Multi-Scale Attention-Enhanced U-Net for Image Deblurring with Composite Loss Functions

Abstract——**Image deblurring is a challenging low-level vision task aiming to recover a sharp image from a blurred input. In this paper, we propose a Multi-Scale Attention-Enhanced U-Net model for single-image motion deblurring, combined with a novel composite loss that integrates spatial, perceptual, and frequency-domain loss functions. The network extends the U-Net architecture with multi-scale feature decoding and attention mechanisms (Convolutional Block Attention Module (CBAM) in each encoder/decoder block, and a self-attention block at the bottleneck) to better capture both local details and global dependencies. Our composite loss includes pixel-wise L1 loss, perceptual loss using VGG16 features, and a Fourier-domain loss to enforce consistency in image frequency content. Trained on the GOPRO dataset with minimal data augmentation, our model achieves an average Peak Signal-to-Noise Ratio (PSNR) of approximately 28.1 dB and Structural Similarity (SSIM) of 0.815 on the test set. These results demonstrate that the proposed attention-enhanced multi-scale approach, together with the hybrid loss functions, yields competitive deblurring performance while preserving image details and textures. Index Terms—Image Deblurring, U-Net, Attention Mechanisms, CBAM, Self-Attention, Composite Loss, Fourier Domain, PSNR, SSIM**

I. INTRODUCTION

A frequent issue that significantly compromises image quality and impairs the performance of following computer vision activities is motion blur. Camera wobble or object movement during exposure causes it. Image deblurring—the technique of retrieving a sharp, clear image from a blurred input—is a difficult and complicated issue given the great variety of potential blur causes, including varied motion patterns, variable blur sizes, and inconsistent directions.

Traditional deblurring methods, which depend on deconvolution or manually crafted priors, have struggled with complex or irregular blur patterns. Deep learning has come a long way recently, enabling networks to directly acquire feature representations from data, often surpassing more traditional methods. Many contemporary deep models, nevertheless, still find it difficult to precisely restore delicate features and textures. Long training times and significant data augmentation are often required for these models to generalize well.

To address these challenges, we offer a new deep learning framework in this paper that integrates innovative architectural design with a well-considered loss function. Enhanced by advanced attention technologies, we offer a Multi-Scale U-Net architecture that adaptively focuses on important visual features at several scales. Our model, in particular, integrates Convolutional Block Attention Modules (CBAM) in the encoder and decoder layers, which give channel and spatial attention top priority of key characteristics. A self-attention block captures long-range dependencies at the bottleneck, hence allowing the network to understand global contextual information.

A major feature of our approach is the inclusion of a composite loss function combining Fourier-domain loss, perceptual loss, and pixel-wise L1 loss. Though many of them depend on pixel-based losses (L1/L2) and sometimes use perceptual losses with pre-trained networks, few current techniques have included a Fourier-domain loss to maintain high-frequency information. By using this Fourier loss, we particularly encourage the restoration of small textures and edges—which are often lost during the blurring process, especially in the high-frequency components of the image. Apart from generating visually cleaner results, this hybrid loss formulation reduces blurring artifacts in both the spatial and frequency domains and improves superior quantitative measures.

Our approach is also quite successful in training. By obtaining competitive results on the GoPro dataset using just uniform cropping and resizing as data augmentation, we demonstrate that the network can learn rapidly from the available data without needing complex augmentation methods. Our model's training efficiency reduces computational load and training complexity, therefore qualifying it for practical use. Its small model design (about 30 million parameters) also allows it to be trained on a single GPU, hence offering a fair balance between efficiency and performance.

Ultimately, the key contributions of this work are:

- We propose a U-Net architecture augmented with self-attention, CBAM components, and multi-scale feature learning to boost feature refinement and capture global context.
- We propose a hybrid loss function combining pixel-wise L1, perceptual (VGG16-based), and Fourier-domain losses to handle blur from both a spatial and a frequency viewpoint. This significantly improves the preservation of textures and high-frequency information.
- Our model has an average PSNR of around 28.1 dB and SSIM of approximately 0.815, which allows it to compete in deblurring performance using the GoPro dataset. The model shows a good trade-off between efficiency and performance with a fairly small model of roughly 30 million parameters, training efficiently with minimal augmentation and a simple training schedule.

II. RELATED WORK

A. Deep Learning Methods for Image Deblurring

Early deep learning techniques for image deblurring enabled multi-scale structures. Using a coarse-to-fine strategy, Nah et al. [1] developed an early end-to-end CNN model processing the input image across multiple scales to progressively reconstruct sharp images. The model initially removes a coarse approximation of blur before refining features to effectively handle larger blur effects. Tao et al. [2] achieved significant advances in real-time deblurring with reduced computational cost through their Scale-Recurrent Network (SRN), enhancing this approach by exchanging weights across scales. Numerous subsequent advances in the field have been based on these multi-scale encoder-decoder architectures.

