Deep Supervised Attention Network for Dynamic Scene Deblurring

Seok-Woo Jang , Limin Yan , Gye-Young Kim *

Posted Date: 17 February 2025

doi: 10.20944/preprints202502.1206.v1

Keywords: dynamic deblurring; multiple loss function; multi-scale network; supervised attention; recurrent network; feature mapping

 doi:10.20944/preprints202502.1206.v1

Article

Deep Supervised Attention Network for Dynamic Scene Deblurring

Seok-Woo Jang [1], Limin Yan [2] and Gye-Young Kim [2,*]

[1] Department of Software, Anyang University, 22, 37-Beongil, Samdeok-Ro, Manan-Gu, Anyang 14028, Republic of Korea

[2] School of Software, Soongsil University, 369, Sangdo-Ro, Dongjak-Gu, Seoul 06978, Republic of Korea

* Correspondence: gykim11@ssu.ac.kr

Abstract: In this study, we propose a dynamic scene deblurring approach using a deep supervised attention network. While existing deep learning-based deblurring methods have significantly outperformed traditional techniques, several challenges remain: (1) Invariant weights: Small conventional neural network (CNN) models struggle to address the spatially variant nature of dynamic scene deblurring, making it difficult to capture the necessary information. A more effective architecture is needed to better extract valuable features; (2) Limitations of standard datasets: Current datasets often suffer from low data volume, unclear ground truth (GT) images, and a single blur scale, which hinders performance. To address these challenges, we propose a multi-scale, end-to-end recurrent network that utilizes supervised attention to recover sharp images. The supervised attention mechanism focuses the model on features most relevant to ambiguous information as data are passed between networks at difference scales. Additionally, we introduce new loss functions to overcome the limitations of the peak signal-to-noise ratio (PSNR) estimation metric. By incorporating a fast Fourier transform (FFT), our method maps features into frequency space, aiding in the recovery of lost high-frequency details. Experimental results demonstrate that our model outperforms previous methods in both quantitative and qualitative evaluations, producing higher-quality deblurring results.

Keywords: dynamic deblurring; multiple loss function; multi-scale network; supervised attention; recurrent network; feature mapping

1. Introduction

Traditional image deblurring algorithms reconstruct images based on a specific image model, and the image degradation process can be expressed as follows:

$$B = K * S + n, \tag{1}$$

where B, S, n, and K represent the blurred image, latent sharp image, noise, and unknown blur kernel, respectively. To reconstruct the latent image S, a deblurring algorithm must accurately estimate the blur kernel K and then perform a deconvolution operation on the blurred image using K to recover the sharp image. However, difference combinations of sharp images and blur kernels can produce the same blurred image after convolution, making the deblurring problem inherently ill-posed.

Kim et al. [1] introduced the concept of dynamic scene deblurring, highlighting that blur is caused by various factors, such as camera shake and object motion, leading to non-uniform blur in dynamic scenes. Despite this, traditional methods that rely on prior knowledge to estimate the blur kernel often overlook the non-uniform nature of blur, making accurate kernel estimation unrealistic and prone to artifacts. Additionally, most traditional image restoration methods based on prior knowledge use iterative optimization, which involves tuning a large number of parameters and significantly increases computational overhead. Consequently, the performance of traditional

 doi:10.20944/preprints202502.1206.v1

deblurring methods could be improved. With the advent of deep learning, dynamic image deblurring has made significant strides in both performance and efficiency. Deep learning-based methods use multiple image pairs to create mapping functions from distorted to sharp images, eliminating the need to estimate complex prior information and reducing errors associated with traditional methods. Sun [2] and Schuler [3] were the first to introduce convolutional neural networks (CNNs) for image deblurring, where CNNs still estimate the blur kernel. Although CNN-based approaches can improve deblurring performance by more accurately predicting the blur kernel, they remain limited to specific types of blur and struggle with spatially varying blurs. To address these limitations, Nah et al. [4] proposed a multi-scale CNN-based image restoration approach that directly recovers latent images without assuming a specific blur kernel model, thereby avoiding artifacts caused by kernel estimation errors. However, their method does not account for the temporal information of blurred images, which is a critical aspect given that blur features span both temporal and spatial dimensions.

Building on the work of Nah et al., Zhang et al. [4,5] introduced an RNN structure to capture temporal information in images. To address the issue of parameter redundancy in multi-scale networks, Xin et al. [6] proposed a multi-scale recurrent network with parameter sharing, significantly reducing network complexity. Subsequently, Kupyn et al. [7] utilized a generative adversarial network (GAN) to input blurred images into the generator and output sharp images, with a discriminator supervising the quality of the generated images. To leverage multi-scale features, they incorporated a feature pyramid network in the generator, combining feature information across different scales to better handle image blurring. Kuldeep et al. [8] further enhanced deblurring by introducing self-attention mechanisms to capture non-local spatial dependencies between features, improving the network's ability to manage spatial variations. However, they overlooked the significant computational overhead introduced by self-attention.

