling factors ($\times 3$ and $\times 4$) lead to lower SSIM scores across all datasets, reflecting the increased challenge of maintaining perceptual quality at larger scales.

- **Dataset Trends:** Set5 consistently achieves the highest SSIM values, followed by Set14 and BSD100, while Urban100 records the lowest SSIM, especially at higher scales. This trend highlights the impact of image complexity and texture on perceptual reconstruction.
- **Stability and Robustness:** The minimal difference between maximum and final SSIM values suggests stable training and good generalization, with the model maintaining perceptual quality throughout the training process.
- **Comprehensive Performance:** The combination of high PSNR and SSIM scores across multiple datasets demonstrates that the model not only achieves strong numerical fidelity but also preserves important structural and perceptual aspects of the images.

In summary, the SSIM analysis confirms that LightSR delivers reliable perceptual quality and structural similarity in addition to high PSNR, further validating its effectiveness for image super-resolution tasks across diverse datasets and scaling factors.

6.3 Results for REALSR

For the RealSR dataset, we focused on the Canon subset, which contains 200 images for training and 50 images for validation. To accelerate convergence and leverage learned representations, the model weights were initialized from a previously trained RealSR model. This transfer learning approach allows the network to benefit from features already adapted to real-world image degradations, as recommended in best practices for using pretrained weights.

The model was then fine-tuned for 50 epochs on the Canon training set. The validation performance was monitored throughout training.

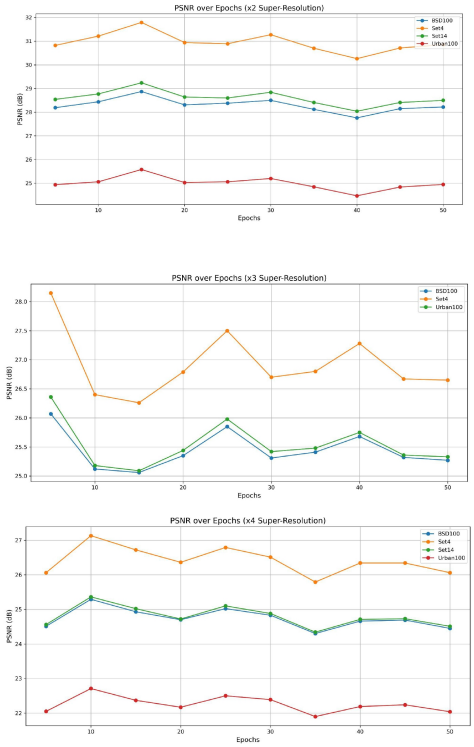

Figure 8: PSNR over epochs for the RealSR Canon validation set at different upscaling factors.

Analysis:

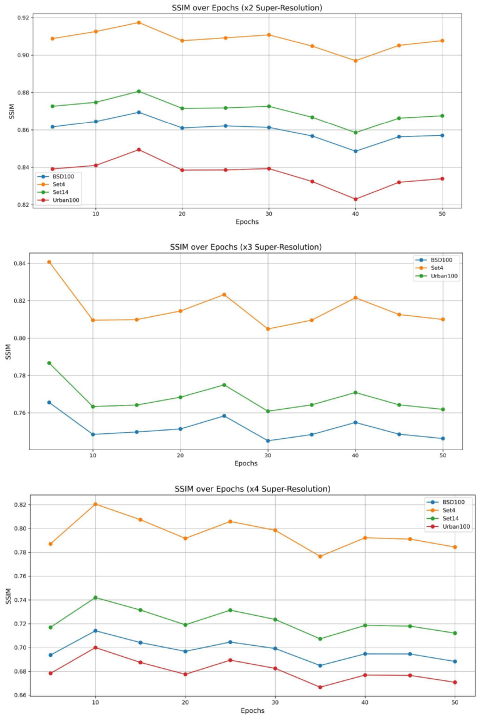

Figure 9: SSIM over epochs for BSD100, Set5, Set14, and Urban100 on the RealSR Canon validation set at $\times 2$, $\times 3$, and $\times 4$ upscaling factors.

Real SR Dataset									
Model	PSNR					SSIM			
	BSD100	Set5	Set14	Urban100		BSD100	Set5	Set14	Urban100
X2	28.87	31.79	29.24	25.58		0.8695	0.9174	0.8807	0.8494
X3	26.07	28.15	26.36	N/A		0.7656	0.8408	0.7867	N/A
X4	25.29	27.13	25.36	22.71		0.7141	0.8205	0.742	0.7

Table 3: PSNR and SSIM values for BSD100, Set5, Set14, and Urban100 on the RealSR Canon validation set at different upscaling factors.