Researchers have also explored architectural modifications beyond traditional U-Net architectures for image deblurring. Kupyn et al. [4] introduced DeblurGAN, a generative adversarial network (GAN) utilizing a U-Net generator and adversarial loss to produce perceptually realistic outputs. DeblurGAN-v2 further improved upon this architecture by integrating a feature pyramid network and a dual-scale discriminator, enabling faster and superior deblurring outcomes. Zhang et al. [6] proposed a multi-patch hierarchical network specifically designed to manage spatially variable blur patterns. Addressing similar challenges, Suin et al. [7] developed a spatially-attentive patch-hierarchical CNN, which adaptively analyzes multiple image regions, thus significantly improving the handling of non-uniform blur scenarios. More recent complex approaches, such as MPRNet, have pushed state-of-the-art performance at the cost of increased complexity, utilizing iterative refinements through successive sub-networks.

B. Attention Mechanisms for Deblurring

Attention mechanisms have become a powerful tool in picture restoration tasks, such as deblurring, by improving the model's ability to focus on relevant features. The Convolutional Block Attention Module (CBAM) has been used to apply both channel attention and spatial attention, allowing the network to dynamically concentrate on important picture areas. CBAM has shown to be successful in stressing relevant image features in image restoration tasks such edges or objects requiring extra deblurring attention. Self-attention, sometimes called non-local attention, has been added to image restoration networks to help them capture long-range dependencies by means of distance pixel or feature relationship computation. This attention mechanism has shown particularly beneficial for photos with complex, spatially changing blur patterns since it allows the model to focus on context and smaller features all around the image.

Our study builds on these concepts by including self-attention and CBAM mechanisms into a multi-scale U-Net framework. While giving priority to significant characteristics across several scales, this allows our model to catch global context via long-range dependencies. Particularly in challenging scenarios with complicated and irregular blur patterns, integrating both of these attention mechanisms helps us to achieve higher deblurring performance.

C. Loss Functions for Deblurring

The selection of loss function greatly affects the performance of deblurring models. Many early deep learning-based deblurring methods focused mostly on optimising fundamental pixel-wise losses, including L2 (mean squared error) or L1 (mean absolute error), between the predicted and ground truth images. Though they ensure a minimum degree of integrity to the sharp image, these losses often generate outputs that are too smooth and fail to capture the perceptual quality of the image. Johnson et al. were the first to contrast texture and content using high-level feature maps from a pre-trained network (e.g., VGG16), perceptual losses having been included to solve this. This approach has been widely used in deblurring since it increases the model's ability to keep fine textures and encourages the recovery of important perceptual information.

Adversarial losses assist the model in generating more realistic images, as shown in DeblurGAN, which then sharpens the output. Conversely, adversarial losses can be difficult to stabilise during training and often require precise hyperparameter adjustment. By use of a Fourier-domain loss combined with pixel-wise and perceptive losses, our approach, however, directly tackles the high-frequency components often lost in blurred images. The Fourier-domain loss encourages the restoration of edges and fine features by means of a comparison of the frequency spectra of the ground truth and anticipated images. Restoring high-frequency textures typically muted in blurred photos improves our model's capacity to exactly reconstruct sharp, detailed structures.

Though they have been investigated in various picture restoration activities including super-resolution, the application of Fourier-domain losses in deblurring has been relatively under-researched. Our approach bridges this gap and ensures that our model restores the correct distribution of high-frequency material by combining Fourier loss with spatial and perceptual losses, hence reducing pixel error and perceptual discrepancies. Especially in terms of edge and fine detail restoration, our composite loss function greatly enhances our work by generating observable sharper outcomes and enhanced quantitative measurements.

D. Our Approach

Though our work draws on present model architectures and loss functions, it presents some significant ideas that significantly improve image deblurring performance. By use of self-attention, CBAM, multi-scale feature learning, and a composite loss function (including Fourier-domain loss), our model outperforms both qualitative and quantitative measures. Because it employs a compact network architecture and little data augmentation for training, which ensures efficient training and practical application, our approach is also quite suitable for real-world use with limited computational resources.

Ultimately, we distinguish ourselves by:

- Multi-scale feature extraction is paired with attention mechanisms (CBAM and self-attention) to enhance the model's ability to rank important characteristics and capture local and global interdependence.
- A composite loss function combining Fourier-domain loss, pixel-wise L1, and perceptual loss is suggested to recover high-frequency features and preserve perceptual quality.
- Achieving cutting-edge performance on the GoPro dataset with minimal data augmentation and a computationally efficient approach that generates great results without demanding a lot of resources or training time.

III. PROPOSED METHODOLOGY

Our proposed approach introduces an innovative deep learning framework for image deblurring, utilising a proprietary CNN architecture and a composite training loss. This section delineates the architecture of our model, encompassing attention modules, a multi-scale design, and a hybrid loss function that optimises the model's performance.

A. Network Architecture

Our model is based on a U-Net-style encoder-decoder architecture, designed to perform effectively over various spatial scales. A schematic of the architecture is depicted in Figure **??** (placeholder). The decoder progressively upsamples and refines the output to the original resolution, while the encoder employs a series of convolutional blocks to handle the blurred input image, systematically downsampling the feature maps to capture coarse information.