To address these challenges in image restoration and feature extraction, we propose a novel network, i.e., deep supervised attention network (DSANet), for dynamic deburring. DSANet is a multi-scale, coarse-to-fine network that embeds a new supervised attention module specifically designed for dynamic deblurring. The main contributions of this study include:

1. An optimized ConvLSTM-based encoder-decoder structure that accelerates network integration and enhances the learning of the spatial-temporal features.
2. A newly proposed supervised attention module that mitigates the high computational overhead of self-attention by guiding the model to focus on features most relevant to blur information when transferring features between different network scales.
3. The introduction of a multi-loss function based on the fast Fourier transform (FFT), enabling the model to learn deblurring features in the frequency domain.
4. The development of a new dataset collected in diverse environments, which outperforms existing datasets in several aspects, reducing the challenges posed by discrepancies between simulated data and real-world images.

2. Related Works

2.1. Encoder-decoder structure

Image segmentation has seen significant advancements, largely due to the adoption of encoder-decoder architectures. Recently, these systems have been widely used in various computer vision tasks [9,10]. As illustrated in Figure 1, an encoder-decoder network in computer vision refers to a symmetric architecture built with CNNs. It typically consists of convolutional layers, pooling layers, and batch normalization layers. During the encoding stage, the input data are gradually transformed into feature maps with smaller spatial dimensions and an increased number of channels. In the decoding stage, these compressed feature maps are converted back to the input format through deconvolution or up-sampling operations [11,12]. The encoder-decoder architecture is optimized for faster network convergence and improved gradient propagation. However, its application in image restoration has been limited, as batch normalization is highly sensitive to batch size.

 doi:10.20944/preprints202502.1206.v1

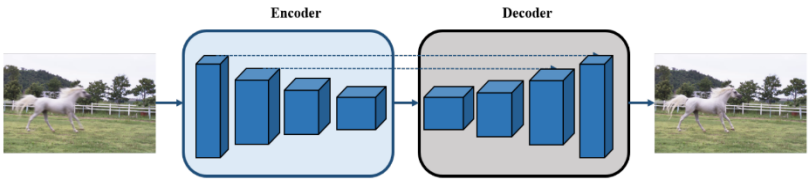

Figure 1. Basic encoder-decoder architecture.

2.2. Multi-scale network

In many computer vision tasks, various forms of coarse-to-fine or multi-scale architectures are commonly employed [13,14]. The core idea behind multi-scale networks is that images at different scales provide varying degrees of feature information. Smaller-scale images are well-suited for capturing global semantic information, which can then guide the analysis of larger-scale images with broader receptive fields and more comprehensive global information. However, the requirements for multi-scale structures can vary across different computer vision tasks. While existing multi-scale structures can address the receptive field needs in deblurring tasks, the use of numerous convolutional layers based on residual blocks to capture long-range dependencies significantly increases the model's complexity.

2.3. Self-attention

Attention mechanisms [15] were initially introduced to enhance the translation performance of neural machine translation (NMT) systems and have recently gained prominence in computer vision. A typical example is CNN. In CNNs, each convolutional layer focuses on a localized region defined by the kernel size. Although the receptive field expands over time, it primarily captures global feature correlations. Researchers have explored integrating attention mechanisms with computer vision, often using masks to enhance the attention process. These masks help identify salient features in an image by applying newly trained weights to each image, enabling deep neural networks to focus on areas that require attention. Wang et al. [16] introduced a combined attention module within an encoder-decode framework, which refines the feature map and improves network performance. However, the need to calculate a 3D focus map results in significant computational overhead. Woo et al. [17] proposed the convolutional block attention module (CBAM), a hybrid attention module that extracts meaningful features across two channel dimensions, enhancing adaptive feature learning. This module can be seamlessly integrated into other networks, but attention mechanisms still incur high computational costs. Hu et al. [18] developed the squeeze-and-excitation network (SENet), which computes global image information at the feature channel level. Tsai et al. [19] introduced the blur-aware attention network (BANet), which achieves accurate and efficient deblurring with a single forward pass. BANet leverages region-based self-attention with multi-kernel strip pooling to address blur patterns of varying magnitudes and orientations and use cascaded parallel dilated convolutions to aggregate multi-scale content features.

2.4. Residual block

He et al. [20] introduced the residual block, a highly successful deep-learning architecture that enables the training of very deep networks and addresses the issue of vanishing gradients. While the problems of vanishing and exploding gradients have been substantially mitigated by batch normalization and suitable activation functions [21], residual blocks play a crucial role in preventing network degradation. However, traditional residual networks are not ideal for image restoration tasks. To enhance image restoration quality, Wang et al. [22] removed the normalization layer from all residual blocks. In contrast, Nah et al. [4] demonstrated that incorporating batch normalization can improve the performance of deblurring networks. Unlike traditional residual networks, the residual module used in their approach does not include a batch normalization layer.

 doi:10.20944/preprints202502.1206.v1

3. Deep Supervised Attention Network (DSANet)

Figure 2 illustrates the overall structure of the proposed DSANet. The network employs a multi-scale recurrent architecture comprising an encoder and a decoder. The encoder extracts features, while the decoder restores the image. A skip connection between the encoder and decoder enhances the receptive field and accelerates the model convergence.

