



A Deep Convolutional Encoder–Decoder–Restorer Architecture for Image Deblurring

Yiqing Fan¹ · Chaoqun Hong¹ · Guanghui Zeng¹ · Lijuan Liu¹

Accepted: 30 October 2023 / Published online: 11 February 2024
© The Author(s) 2024

Abstract

The accuracy of many computer vision tasks is reduced by blurred images, so deblur is important. More details of the image can be captured by a common multi-stage network, but the computational complexity of this method is higher compared with a single-stage network. However, a single-stage network cannot capture multi-scale information well. To tackle the problem, a novel convolutional encoder–decoder–restorer architecture is proposed. In this architecture, a multi-scale input structure is used in the encoder. Improved supervised attention module is inserted into the encoder for enhanced feature acquisition. In decoder, information supplement block is proposed to fuse multi-scale features. Finally, the fused features are used for image recovery in the restorer. In order to optimise the model in multiple domains, the loss function is calculated separately in the spatial and frequency domains. Our method is compared with existing methods on the GOPRO dataset. In addition, to verify the applications of our proposed method, we conduct experiments on the Real image dataset, the VOC2007 dataset and the LFW dataset. Experimental results show that our proposed method outperforms state-of-the-art deblurring methods and improves the accuracy of different vision tasks.

Keywords Image deblurring · Encoder–decoder–restorer architecture · Multi-Scale · Deep learning

✉ Guanghui Zeng
zenggh@gx-credit.com

Yiqing Fan
yqfan@stu.xmut.edu.cn

Chaoqun Hong
cqhong@xmut.edu.cn

Lijuan Liu
ljliu@xmut.edu.cn

¹ Key Laboratory of Fujian Universities for Virtual Reality and 3D Visualization, Xiamen University of Technology, Xiamen 361024, Fujian, China

1 Introduction

A clear image preserves more information than a blurred image. However, motion blur is very common in our daily lives. It is often caused by camera shake and the movement of objects. This causes the edges of the image to become blurred, which greatly reduces the accuracy of computer vision tasks. Therefore, image deblurring tasks need to be tackled urgently.

In the past decades, a lot of research work has been carried out by many scholars in order to obtain clear latent image. Traditional image deblurring methods are usually based on the premise of simplified ideal conditions [1–4]. In recent years, deep learning methods have been widely used in the field of computer vision, such as image deraining [5, 6], image denoising [7, 8], image deblurring [9, 10] and recently the particularly popular human activity recognition [11–13]. Learning-based methods have achieved good results in image deblurring tasks. Most of the existing estimation methods for learning blur kernels use convolutional neural networks to extract blur kernels [14, 15]. However, such methods rely excessively on the accuracy of blur kernel estimation. In the case of inaccurate blur kernel estimation, the quality of recovered image is poor. As a result, more and more scholars have recently been devoted to the study of end-to-end image deblurring methods [16–19]. To recover a latent clear image directly from a blurred image, these methods use convolutional neural networks to learn the mapping relationship between the blurred image and the clear image. Since the methods do not require the estimation of blur kernels, the errors caused by the estimation of blur kernels are reduced. However there are problems with existing end-to-end networks. Single-scale end-to-end networks [16, 17] take less time to recover images, but are unable to recover images by acquiring multi-scale information. Multi-scale end-to-end networks [18, 19] provide multi-scale acquisition of images by stacking sub-networks, but inevitably with an increase in computational time.

To tackle the above problem, we propose a new Encoder–Decoder–Restorer network structure for image deblurring. The single UNET is used as the backbone network and the multiple input encoder is used [20]. Multi-scale information can be captured by this structure. For the enhanced feature extraction capability of the network, we improve the existing supervised attention module and add it to the encoding blocks. We propose that the information supplement block can effectively help fusion of different scale feature maps. To make use of the multi-scale information for the restoration of multi-scale images, we specially propose the restorer. Finally, as shown in Fig. 1a, b, the images in the spatial domain can clearly express the spatial structure of the images, while the images in the frequency domain can more clearly detect the disparity between the high and low frequencies of the images. Therefore, we propose the multi-scale content reconstruction loss function and the multi-scale frequency reconstruction loss function in the spatial and frequency domains respectively.