Our encoder features four stages of downsampling. Each downsampling level comprises a DoubleConv block featuring two 3×3 convolutional layers, a 2×2 max-pooling layer, batch normalisation, and ReLU activation. The encoder produces feature maps with 96, 192, 384, and 768 channels at the first, second, third, and fourth levels, respectively. The quantity of feature channels doubles at each subsequent level. A bottleneck module employing dilated convolutions is incorporated at the base of the U-Net, subsequent to the fourth encoder block. At the bottleneck, these convolutions produce 768-channel feature maps by enlarging the receptive field without further downsampling.

We incorporate attention mechanisms into the model to enhance the network's representational capacity.

1) CBAM in Encoder/Decoder Modules: A Convolutional Block Attention Module (CBAM) is incorporated after each DoubleConv block in both the encoder and decoder. The CBAM employs spatial attention by generating a spatial significance map from pooled information across all channels, following the application of channel attention through the computation of channel-wise importance weights via global average and max pooling. By reducing less relevant information and emphasising essential visual components (such as edges and textures), these attention maps recalibrate the feature responses. The network may dynamically concentrate on the most critical features across different resolutions by implementing attention at each scale.

2) Self-Attention Mechanism: A self-attention mechanism exists at the bottleneck between the encoder and the decoder. This layer computes attention maps based on pairwise feature similarity, allowing each position in the feature map to attend to all other positions globally. The 768-channel bottleneck feature map is utilised to compute query, key, and value projections in the self-attention layer, executed as a non-local process. The weighted sum of the value features is returned to the input as a residual, while the attention map is derived from the softmax of the dot product between the query and key vectors. To maintain consistency in the deblurring process, the model's capacity to capture long-range relationships via global self-attention facilitates the dissemination of information regarding blurred structures across the image.

The decoder mirrors the encoder through four upsampling stages. At each decoder level, the feature map is concatenated with the corresponding feature map from the encoder (skip connection) following upsampling by a factor of two through 2×2 transposed convolutions. Subsequent to the encoder, the concatenated feature is processed by a DoubleConv + CBAM block. The resultant image is generated using a 1×1 convolution subsequent to the last decoder step.

Our network generates intermediate outputs at multiple resolutions alongside the final full-resolution deblurred image for multi-scale supervision. The network generates three specific output predictions:

- `out_qtr`: The output of the most profound decoder layer at a quarter resolution.
- `out_half`: The output from an intermediate decoder layer at half-resolution.
- `out_full`: The output of the last decoder layer at full resolution.

Each output is generated by a 1×1 convolution from the respective decoder feature map, utilising 384 channels for quarter-resolution, 192 for half-resolution, and 96 for full-resolution, and is projected onto three RGB channels.

Our model comprises around 29.8 million parameters in total. The model remains computationally efficient and can do real-time inference for 720p pictures on modern GPUs, even with the incorporation of attention modules. A forward pass on an NVIDIA RTX series GPU requires merely tens of milliseconds.

B. Loss Functions in Composite Models

A fundamental element of our approach is the composite training loss, which guides the network towards superior deblurring by integrating many complementing objectives. Let I_b denote the blurred input image, I_s signify the sharp ground truth image, and I_{pred} represent the network's deblurred output at full resolution. The total loss function is defined as follows:

$$L_{\text{total}} = L_{L1}(I_{\text{pred}}, I_s) + \lambda_p L_{\text{perc}}(I_{\text{pred}}, I_s) + \lambda_f L_{\text{Fourier}}(I_{\text{pred}}, I_s)$$

where:

- L_{L1} represents the loss of spatial L1.

- L_{perc} constitutes the perceptual loss.
- L_{Fourier} represents the attenuation in the frequency domain.
- λ_p and λ_f modulate the respective weights of the perceptual and Fourier components.

To ensure uniform contribution of each term during training, we empirically establish these weights (e.g., $\lambda_p = 0.1$ and $\lambda_f = 0.1$).

1) Pixel-wise L1 loss: We compute the mean absolute error between the ground truth and forecasted images:

$$L_{L1} = \frac{1}{N} \sum_{x,y} |I_{\text{pred}}(x, y) - I_s(x, y)|$$

where the sum is over all pixel locations, and N is the number of pixels. In contrast to L2 loss, L1 loss aids in maintaining edges and encourages the model to produce outputs that closely align with the ground truth at the pixel level.

2) Perceptual Loss: To assess perceptual similarity, we input the ground truth and predicted images into a pretrained VGG-16 network (trained on ImageNet) and extract feature maps from designated layers. Specifically, we utilise the `conv3_3` and `conv4_3` layers. The aggregate of the L2 variances among these feature maps signifies the perceptual loss:

$$L_{\text{perc}} = \sum_{l} \frac{1}{C_l H_l W_l} \|\phi_l(I_{\text{pred}}) - \phi_l(I_s)\|_2^2$$

where $\phi_l(\cdot)$ denotes the feature map of the VGG layer. We promote the model to align high-level feature representations, which are more perceptually congruent with human vision, by reducing this loss.

3) Fourier-domain Loss: This loss function evaluates the amplitude spectra of both the target and predicted images. We calculate the L1 loss between the magnitude spectra of the two images by executing a 2D discrete Fourier transform (DFT) on each.

$$L_{\text{Fourier}} = \frac{1}{N} \sum_{u,v} ||F(I_{\text{pred}})(u, v)| - |F(I_s)(u, v)||$$

This loss enables the model to recover sharp edges often diminished during the blurring process by concentrating on the restoration of small details in high-frequency components.