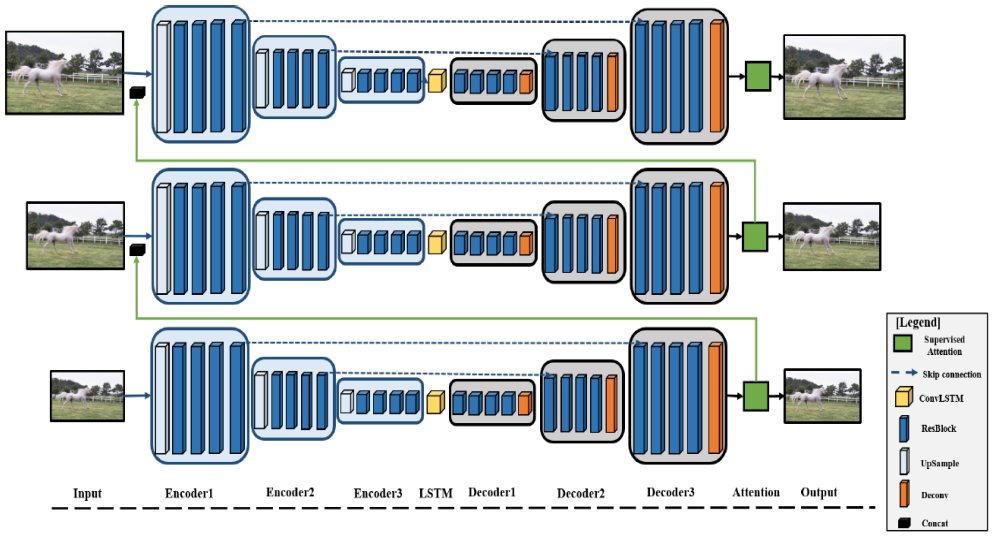

Figure 2. Proposed encoder-decoder architecture.

In addition, the network incorporates a blurred attention block that selectively ignores irrelevant features, focusing instead on connectivity features of varying magnitudes to suppress redundant information. Equation (2) shows how the network processes a blurred image to estimate a sharp image.

$$R^i, h^i = \mathrm{Net}_{\mathrm{DSANet}}(B^i, R^{(i+1)'}, h^{(i+1)'}; \theta_{\mathrm{DSANet}}), \tag{2}$$

where i represents the scale index, with i = 1 indicating the smallest scale; R^i and h^i are the blurred image and estimated latent image at the i-th scale, respectively; $\mathrm{Net}_{\mathrm{DSANet}}$ is a multi-scale recurrent supervised attention network with training parameters denoted as θ_{DSANet}; the hidden state features, h_i, flow across scales and captures both image structure information and blur information; and $(i+1)'$ refers to the operation of the adjustment of image or feature size.

When the blurred image is down-sampled to different scales, the sampling coefficient between adjacent scales is 1/2. The process begins by loading the smallest-scale image R^3 into the Encoder1. A down-sampling layer with a stride of 2 reduces the size of the feature map to half its original dimensions and increases the number of channels from 3 to 32, (i.e., [H×W×3] → [H/2×W/2×32]), where H and W represent the height and width of the feature map, respectively. The number three indicates that the original image has three channels. Four residual blocks, without additional down- or up-sampling, are used to enhance the receptive field of the network and improve image restoration.

The feature map is then passed to Encoder2, which also employs a down-sampling layer with a stride of 2 to reduce the size of the feature map to half of its original size and increase the number of channels from 32 to 64, (i.e., [H/2×W/2×32] → [H/4×W/4×64]). Four residual blocks, without down- or up-sampling, are used to further expand the receptive field and enhance the image restoration effect.

Next, Encoder3 uses a down-sampling layer with a stride of 2 to reduce the feature map size by half and increase the number of channels from 64 to 128, (i.e., [H/4×W/4×64] → [H/8×W/8 ×128]). Four residual blocks, without down- or up-sampling, are again employed to increase the receptive

 doi:10.20944/preprints202502.1206.v1

field and improve image restoration. Subsequently, the ConvLSTM module is introduced to enhance the model's ability to capture spatial-temporal dependencies.

In the Decoder stage, the feature map is input by Decoder1. This stage uses four residual blocks, without down- or up-sampling, to maintain the receptive field and improve image restoration. An up-sampling layer with a stride of 2 is applied to double the size of the feature map and reduce the number of channels from 128 to 64, (i.e., [H/8×W/8×128] → [H/4×W/4×64]).

Decoder2 employs four residual blocks without additional down- or up-sampling to expand the network's receptive field and enhance image restoration. An up-sampling layer with a stride of 2 is then used to double the size of the feature and reduce the number of channels from 64 to 32, (i.e., [H/4×W/4×64] → [H/2×W/2×32].

Decoder3 also uses four residual blocks without down- or up-sampling to further increase the receptive field and improve image restoration. Another up-sampling layer with a stride of 2 doubles the size of the feature map while keeping the number of channels unchanged, (i.e., [H/2×W/2×32] → [H×W×32]). Finally, in the supervised attention layer, the image is restored to its original image scale [H×W×3]).