The contributions of the proposed method can be summarized as follows:

1. We propose a new network structure for image deblurring. It contains a multi-scale feature encoder, a multi-scale feature decoder and a multi-scale image restorer.
2. To enhance the feature extraction capability of the multi-scale feature encoder, an improved supervised attention module is added to the encoding block. In the multi-scale feature decoder, an information supplement block is proposed for better fusion of feature information of different scales.
3. Finally, in order to train the model to produce clearer latent images, we use a multi-scale content reconstruction loss function and a multi-scale frequency reconstruction loss function in the spatial and frequency domain respectively.

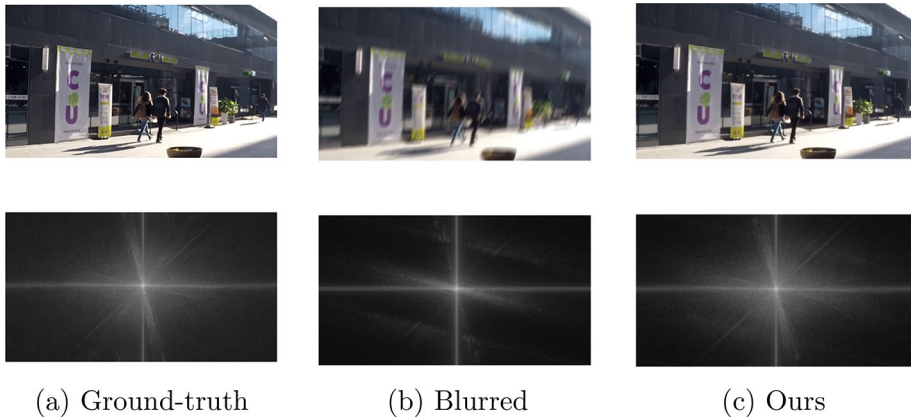


Fig. 1 Spectrograms after centring of different images

The remainder of our paper is organized below. In Sect. 2, we introduce the existing methods on image deblurring. In Sect. 3, we outline the proposed method first and then introduce it in detail. After theoretical introduction, in Sect. 4 we perform relevant experiments to demonstrate the superiority of our method. In Sect. 5, we provide some discussion about the novelty and improvements of the proposed method. Finally, in Sect. 6, we conclude the paper with a summary of the method.

2 Related Works

In this section, we briefly introduce existing image deblurring methods, which include traditional image deblurring methods, learning-based blurring kernel estimation methods and end-to-end image deblurring methods.

Traditional image deblurring methods. Traditional image deblurring methods are divided into blind and non-blind deblurring. In the non-blind deblurring methods, the blurred image is recovered by assuming a blurring kernel. Such as the common Lucy-Richardson algorithm [21] and the Wiener filtering algorithm [20], these algorithms restored a clear image using deconvolution. The latent image obtained is not clear enough because the a priori information of the image was not fully utilized. Many later methods [22, 23] have been based on these two methods with improvements to achieve image deblurring. However, the real blur kernel is often different from what we expect, so the traditional non-blind deblurring methods have some limitations in dealing with real blurred images. The traditional blind deblurring methods are the maximum posterior probability based deblurring method algorithm [1, 2] and the variable Bayesian framework based deblurring algorithm [3, 4]. The deblurring algorithm based on the maximum posterior probability was more efficient, but it did not take advantage of the inherent distribution of the data and had some limitations. The variational Bayesian-based image deblurring algorithm was theoretically more robust by considering all possible solutions. However, as all solutions needed to be considered, the algorithm was too slow to meet the practical requirements.