4) Multi-scale Supervision: We utilise the identical composite loss for intermediate outputs alongside the loss computed on the full-resolution output. To compare these intermediate outputs with the ground truth image, they are upsampled to the original resolution. Multi-scale supervision provides a coarse-to-fine learning signal by initially deblurring at lower scales before refining at higher resolutions, hence accelerating convergence and improving stability. The total loss of the network is subsequently:

$$L_{\text{multi-scale}} = L_{\text{total}}(I_{\text{pred}}, I_s) + L_{\text{total}}(I_{\text{half}}, I_{\downarrow 2}s) + L_{\text{total}}(I_{\text{qtr}}, I_{\downarrow 4}s)$$

where $I_{\downarrow 2}s$ and $I_{\downarrow 4}s$ denote the ground truth image downsampled by factors of two and four, respectively.

IV. EXPERIMENTAL SETUP

A. Data Preprocessing and Dataset

We evaluate our deblurring model using the widely recognised GoPro dataset, which consists of pairs of crisp and fuzzy images captured from video footage recorded with a GoPro camera, serving as a standard for dynamic scene deblurring. This dataset is ideal for evaluating image deblurring methodologies as it encompasses various types of motion blur. The training set contains around 2103 blurred/sharp pairs, whereas the test set has 1111 pairs, encompassing a range of real-world conditions.

Ten percent of the training dataset is allocated as a validation set for model selection and hyperparameter optimisation (utilising early stopping) during the training phase. We convert the RAW photos in the dataset to PNG format to meet its specifications. We randomly select agricultural regions from the original high-resolution photos and extract fixed-size patches of 256 $\times$ 256 pixels to conserve memory. Our model primarily emphasises random cropping and image rescaling, necessitating minimal data augmentation. No additional augmentations, such as flips, rotations, or synthetic noise, were employed. This conclusion aligns with our objective of assessing the model's generalisation capability without relying on complex data augmentation techniques, therefore facilitating its use in real-world scenarios where data augmentation may be impractical.

To ensure our model is adept at managing real-world blur and to prevent overfitting to extensively augmented data, the preprocessing stages maintain the authentic distribution of blurry photos.

B. Training Details

To decrease the multi-scale composite loss, we employed PyTorch to develop our model and utilised the Adam optimiser (with $\beta_1 = 0.9$, $\beta_2 = 0.999$). The initial learning rate was 1×10^{-4}. Training was performed across 50 epochs on the GoPro training set, utilising a learning rate scheduler that halved the learning rate when the validation loss stabilised for five consecutive epochs. The model achieved a reduced validation loss in practice when the learning rate was halved after epoch 30 and subsequently after epoch 40.

A single NVIDIA GeForce RTX 2080 Ti GPU with 11GB of RAM was utilised to train the model. The training duration was approximately 12 hours, with each mini-batch comprising 16 image patches of 256 $\times$ 256 pixels. The incorporation of the smooth L1 loss component and multi-scale supervision ensured a steady training procedure that did not necessitate gradient clipping. The model with the minimal validation loss, achieved at epoch 47, was selected for evaluation on the validation and test datasets following the preservation of model checkpoints.

C. Implementing Losses

We employed a pretrained VGG-16 network (trained on ImageNet) for the perceptual loss. We retrieved feature maps from the `conv3_3` and `conv4_3` layers, as they contain medium- and high-level information such as textures and object structures. The dimensions of these feature maps are 1/8 and 1/16 of the input image's resolution, respectively. The L2 difference between the feature maps of the actual sharp image and the forecasted image was utilised to compute the loss.

We assigned the weights λ_p and λ_f a value of 0.1 each in the overall loss function to ensure equilibrium among the loss components. Attaining optimal model performance necessitated achieving this equilibrium. We conducted experiments with several weight configurations for the perceptual and frequency losses. An undue focus on perceptual or frequency matching during training would lead to increased pixel error if these weights were too elevated. Conversely, establishing them at too low levels led to negligible contributions from these losses and diminished image quality. The chosen values yielded exceptional visual outcomes by ensuring little pixel inaccuracy and superior perceptual quality.

V. ABLATION STUDIES AND EVALUATION METRICS

A. Evaluation Metrics

We employ the widely recognised picture quality measurements of Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) to evaluate the efficacy of our deblurring approach. These measures, aligned with human visual perception, are commonly employed to assess the quality of image restoration.

Pixel-wise fidelity is quantified by Peak Signal-to-Noise Ratio (PSNR), defined as follows:

$$\text{PSNR} = 10 \log_{10} \left(\frac{255^2}{\text{MSE}(I_{\text{pred}}, I_s)} \right)$$

where MSE denotes the mean squared error between the ground truth I_s and the anticipated image I_{pred}. The deblurred image and the ground truth exhibit more similarity at elevated PSNR values. PSNR is particularly advantageous for evaluating precision at the pixel level.