This process can be expressed as follows:

$$R^3, h^3 = \mathrm{Net}_{\mathrm{DSANet}}(B^3, \theta_{\mathrm{DSANet}}), \tag{3}$$

where R^3 and h^3 represent the smallest scale of the restored image and the learned hidden state features, respectively. The supervised attention layer selectively refines h^3 and then concatenates it with the up-sampled features from the previous scale. Similarly, the processing at the next-scale is defined as:

$$R^2, h^2 = \mathrm{Net}_{\mathrm{DSANet}}(B^2, R^{(3)'}, h^{(3)'}; \theta_{\mathrm{DSANet}}), \tag{4}$$

where R^2 and h^2 denote the restored image and learned hidden state features at the second scale, respectively, while $R^{3'}$ and $h^{3'}$ represent the up-sampled features from the smallest scale. After processing through the final scale network, the output, which combines the features from the previous scales, is given by:

$$R^1, h^1 = \mathrm{Net}_{\mathrm{DSANet}}(B^1, R^{(2)'}, h^{(2)'}; \theta_{\mathrm{DSANet}}) \tag{5}$$

The proposed network addresses the limitations of previous deblurring networks by incorporating a supervised attention module to achieve enhanced feature extraction capabilities.

3.1. Supervised attenuation module

Recent work in super-resolution has implemented self-attention mechanisms [23], which, while enhancing feature extraction capabilities, have also introduced significant computational overhead.

To improve the restoration network's ability to perceive blurred features with spatial variations while mitigating computational costs, we propose a blur module based on supervised attention. This module utilizes a supervised attention mechanism to dynamically capture blurred features with spatial variations, making it easier to retrieve a clear latent image. Figure 3 illustrates the supervised attention module.

 doi:10.20944/preprints202502.1206.v1

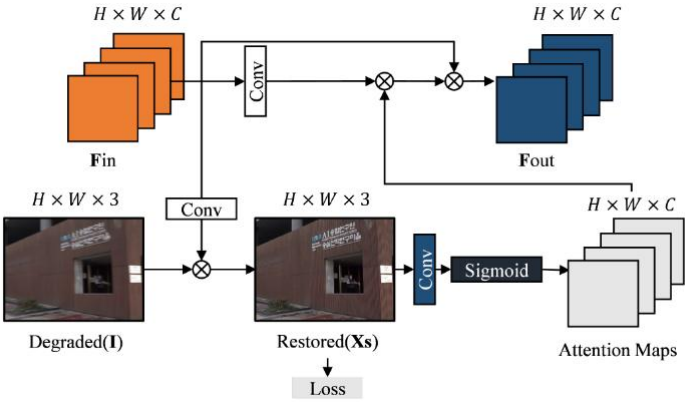

Figure 3. Supervised attention module.

In the final stage of the multi-scale network, the degraded image I is restored using the learned feature map F_i^n, resulting in the restored image Restored(x_s). The supervised attention module generates an attention feature map using the Sigmoid activation function applied to the recovered image Restored(x_s). The attention feature map's values ranges from 0 to 1, where higher values indicate greater attention. This attention feature map guides the fusion of features from different scale networks. Since blurring varies spatially, the module's significance lies in applying different levels of attention to various positions in the blurred image. It follows a selective memory mechanism similar to the forgotten gate in LSTM [24]. Before concatenating features from different scales, unimportant features are selectively forgotten, important features are emphasized, and redundant information is suppressed.

3.2. Multi-loss function

The rapid advancement in detection and segmentation is largely attributed to effective evaluation metrics. However, a similar robust metric is still needed for low-level vision tasks. Peak signal-to-noise ratio (PSNR) is the primary evaluation metric used in image deblurring. It measures the content loss of an image, with higher PSNR values indicating lower content loss. Despite its widespread use, PSNR has limitations in accurately assessing image quality. For instance, as shown in Figure 4, an image with the highest PSNR value may not necessarily have the best perceptual quality.

Figure 4. From left to right: degraded image, SRResNet, SRGAN, and GT.

The super-resolution generative adversarial network (SRGAN) [25] addresses this by optimizing loss functions that are more sensitive to human perception. Even if the PSNR value is not exceptionally high, SRGAN aims to improve the perceptual quality of the restored image. Johnson et al. [26] proposed a multi-loss function based on perceptual loss for style transfer tasks, which improves content representation at the expense of reduced PSNR precision. Later, Jiao et al. [27] and

 doi:10.20944/preprints202502.1206.v1

Ignatov et al. [28] introduced auxiliary loss functions to enhance image quality. However, these approaches did not significantly improve perceptual quality compared to single loss function models and often resulted in lower PSNR values.

As with most multi-scale deblurring networks, we initially experimented with a multi-scale single loss function, such as mean squared error (MSE). However, a single loss function alone did not improve the subjective visual quality of the restored images. Therefore, we developed a new multi-loss function designed to enhance the subjective evaluation of image quality while maintaining or improving the PSNR value. The multi-loss function is defined as follows:

$$Loss_{total} = MSE + \mu \times FFT_{MSE}, \tag{6}$$

where FFT_{MSE} uses the Fourier transform to convert the image signal to the frequency domain, and μ represents the loss weight. We conducted numerous experiments with different loss weights, such as 0.1, 0.4, and 0.6. The results showed that a loss weight of $\mu = 0.1$ provided the best performance.