The RealSR Canon results provide insight into the model's ability to generalize to real-world image degradations:

- **PSNR and SSIM Trends:** Both PSNR and SSIM values show clear trends across epochs and scaling factors. For all datasets, the metrics are highest at $\times 2$ upscaling and decrease as the scaling factor increases to $\times 3$ and $\times 4$, reflecting the greater difficulty of reconstructing fine details at larger scales.
- **Dataset Comparison:** Set5 consistently achieves the highest PSNR and SSIM values, indicating that the model performs best on simpler images. BSD100 and Set14 show

moderate performance, while Urban100 presents the greatest challenge, with the lowest scores, especially at higher scales.

- **Stability and Convergence:** The SSIM curves (Figure 9) and the PSNR curves (Figure 8) demonstrate that the model converges quickly and maintains stable performance throughout the 50 epochs, with minimal overfitting observed.
- **Real-World Challenges:** Compared to synthetic datasets, both PSNR and SSIM are generally lower and show more fluctuation, highlighting the increased complexity and variability of real-world data. The limited size of the Canon subset (200 training images) further emphasizes the importance of using pretrained weights for robust performance.
- **Comprehensive Assessment:** The combined PSNR and SSIM results in Table 3 confirm that the model not only achieves reasonable numerical fidelity but also preserves structural similarity and perceptual quality on real-world images.

Overall, these results demonstrate that transfer learning from pretrained RealSR weights is effective for adapting to the Canon subset, and that the model is able to deliver stable and competitive performance on challenging real-world super-resolution tasks.

6.4 Qualitative Results

We present qualitative comparisons of our SRConvNet outputs against bilinear interpolation on both synthetic (DIV2K) and real-world (RealSR) datasets. The visual examples highlight model behavior across different scaling factors ($2\times$, $3\times$, and $4\times$).

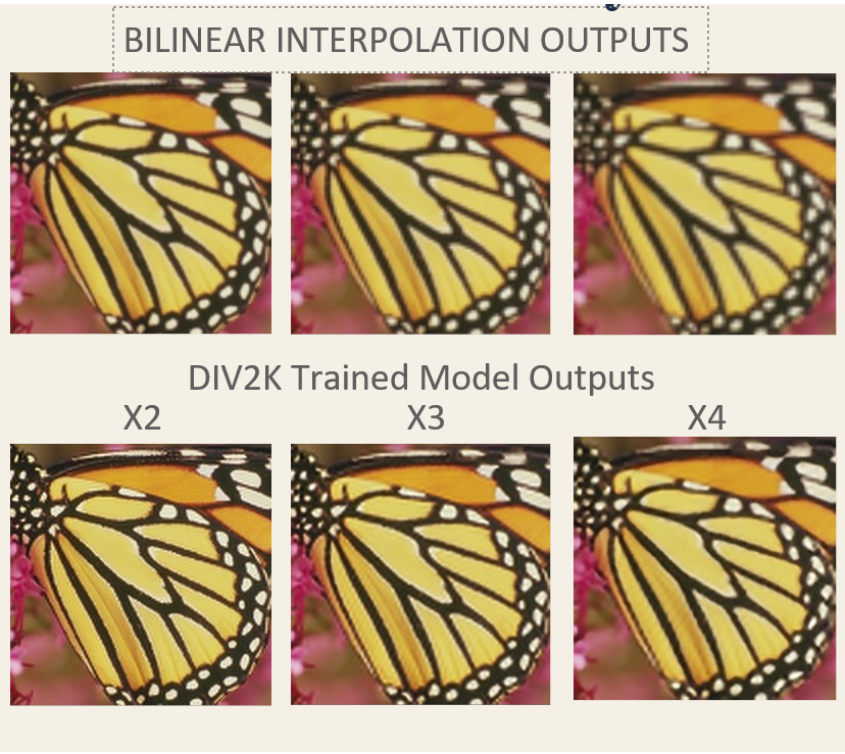

Figure 10: DIV2K model outputs compared to bilinear interpolation across multiple scaling factors.

For the **DIV2K dataset**, the model performs well at $2\times$ upscaling, producing sharp and natural-looking outputs. However, as the scaling factor increases to $3\times$ and $4\times$, the reconstruction quality gradually degrades, especially in fine-textured regions. This decline can be attributed to the increased ill-posedness of the problem more pixel information is lost in higher downscaling, making recovery harder. Despite this, the outputs remain visually realistic and avoid over-smoothing, preserving global structures more effectively than naive interpolation methods.