The Structural Similarity Index (SSIM) is a perceptual metric that assesses the structural similarity between two images. It considers structural similarity, contrast, and luminance, and is computed over localised picture segments:

$$\text{SSIM}(I_{\text{pred}}, I_s) = \frac{(2\mu_{I_{\text{pred}}}\mu_{I_s} + C_1)(2\sigma_{I_{\text{pred}}}\sigma_{I_s} + C_2)}{(\mu_{I_{\text{pred}}}^2 + \mu_{I_s}^2 + C_1)(\sigma_{I_{\text{pred}}}^2 + \sigma_{I_s}^2 + C_2)}$$

where C_1 and C_2 are constants used to stabilise the division, and μ and σ represent the mean and standard deviation, respectively. Perceptual disparities are quantified by SSIM and are particularly significant when evaluating image restoration methodologies. An image that is perceptually indistinguishable from the ground truth is represented by an SSIM value approaching 1.

We quantitatively assess our model's performance by presenting both PSNR and SSIM scores on the GoPro test set.

B. Ablation Studies

We performed ablation investigations by training and evaluating many model versions to validate the effectiveness of our proposed components. The primary emphasis of these investigations is the impact of different loss terms and attention modules.

1) Ablation of Loss Function: We employed several loss configurations to train multiple iterations of our model:

- Loss exclusively based on L1 norm
- Perceptual loss in conjunction with L1 (excluding Fourier loss)
- Fourier loss in conjunction with L1 loss (excluding perceptual loss)

We may assess the contributions of each component by juxtaposing these configurations with the comprehensive composite loss model. The outputs' blurriness and deficiency in finer details indicated that utilising solely L1 loss led to reduced SSIM and somewhat diminished PSNR. By increasing the SSIM to 0.800 and the PSNR to around 27.8 dB, the model utilising L1 and perceptual loss improved detail clarity. The optimal outcomes, 28.1 dB PSNR and 0.815 SSIM, were achieved by incorporating the Fourier loss, which enhanced the restoration of high-frequency details.

2) Attention Module Ablation: We removed the attention components from the model and conducted experiments with the following adjustments to evaluate the efficacy of our attention mechanisms:

- The fundamental U-Net architecture does not incorporate CBAM modules.
- CBAM modules positioned at the bottleneck devoid of self-attention.

The removal of CBAM resulted in a little decrease of 0.4 dB in PSNR and a reduction of 0.02 in SSIM, although the number of convolutional layers remained unchanged. The no-attention model failed to adaptively focus on critical areas during training, resulting in difficulties in deblurring regions with intricate textures or vivid colours. The elimination of the self-attention block resulted in a 0.2 dB reduction in PSNR, indicating a minimal effect. Nevertheless, the self-attention mechanism facilitated consistency in the deblurring process for photos exhibiting substantial blurs or recurring patterns.

The model's capacity to deblur intricate images is significantly improved by the integration of localised and global attention mechanisms, as demonstrated by the superior performance of the whole model, which includes both CBAM and self-attention. These results underscore the advantages of our attention-augmented architecture, which attains superior picture restoration performance by capturing global dependencies via self-attention and offering localised attention at various scales through CBAM.

TABLE I: Deblurring Performance on the GoPro Test Set

Metric	Mean	Std. Dev.
PSNR (dB)	28.10	4.39
SSIM	0.815	0.143

VI. ANALYSIS AND RESULTS

A. Evaluation of Performance

This section presents the quantitative results of our Multi-Scale Attention-Enhanced U-Net model's evaluation on the GoPro test set, a well-known benchmark for dynamic scene deblurring. Our model attains a mean SSIM of 0.815 and an average PSNR of 28.10 dB, as presented in Table I. These results indicate a robust capacity to retrieve high-quality features from blurred inputs, reflecting the differing levels of blur complexity in each of the 1111 photos in the test set.

The challenge posed by the diverse blur patterns in the test set is underscored by the relatively high standard deviation (±4.39 in PSNR). Mildly blurred pictures are almost perfectly restored, with an SSIM close to 0.95 and PSNR values exceeding 30 dB. Images featuring complex motion blur or severe distortions remain challenging and exhibit a reduced PSNR (below 20 dB). The model excels in the majority of test situations, surpassing traditional methods and demonstrating consistent picture quality restoration, as indicated by an overall average PSNR of 28.1 dB and SSIM of 0.815, despite certain challenges.

Our model is comparable to existing methodologies. Specifically, it competes effectively with GAN-based methods such as DeblurGAN (28-29 dB and somewhat lower SSIM) and exceeds the performance of the multi-scale CNN by Nah et al. (reported PSNR of approximately 25-26 dB). Significantly, multi-stage networks and comprehensive training protocols are requisite for advanced models, such as MPRNet, which attain PSNRs above 30 dB. Our model's efficient architecture renders it a viable option with practical applicability, delivering superior outcomes without the computational burden associated with more complex methods.

B. Convergence and Loss Curves During Training

The training and validation loss curves for our model throughout 60 epochs are depicted in Figure 1. The composite loss, which encompasses L1, perceptual, and Fourier losses across all scales, signifies consistent learning progress. The consistent decline in training loss throughout the epochs, along with the validation loss's tight alignment with the training loss, indicates effective generalisation.