The FFT_{MSE} loss is defined as:

$$FFT_{MSE} = \frac{1}{M \times N} \sum_{0 \le i \le N} \sum_{0 \le j \le M} \left(FFT(f_{ij}) - FFT(f_{ij}^{'}) \right)^2, \tag{7}$$

where M and N denote the length and width of the image, respectively; f_{ij} represents the pixel value of the original image, and $f_{ij'}$ is the pixel value of the target image at position (i, j). $FFT(\cdot)$ is the fast Fourier transform function. A smaller FFT_{MSE} value indicates a higher quality of the restored image, while a larger value suggests significant distortion and poor quality. By applying the FFT, we can observe features in the frequency domain that are otherwise undetectable. The resulting spectrogram reveals the frequency components of the image, where high-frequency signals correspond to edges and noise, and low-frequency signals correspond to background. In the frequency domain, we can effectively manipulate high- and low-frequency information for tasks such as image denoising, enhancement, and edge extraction.

3.3. A new dataset

The quality of a dataset significantly impacts the advancement of image restoration and the broader field of computer vision research, directly influencing network performance. However, a perfect dataset that includes both degraded real-world images and corresponding ground-truth images has not yet been proposed, due to the difficulty in capturing both simultaneously. Most existing datasets rely on synthetic methods [29,30], which often differ significantly from real-world degraded images. To address this issue, Lu et al. [31] and Nimisha et al. [32] explored unsupervised learning methods to mitigate performance problems associated with synthetic datasets. However, their approaches were limited to specific areas such as face and text deblurring. Sun et al. [2] introduced a synthetic dataset comprising 80 natural images and eight blur kernels.

Lai et al. [33] developed a dataset of 100 blurred images collected from real-world scenarios to evaluate deblurring methods. Despite this, none of the available datasets fully address the gap between synthetic and real datasets. Recently, Rim et al. [34] introduced the Realblur dataset, which includes pairs of degraded real-world images and ground truth (GT) images captured simultaneously using an image acquisition system. This system, which uses two cameras to capture blurry and sharp images at different shutter speeds, reduces the discrepancy between synthetic and real datasets. However, there is often a significant offset between the sharp and blurred images, leading to structural differences that make such image pairs unsuitable for training. To address this, Rim et al. [34] suggested photographing only static objects and applying several post-processing steps, including photometric and geometric alignment. Nevertheless, research on deblurring has frequently been conducted in dynamic environments, and a static-only dataset is insufficient for improving deblurring network performance. Additionally, extensive post-processing introduces inefficiencies and high costs. Zhang et al. [35] recently proposed a deblurring network that combines two generative adversarial models [36]: one to learn to blur real images and another to learn to deblur

 doi:10.20944/preprints202502.1206.v1

synthetic images. This approach helps mitigate the complexities associated with synthetic and real dataset differences but faces limitations due to the lack of large-scale datasets.

We reviewed the limitations of current synthetic datasets and identified various issues contributing to the gap between synthetic and real datasets. To address these challenges, we propose a new synthetic-based dataset designed to significantly reduce the complexities associated with synthetic data. The new dataset is collected using the method illustrated in Figure 5.

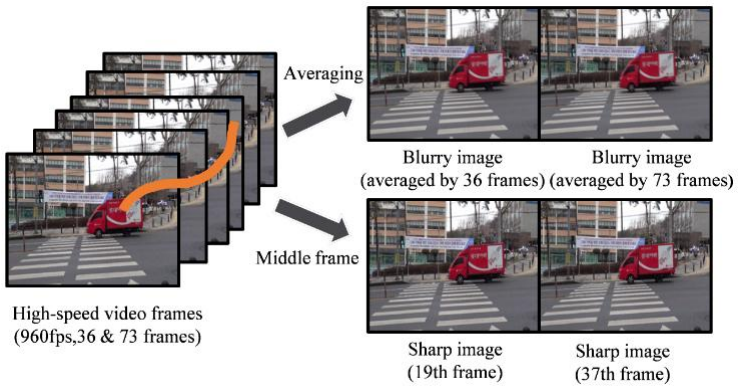

Figure 5. New dataset collection.

4. Experimental Results

The quality of the dataset is crucial in image restoration tasks. Generating a dataset for image blur is particularly challenging due to the presence of both uniform and non-uniform blurs, with the latter also known as dynamic blurs. Traditional methods for creating such datasets involve using various blurring kernels to synthesize both uniform and non-uniform blurred images from sharp images. However, this approach often oversimplifies the imaging model, leading to significant discrepancies between synthesized data and real-world images.

Recently, a method involving the averaging of consecutive frames has been proposed to generate blurred images more effectively. Based on this approach, Nah et al. [4] created the GOPRO dataset, which consists of 3214 pairs of sharp and blurred images. Of these, 2103 pairs are used for training and model development, while the remaining 1111 pairs are reserved for model evaluation. We evaluate the performance of our proposed network by comparing its performance with those of state-of-the-art methods using the GOPRO dataset. Additionally, we assess the network's performance both with the GOPRO dataset alone and with the inclusion of our dataset. Our results demonstrate that the new dataset significantly enhances deblurring performance.