Learning-based kernel estimation methods. Early learning-based blur kernel estimation methods [15] were used to deblur images by predicting the probability distribution of patch

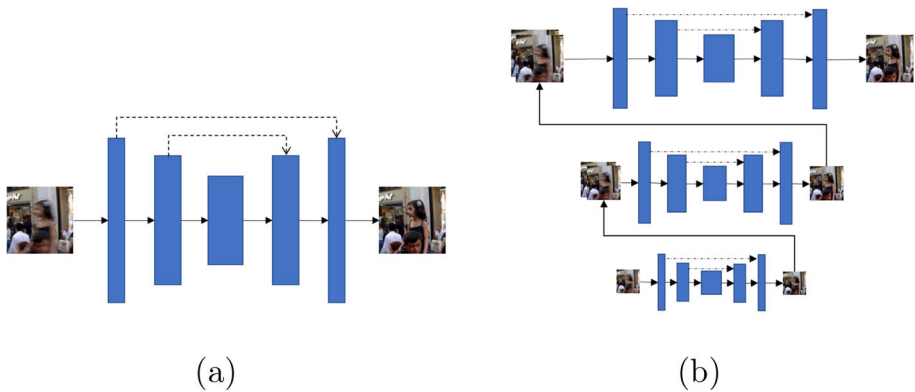


Fig. 2 End-to-end networks architecture for image deblurring: **a** Single-stage networks, **b** multi-stage networks

motion blur. Recently, Kaufman et al. [24] proposed a new method for blind image deblurring. This method consists of an analysis network and a synthesis network. The analysis network is used to estimate the blur kernel. The synthesis network performs the deblurring based on the obtained blur kernel to the image. The clarity of the recovered image is influenced by the accuracy of the blur kernel estimation. If the blur kernel is incorrectly estimated, the image will be poorly recovered.

End-to-end image deblurring methods. Since the estimation of blur kernels is more difficult, more and more scholars are working on end-to-end image deblurring methods. The methods have two main network structures. One is a single-stage deblurring network. In a single-stage network, the original-scale blurred image is used as input, as shown in Fig. 2a. Yuan et al. [25] proposed a lightweight single-stage network. This network introduces bi-directional optical flow to guide the learning of deformable convolution and uses the sampled points of deformable convolution to approximate the blurred kernel. However, this network does not achieve the best performance. The other is a multi-stage deblurring network. The structure is shown in Fig. 2b, the image is scaled to different scales and is recovered starting from the smallest scale. This structure was shown to be effective through experiments by Nah et al. [18]. However, the computational complexity of this method increases with it.

In order to tackle these problems, an efficient deblurring network has been proposed. This network structure enables the acquisition of information from multi-scale images and saves the time consumed by stacking sub-networks in order to acquire multi-scale information. We will describe our proposed network in detail in Sect. 3.

3 Methodology

In this paper, we propose restorer and use it together with UNET to form Encoder–Decoder–Restorer network structure. In order to enhance the feature extraction and fusion capabilities of the single-stage network, we improve the supervised attention modules and propose the information supplement block for the encoder and decoder respectively. In addition, we propose new loss functions in the spatial and frequency domains. In the following, we describe our approach in detail.