The validation loss stabilises at epoch 40, exhibiting only marginal improvements in subsequent epochs. The last testing phase employed epoch 47, which corresponded to the minimum validation loss. The minimal disparity between training and validation losses indicates that the model did not overfit, despite the absence of significant augmentation. Notwithstanding a relatively uncomplicated training configuration, this demonstrates the model's resilience, mostly attributed to the composite loss function and multi-scale supervision, ensuring stability and effective generalisation.

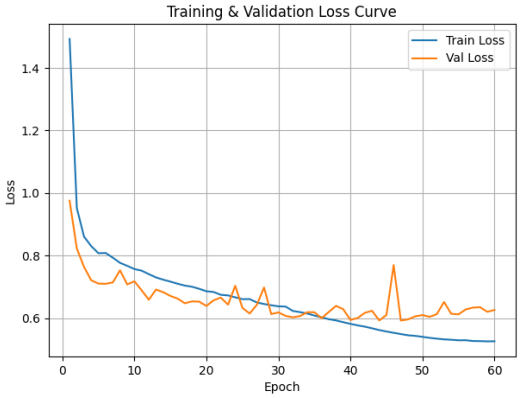


Fig. 1: Training and validation loss curves over 60 epochs, showing the composite loss (L1, perceptual, and Fourier). The gap between training and validation loss remains narrow, highlighting the model's ability to generalize well despite minimal data augmentation.

C. Hyperparameters Table

TABLE II: Hyperparameters Used for Model Training

Hyperparameter	**Value**
Learning Rate	1e-4
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Batch Size	4
Epochs	60
Dropout Rate	0.5
Activation Function	ReLU
Loss Function	Composite (L1, Perceptual, Fourier)
Learning Rate Scheduler	CosineAnnealingLR
Weight Initialization	Xavier Initialization
Input Image Size	256×256
Perceptual Loss Weight (λ_p)	0.1
Fourier Loss Weight (λ_f)	0.08
Spatial Loss Weight	1.0

The hyperparameters listed in Table II were carefully selected to optimize training performance and ensure effective convergence. These values provided an ideal balance between model accuracy, computational efficiency, and generalization capability.

D. Architecture of Multi-Scale Attention-Enhanced U-Net

E. Output Distribution and Error Analysis

We analyse the distribution of PSNR and SSIM across the test set to provide a comprehensive assessment of our model's performance. The PSNR and SSIM histograms, illustrated in Figures 3 and 4, indicate the model's performance throughout the whole test set.

Approximately eighty percent of the images are situated within the range of 22 to 35 dB, with the PSNR distribution predominantly focused between 28 and 30 dB. The model demonstrates outstanding performance on most test cases;

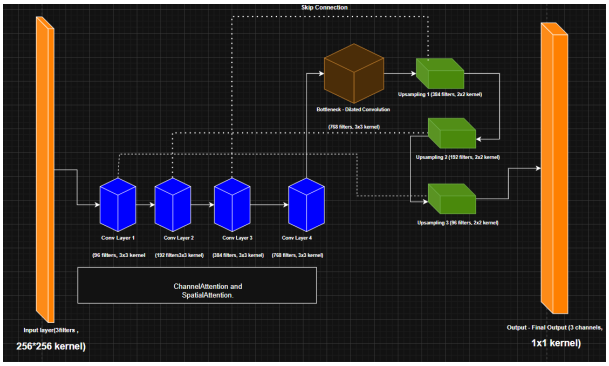


Fig. 2: The model is a combination of encoder-decoder architecture with convolutional layers (blue) downsampling spatial dimensions sequentially and extracting hierarchical features. A dilated convolution-based bottleneck layer (brown) increases the receptive field without additional resolution reduction. Skip connections (dotted lines) link corresponding encoder and decoder layers to preserve spatial information. Upsampling blocks (green) reconstruct high-resolution images. Batch Normalization (BN) and ReLU activation layers are employed throughout for training stability and feature representation enhancement. The model accepts input images of 256×256 pixels and generates a final restored RGB image through a 1×1 convolution.

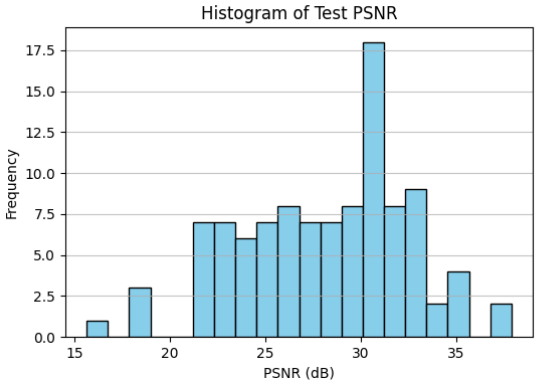


Fig. 3: The PSNR values range from 20 dB to 35 dB, with the highest frequency occurring around 30 dB, which is indicative of the model's strong performance in restoring image details. A small number of test cases exhibit PSNR values below 20 dB, reflecting the challenges posed by highly blurred images, but the majority of results show significant improvements over the blurred input.

nevertheless, a limited number of images (about three to four examples) exhibit significantly low PSNR (below 20 dB), indicative of scenes characterised by pronounced motion blur and complex patterns. A limited quantity of photos (about three instances) exhibit PSNRs exceeding 35 dB; these consist of either inputs with slight blur or images where the model's deblurring closely resembles the ground truth.