Figure 6. GOPRO dataset pairs of sharp and blur images.

The experiments were implemented using PyTorch and trained on two NVIDIA Titan V GPUs. For each training iteration, randomly cropped 256×256-pixel patches were used as input. All training variables were initialized using Kaiming initialization [37]. The Adam optimizer [38] was employed

 doi:10.20944/preprints202502.1206.v1

with an initial learning rate of 2×10^{-4}, which gradually decreases to 1×10^{-6}. The models were trained for 4000 epochs using the cosine annealing strategy.

To evaluate the performance of the proposed model in dynamic deblurring tasks, we compared our network's PSNR, SSIM, parameter size, and single-image processing time with those of MSCNN [5], RNNDeblur [1], SRN [6], and DMPHN [39]. For a fair comparison, we assessed the quantitative performance of DSANet2 with data augmentation and DSANet1 without data augmentation. Additionally, we evaluated DSANet++ with the new dataset to demonstrate its validity. The quantitative results are summarized in Table 1. Even DSANet1, which does not use data augmentation, outperformed DMPHN in terms of both metric scores and processing speed. The inclusion of data augmentation improved both PSNR and SSIM. Furthermore, DSANet++ achieved even better performance with the introduction of the new dataset. These experimental results indicate that the new dataset significantly enhances both quantitative and qualitative performance.

Table 1. Evaluation results on the benchmark GOPRO testing set.

Model	PSNR	SSIM	Time (ms)	Size (MB)
MSCNN	29.23	0.9162	4300	303.6
RNNDeblur	29.19	0.9306	1400	37.1
SRN	30.60	0.9323	1600	33.6
DMPHN	31.25	0.9483	424	86.8
DSANet1	31.38	0.9485	254	32.2
DSANet2	31.55	0.9490	254	32.2
DSANet++	**31.65**	**0.9492**	254	32.2

Figure 7 presents the qualitative results obtained using our model. The tested images are blurred images from the test set, without corresponding GT images. The performance of our model is compared with those of MSCNN [5], RNNDeblur [1], SRN [6], and DMPHN [39]. DSANet performs better in recovering sharp images from complex dynamic scenes compared to previous methods.

We conducted ablation studies on various modules of the proposed network. DSANet operates at three different scales (K=1,2,3), corresponding to input image sizes of 256×256, 128×128, and 64×64, respectively. When K=1, only the original-scale image is input into the network. When K=2, images at two different scales are input. When K=3, images at three scales are processed. Table 2 provides the quantitative evaluation results of DSANet at different scales, while Figure 8 shows the qualitative evaluation results.

We observe that the performance improvement with a single-scale network is minimal. When K = 2, the network achieves good performance but has suboptimal processing speed. The three-scale network brings only slight performance improvements but significantly increases computation speed. These results demonstrate that the multi-scale network is highly effective for image deblurring tasks.

Figure 7. Visual comparison of the benchmark GoPro testing set.

 doi:10.20944/preprints202502.1206.v1

Table 2. Quantitative evaluation results of DSANet at different scales.

	k = 1	k = 2	k = 3
PSNR	30.5	31.4	31.65
SSIM	0.9402	0.9485	0.9492
Time (ms)	921	534	253

Figure 8. Visual comparison of the multi-scale networks.

To evaluate the effectiveness of the encoder-decoder structure and the supervised attention module, we designed and compared different models, as follows (outlined in Table 3). **1Ed Model**: This network consists of a single encoder and a single decoder, without the supervised attention module. **2Ed Model**: This network features two encoders and two decoders, but does not include the supervised attention module. **3Ed Model**: This network is composed of three encoders and three decoders, excluding the supervised attention module. In contrast, DSANet represents the final network configuration, which includes three encoders and three decoders, and incorporates the supervised attention module.

Table 3. Quantitative results for different models.

Model	PSNR	SSIM	Time (ms)	Size (MB)
1Ed Model	27.56	0.9255	80	5.3
2Ed Model	29.4	0.9358	145	12.4
3Ed Model	30.24	0.9401	280	31.5
DSANet	**31.65**	**0.9492**	**254**	32.2

In all networks, each encoder and decoder typically consist of an up- or down-sampling layer along with four residual blocks. Adding more residual blocks can expand the receptive field of each scale network. However, stacking too many residual blocks will increase the number of parameters, which can be inefficient. In DSANet, we use four residual blocks in each encoder and decoder, as adding more than four blocks results in minimal improvement. This design aims to balance efficiency and performance. Additionally, all networks employ the multi-scale architecture introduced by ConvLSTM [40] for a fair comparison.

As the number of encoders and decoders increases, the network's performance improves due to the larger receptive field. However, this also increases the number of parameters. To balance efficiency and effectiveness, we opted for a network with three encoders and three decoders. DSANet, with its three encoders, three decoders, and the supervised attention module, demonstrates superior performance. Notably, DSANet achieves a significant performance boost compared to the 3Ed Model, with only a 0.3 MB increase in parameters. This highlights the effectiveness of the supervised

 doi:10.20944/preprints202502.1206.v1

attention module, which enhances feature extraction as a lightweight gating mechanism while reducing redundant features.