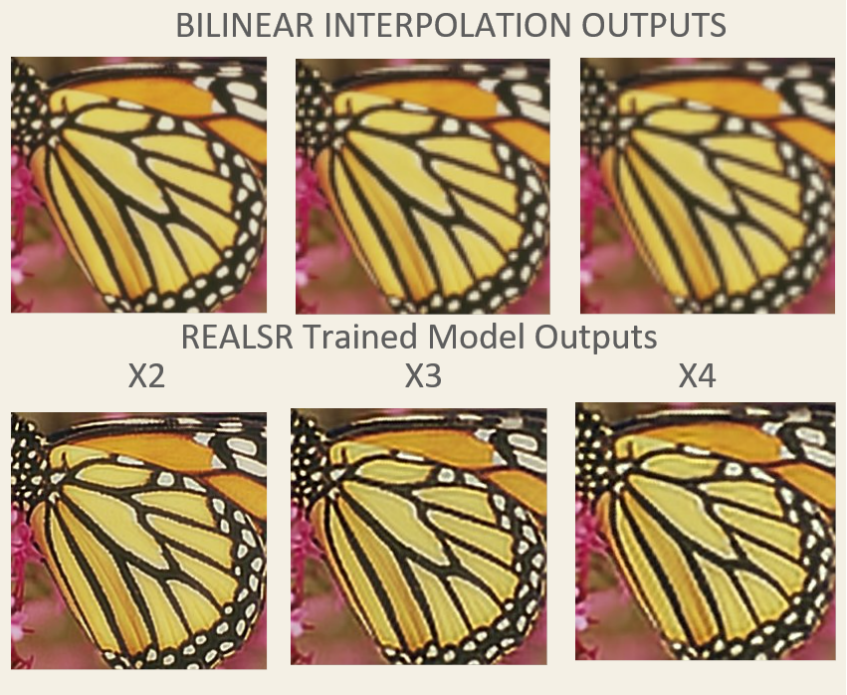

Figure 11: RealSR model outputs compared to bilinear interpolation across multiple scaling factors.

For the **RealSR dataset**, similar trends are observed. At $2\times$ upscaling, the model is able to restore clean and coherent details. However, with $3\times$ and $4\times$ scaling, edge artifacts start to appear particularly in high-contrast and boundary regions. This is likely due to the presence of complex and uncontrolled degradations in the RealSR dataset, such as optical blur, misalignment, and device-dependent noise, which compound at higher scales. While the model still recovers semantically faithful outputs, artifact suppression remains a challenge for real-world high-scale super-resolution.

7 Conclusion and Future Work

7.1 Conclusion

In this work, we implemented **SRConvNet**, a transformer-style convolutional neural network designed for lightweight Single Image Super-Resolution (SISR). The architecture is tailored for deployment on resource-constrained edge devices, offering a strong balance between reconstruction quality and computational efficiency.

Key components of SRConvNet include:

- **Fourier Modulated Attention (FMA):** Efficiently models both global and local dependencies by leveraging frequency-domain modulation.
- **Dynamic Mixing Layer (DML):** Enables spatial adaptability through dynamic multi-scale feature interaction.
- **Attentive Convolutional Blocks (ACBs):** Capture fine-grained local features while maintaining low computational overhead.

Together, these components enable SRConvNet to deliver visually realistic and high-fidelity super-resolved images with a minimal footprint, making it suitable for real-time applications on mobile and embedded platforms.

7.2 Future Work

Future extensions of this work can explore integrating SRConvNet into higher-level computer vision pipelines:

- **Semantic Segmentation:** Using SRConvNet as a pre-processing step for low-resolution or noisy inputs can improve boundary clarity and preserve feature integrity, potentially enhancing segmentation accuracy.
- **Object Detection:** SRConvNet can be incorporated as a pre-upscaling module within object detection frameworks (e.g., YOLO, Faster R-CNN). Improved resolution is expected to aid in detecting small objects and maintaining spatial consistency, which is critical in applications such as surveillance and autonomous driving.

8 Project Resources

The following resources are publicly available for further reference (UCSD ID), demonstration, and reproducibility:

Project Page
Project Video

9 References

1. Jinsheng Fang, Hanjiang Lin, Xinyu Chen, and Kun Zeng. "A Hybrid Network of CNN and Transformer for Lightweight Image Super-Resolution." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022, pp. 1102–1110.
2. Wenbin Zou, Tian Ye, Weixin Zheng, Yunchen Zhang, Liang Chen, and Yi Wu. "Self-Calibrated Efficient Transformer for Lightweight Super-Resolution." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022, pp. 930–940.
3. Garas Gendy, Nabil Sabor, Jingchao Hou, and Guanghui He. "A Simple Transformer-Style Network for Lightweight Image Super-Resolution." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR), 2023, pp. 1484–1494.
4. F. Li, R. Cong, J. Wu, et al. "SRConvNet: A Transformer-Style ConvNet for Lightweight Image Super-Resolution." International Journal of Computer Vision, 133, 173–189 (2025). https://doi.org/10.1007/s11263-024-02147-y.
5. Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. "Real-World Super-Resolution via Kernel Estimation and Noise Injection." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 466–467.
6. Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. "Toward Real-World Single Image Super-Resolution: A New Benchmark and A New Model." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3086–3095.
7. Eirikur Agustsson and Radu Timofte. "NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1122–1131.