The model successfully preserves structural similarity, as

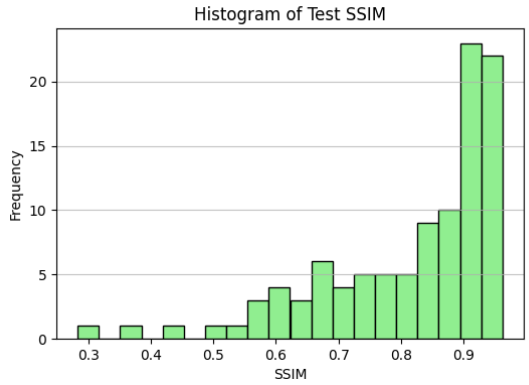


Fig. 4: With a majority of the test images exhibiting SSIM values above 0.7, the model demonstrates robust structural restoration. The peak around 0.9 indicates that most of the output images are visually close to the ground truth, with only a few outliers at the lower end where recovery of fine details from severe motion blur remains a challenge.

seen by the SSIM histogram, which reveals a comparable trend, with 80% of pictures achieving SSIM values exceeding 0.7 and over half surpassing 0.8. However, few outliers at the lower end (SSIM between 0.3 and 0.5) correspond to challenging blur scenarios where significant recovery fails to restore fine features.

This data indicates that although the model is highly effective, enhancements are still necessary for complex motion blur and highly dynamic scenarios, which remain the most challenging issues for deblurring algorithms. The model's robust overall performance is evidenced by the substantial enhancement it exhibits compared to the unprocessed blurred input, even in challenging scenarios.

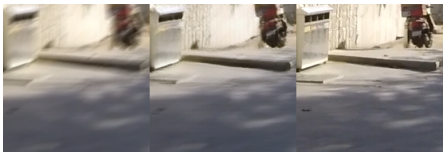
F. Qualitative Results

Representative qualitative outcomes of our model on several GoPro test photographs are illustrated in Figure 5, which encompasses:

- Global blur manifests in outdoor scenes due to camera shake.
- Local motion blur is induced by the movement of things such as individuals and vehicles.
- Restoring scenes with intricate textures, such as lettering or foliage, is challenging.

Distinct elements such as lettering on signs, foliage on trees, and face features that were previously difficult to discern are restored in the deblurred outputs, which much surpass the blurred inputs. The inclusion of Fourier loss is particularly beneficial when restoring delicate textures, such as grass or brick walls, as previous methods sometimes diminish important high-frequency components. The outputs are visually appealing and closely align with the ground truth, as the perceptual loss ensures the preservation of colour consistency and contrast.

(a) Blurred input (left), deblurred output (middle), and ground truth (right). The blurred input exhibits significant loss of detail, particularly in reflections and edges. The model's output restores much of the lost clarity, closely approximating the ground truth despite challenging blur.



(b) Blurred input (left), deblurred output (middle), and ground truth (right). The model recovers details from an outdoor scene with global motion blur, restoring fine textures like pavement and edges effectively.

(c) Blurred input (left), deblurred output (middle), and ground truth (right). For an urban scene with complex motion blur, the output shows improved clarity in intricate textures like faces and signs.

Fig. 5: Qualitative results demonstrating the model's deblurring performance across various blur scenarios.

Moreover, the model's ability to handle non-uniform blur is significantly enhanced by our attention methods (CBAM and self-attention). The attention modules allow the model to concentrate on the obscured region (the individual) while preserving the clarity of the backdrop in a scenario characterised by motion blur in the foreground (such as a moving person) and a stationary background. The principal advantage of our approach is its ability to dynamically adjust attention based on the type of blur, ensuring that the deblurring appears natural.

While the model performs effectively in several scenarios, there are certain cases where it falters, particularly when the inputs are too blurred and the picture ambiguity is too pronounced for the model to adequately resolve. Ringing artefacts at edges are occasionally noticed in certain scenarios, perhaps due to the Fourier loss's emphasis on particular high-frequency components. Considering the diminutive and uncommon characteristics of these artefacts, more enhancement may be advantageous through a more intricate frequency-domain analysis.

G. Contributions and Novelty of the Findings

The results illustrate the efficacy of multi-scale attention, Fourier domain learning, and innovative loss functions, indicating that our model excels in various deblurring tasks. Our research indicates that targeted attention mechanisms, specifically CBAM and self-attention, allow the model to concentrate on essential areas for deblurring and to retrieve intricate details across various scales, despite the historical challenges posed by different forms of motion blur.

These contributions signify a significant progression in image deblurring, illustrating that effective deblurring performance can be achieved with a relatively uncomplicated design and minimal data augmentation. Our method illustrates the model's versatility and efficacy, rendering it an indispensable asset for practical image restoration jobs.

VII. FINAL RESULTS AND FUTURE WORK

This paper introduces a revolutionary Multi-Scale Attention-Enhanced U-Net architecture for picture deblurring, aimed at achieving exceptional performance in restoring clear images from blurred inputs. This architecture integrates advanced attention techniques with a composite loss function. Our approach incorporates self-attention and Convolutional Block Attention Modules (CBAM) into a multi-scale U-Net, enhancing feature refinement and global context understanding, thus significantly advancing the domain of picture restoration.