Object detection is a major research area in computer vision, with significant advances driven by deep learning methods. However, image blurring remains a challenge that limits object detection performance. To assess the impact of deblurring on object detection, we selected 100 images from the GoPro dataset and evaluated object detection performance before and after deblurring using the Yolov4 algorithm. The qualitative results are shown in Figure 9.

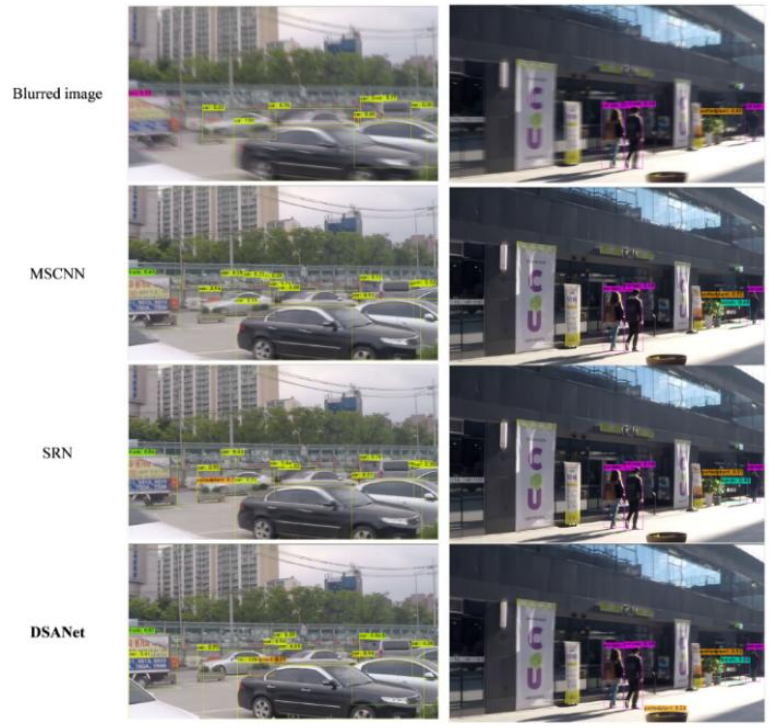

Figure 9. Qualitative evaluation of object detection.

5. Conclusions

Images are crucial for information acquisition, but distorted images can hinder this process, making image restoration essential. Image deblurring not only helps users retrieve information more effectively but also enhances performance in tasks such as object detection, segmentation, and classification. However, current deep learning-based deblurring methods face persistent issues. To address these challenges, this paper introduces a new network. The network utilizes a ConvLSTM-based encoder-decoder architecture to accelerate convergence and capture spatio-temporal features. Additionally, we propose a novel supervised attention module to mitigate the high computational costs associated with self-attention mechanisms. This module uses a lightweight gating mechanism to direct the model's focus towards highly correlated features of the blurred information as it transfers features across networks of varying sizes. To address limitations in traditional image reconstruction evaluation measures, we introduce several loss functions based on the fast Fourier transform. These functions enable the model to learn ambiguity features in the frequency domain effectively. Finally, we have compiled a new dataset that outperforms existing datasets, reducing challenges arising from discrepancies between synthetic and real images. A series of ablation experiments demonstrate the effectiveness of different modules within the DSANet framework.

 doi:10.20944/preprints202502.1206.v1

6. Patents

Acknowledgments: This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (IITP-2024-RS-2022-00156360).

Conflicts of Interest: The authors declare that there is no conflict of interests regarding the publication of this article.

References

1. Kim, T.H.; Ahn, B.; Lee, K.M. Dynamic scene deblurring. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 2013; pp. 3160–3167. https://doi.org/10.1109/ICCV.2013.392.
2. Sun, J.; Cao, W.; Xu, Z.; Ponce, J. Learning a convolutional neural network for non-uniform motion blur removal. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Boston, USA, 2015; pp. 769–777. https://doi.org/10.1109/CVPR.2015.7298677.
3. Schuler, C.J.; Hirsch, M.; Harmeling, S.; Scholkopf, B. Learning to deblur. IEEE Transactions on Pattern Analysis and Machine Intelligence **2016**, 38, 1439–1451. https://doi.org/10.1109/TPAMI.2015.2481418.
4. Nah, S.; Kim, T.H.; Lee, K.M. Deep multi-scale convolutional neural network for dynamic scene deblurring. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Honolulu, USA, 2017; pp. 257–265. https://doi.org/10.1109/CVPR.2017.35.
5. Ren, W.; Zhang, J.; Pan, J.; Liu, S.; Ren, J.S.; Du, J.; Cao, X.; Yan, M.H. Deblurring dynamic scenes via spatially varying recurrent neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 2022, 44, 3974–3987. https://doi.org/10.1109/TPAMI.2021.3061604.
6. Tao, X.; Gao, H.; Shen, X.; Wang, J.; Jia, J. Scale-recurrent network for deep image deblurring. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018; pp. 8174–8182. https://doi.org/10.1109/CVPR.2018.00853.
7. Kupyn, O.; Martyniuk, T.; Wu, J.; Wang, Z. DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 2019; pp. 8877–8886. https://doi.org/10.1109/ICCV.2019.00897.
8. Kuldeep, P.; Rajagopalan, A.N. Region-adaptive dense network for efficient motion deblurring. In Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34, 11882–11889. https://doi.org/10.1609/aaai.v34i07.6862.
9. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 2017, 39, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.
10. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 2015; pp. 234–241.
11. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015; pp. 1520–1528. https://doi.org/10.1109/ICCV.2015.178.
12. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Boston, USA, 2015; pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965.
13. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015; pp. 2650–2658. https://doi.org/10.1109/ICCV.2015.304.
14. Mathieu, M.; Couprie, C.; LeCun, Y. Deep multiscale video prediction beyond mean square error. arXiv 2015, arXiv:1511.05440.

 doi:10.20944/preprints202502.1206.v1

15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the NIPS, Long Beach, USA, 2017; pp. 5998–6008. arXiv:1706.03762.
16. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Honolulu, USA, 2017; pp. 6450–6458. https://doi.org/10.1109/CVPR.2017.683.
17. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018; pp. 3–19.
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018; pp. 7132–7141. https://doi.org/10.1109/CVPR.2018.00745.
19. Tsai, F.J.; Peng, Y.T.; Tsai, C.C.; Lin, Y.Y. BANet: A blur-aware attention network for dynamic scene deblurring. IEEE Transactions on Image Processing 2022, 31, 6789–6799. https://doi.org/10.1109/TIP.2022.3216216.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Las Vegas, USA, 2016; pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
21. Nair, V.; Hinton, G.E. Rectified linear units improve restricted Boltzmann machines. In Proceedings of the International Conference on Machine Learning, Haifa, Israel, 2010; pp. 807–814.
22. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change, C.R. ESRGAN: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision, 2018; pp. 1–16.
23. Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; Ma, J. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 2019; pp. 3106–3115. https://doi.org/10.1109/ICCV.2019.00320.
24. Hochreiter, S.; Schmidhuber, J. Long short-term memory. Neural Computation **1997**, 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
25. Ledig, C.; Theis, L.; Twitter, W.S. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Honolulu, USA, 2017; pp. 105–114. https://doi.org/10.1109/CVPR.2017.19.
26. Johnson, J.; Alahi, A.; Li, F. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 2016; pp. 694–711.
27. Jianbo, J.; Cao, Y.; Song, Y.; Lau, R. Look deeper into depth: Monocular depth estimation with semantic booster and attention driven loss. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018; pp. 53–69.
28. Ignatov, A.; Kobyshev, N.; Timofte, R.; Vanhoey, K. DSLR-quality photos on mobile devices with deep convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017; pp. 3297–3305. https://doi.org/10.1109/ICCV.2017.355.
29. Su, S.; Delbracio, M.; Wang, J.; Sapiro, G.; Heidrich, W.; Wang, O. Deep video deblurring for hand-held cameras. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Honolulu, USA, 2017; pp. 237–246. https://doi.org/10.1109/CVPR.2017.33.
30. Zhou, S.; Zhang, J.; Zuo, W.; Xie, H.; Pan, J.; Ren, J.S. DAVANet: Stereo deblurring with view aggregation. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Long Beach, USA, 2019; pp. 10988–10997. https://doi.org/10.1109/CVPR.2019.01125.
31. Lu, B.; Chen, J.C.; Chellappa, R. Unsupervised domain-specific deblurring via disentangled representations. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Long Beach, USA, 2019; pp. 10217–10226. https://doi.org/10.1109/CVPR.2019.01047.
32. Nimisha, T.M.; Sunil, K.; Rajagopalan, A.N. Unsupervised class-specific deblurring. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018; pp. 353–369.

 doi:10.20944/preprints202502.1206.v1

33. Lai, W.S.; Huang, J.B.; Hu, Z.; Ahuja, N.; Yang, M.H. A comparative study for single image blind deblurring. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Las Vegas, USA, 2016; pp. 1701–1709. https://doi.org/10.1109/CVPR.2016.188.
34. Rim, J.; Lee, H.; Won, J.; Cho, S. Real-world blur dataset for learning and benchmarking deblurring algorithms. In Proceedings of the European Conference on Computer Vision, 2020; pp. 184–201.
35. Zhang, K. Deblurring by realistic blurring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; pp. 2734–2743. https://doi.org/10.1109/CVPR42600.2020.00281.
36. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the NIPS, 2014.
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015; pp. 1026–1034. https://doi.org/10.1109/ICCV.2015.123.
38. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. arXiv **2014**, arXiv:1412.6980.
39. Zhang, H.; Dai, Y.; Li, H.; Koniusz, P. Deep stacked hierarchical multi-patch network for image deblurring. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Long Beach, USA, 2019; pp. 5978–5986.
40. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Proceedings of the International Conference on Neural Information Processing Systems, 2015; pp. 802–810.