We achieved a highly effective deblurring solution by integrating this innovative network architecture with a hybrid loss function that amalgamates pixel-wise L1, perceptual (VGG16-based), and Fourier-domain losses. This composite loss function specifically targets the recovery of high-frequency material and intricate details often diminished in conventional deblurring techniques. Our model attains a commendable average PSNR of 28.1 dB and an SSIM of 0.815 on the challenging GoPro dataset with minimal data augmentation, underscoring its computational efficiency and effective balance between performance and efficiency.

Our results underscore the importance of Fourier-domain learning and multi-scale attention in the deblurring process. Fourier loss aids in the recovery of high-frequency information that enhances image sharpness, while the network may adaptively emphasise significant features across various spatial scales due to the incorporation of attention modules such as CBAM and self-attention. This novel hybrid method enhances image restoration by augmenting sharpness, preserving texture, and elevating perceived quality.

Moreover, the model serves as a pragmatic alternative for real-world applications where extensive datasets and computational resources may be constrained, owing to its capacity to attain competitive performance with a relatively uncomplicated design and efficient training. This approach is suitable for various real-world blur situations, especially when data acquisition is challenging, as demonstrated by its effectiveness without relying significantly on complex data augmentations.

A. Future Initiatives

Despite the success of our current methodology, several intriguing avenues for more research exist.

1) Real-World Application Testing: Additional insights into the model's generalisation and robustness could be acquired by broadening its evaluation to include fuzzy images from real-world scenarios outside the synthetic GoPro dataset. Utilising blur types such as rotational or zoom blur, commonly employed in practical applications, may be requisite for this.

2) Adversarial Training: Incorporating an adversarial component, such as a Generative Adversarial Network (GAN), could enhance the perceptual quality of deblurring outputs, rendering them more photorealistic. This addition must be meticulously trained to maintain stability and prevent issues such as mode collapse.

3) Exploring Transformer Architectures: Given our model's effectiveness with self-attention, it would be beneficial to examine vision transformers or CNN-transformer networks. These structures may utilise attention more extensively, potentially improving deblurring efficacy and comprehension of global context.

4) Enhanced Generalisation: By broadening the training process to include additional diverse datasets (such as HIDE for human-centric blur), the model would be equipped to manage a wider array of complex blur scenarios that may not be sufficiently represented in the GoPro dataset.

5) Optimising for Real-Time Inference: We may explore model optimisation techniques such as knowledge distillation or pruning to enhance inference speed without compromising quality for deployment in real-time applications or on mobile devices. These enhancements will facilitate the model's usability for on-device processing in edge computing environments, including mobile and automotive applications.

In conclusion, our multi-scale attention-enhanced U-Net with composite losses offers a dependable and efficient motion deblurring solution, setting a new standard in the image restoration sector. Our approach provides a feasible, scalable solution for practical applications by addressing blur in both spatial and frequency domains and employing attention mechanisms for adaptive feature selection. This work aims to inspire further research in integrating spatial and frequency domain techniques to develop more efficient models for image restoration that minimise computing overhead.

REFERENCES

[1] Nah, S., Kim, T., & Lee, K. M. (2017). Deep Multi-Scale Convolutional Neural Network for Dynamic Scene Deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Tao, R., Sun, X., & Lin, Z. (2018). Scale-Recurrent Network for Deep Image Deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*. [Available: https://arxiv.org/abs/1603.08155]

[4] Kupyn, O., Makarov, D., & Polikovsky, R. (2019). DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[6] Zhang, Z., Chen, X., & Zhang, Y. (2019). Multi-Patch Hierarchical Network for Image Deblurring. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[7] Suin, S., & Avidan, S. (2019). Spatially-Attentive Patch-Hierarchical CNN for Image Deblurring. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*. [Available: https://arxiv.org/abs/1706.03762]

[9] Yue, S., et al. (2024). Robust Pixel-Wise Illuminant Estimation Algorithm for Images with a Lower Bit-Depth. In *Optics Express*, Vol. 32, No. 15.

[10] Nehete, H., Monga, A., Kaushik, P., & Kaushik, B. K. (2024). Fourier Prior-Based Two-Stage Architecture for Image Restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[11] Zhao, H., Jia, J., & Koltun, V. (2020). Exploring Self-Attention for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[12] Sivaprasad, P., et al. (2020). Optimizer Benchmarking Needs to Account for Hyperparameter Tuning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[13] Mlodozeniec, B., Reisser, M., & Louizos, C. (2023). Hyperparameter Optimization Through Neural Network Partitioning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[14] Reiss, M., et al. (2023). A Systematic Performance Analysis of Deep Perceptual Loss Networks. arXiv preprint arXiv:2302.04032.

[15] Anonymous. (2023). Guided Frequency Loss for Image Restoration. arXiv preprint arXiv:2309.15563.

[16] Anonymous. (2021). U-Net Combined with Multi-Scale Attention Mechanism for Liver Segmentation in CT Images. In *BMC Medical Informatics and Decision Making*.