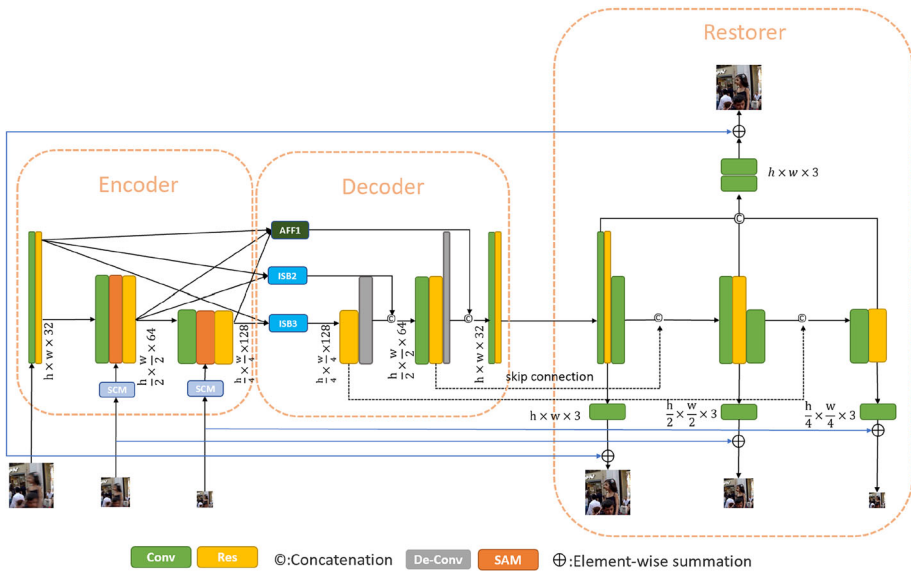


Fig. 3 The architecture of our proposed network

3.1 Network Architecture

In our network, UNET is used as the backbone network containing a multi-scale feature encoder, a multi-scale feature decoder and a multi-scale image restorer, as shown in Fig. 3. We find that blur of larger degree is more easily removed in smaller scale images, while blur of smaller degree is more easily found in larger scale images. So images of different scales are fed into the encoder. Due to the weak feature extraction ability of the single-stage network, it is not able to extract useful information well. So we improve the supervised attention module(SAM) and add it to the encoder. To enhance the information fusion capability of the decoding blocks, we use an information supplement block(ISB) and an asymmetric feature fusion(AFF) [17]. In order to recover images with more detail, the multi-scale image restorer is proposed and used in the network.

Multi-scale feature encoder. Features of different image scales are extracted in a multi-scale feature encoder, which consists of three encoding blocks. The input image is adjusted to different scales and then fed into different encoding blocks. The output of each encoding block layer is shown below:

$$EB_i^{out} = \begin{cases} EB_i(o(B_i)), & i = 1, \\ EB_i(SAM(o(EB_{i-1}^{out}); SCM(B_i)_i^{out})_i^{out}), & i = 2, 3, \end{cases} \quad (1)$$

where EB_i^{out} denotes the output of the i -th encoding block. B_i , SAM_i^{out} and SCM_i^{out} denote the blurred image and the outputs of SAM and SCM at layer i , respectively. o is the mapping function used to generate the input feature images.

In the first encoding block, a set of feature maps are generated for the input image using a 3×3 convolutional layer. These feature maps are fed into the encoding block for calculation and passed to the next layer. In the second and third encoding blocks, the initial features are extracted from the input blurred image using a shallow convolution module(SCM). The

features output from the upper layer are fed into the SAM together with the initial features, which are used to obtain and propagate useful features. Finally, the features are computed and passed through the encoding block to the next layer.

Multi-scale feature decoder. The obtained feature information is fused in a multi-scale feature decoder consisting of three decoding blocks. ISB or AFF is used to assist in the fusion of feature information. The output of each layer of the decoding block can be expressed as:

$$DB_i^{out} = \begin{cases} DB(AF F_i^{out}; DB_{i+1}^{out}), & i = 1, \\ DB(ISB_i^{out}; DB_{i+1}^{out}), & i = 2, \\ DB(ISB_i^{out}), & i = 3, \end{cases} \tag{2}$$

where DB_i^{out} denotes the output of the i -th decoding block. $AF F_i^{out}$ and ISB_i^{out} denote the output of the i -th AFF and ISB respectively.

Multi-scale image restorer. We propose the multi-scale image restorer for recovering images of different scales. The multi-scale image restorer contains three restoration blocks, each of which is used to generate a latent image of a different scale and the corresponding feature information. Decoded features of the same scale are used to recover the image. The process of image recovery can be expressed as follows:

$$\hat{S}_i^{out} = \begin{cases} \sigma(RB(DB_i^{out})) + B_i, & i = 1, \\ \sigma(RB(DB_i^{out}; RB_{i-1}^{out})) + B_i, & i = 2, 3, \end{cases} \tag{3}$$

where B_i and \hat{S}_i^{out} denote the input blurred image and the output latent image at layer i . RB_i^{out} indicates the output of the layer i image restoration block. σ denotes the mapping function that maps the feature map into an image.

In order to be able to remove different degrees of blur from the image, the feature information extracted from the three restoration blocks is used to generate the final latent image. The generation process is shown below:

$$\hat{S}^{out} = \sigma(RB_1^{out}; RB_2^{out}; RB_3^{out}) + B_1, \tag{4}$$

where \hat{S}^{out} denotes the final output image.

3.2 Information Supplement Block (ISB) and Asymmetric Feature Fusion (AFF)

To enhance the ability of this network to fuse information at all scales, inspired by Asymmetric Feature Fusion(AFF) [17], we propose the Information Supplement Block(ISB) and use both AFF and ISB for the network to perform feature fusion at different scales. The structure is shown in Fig. 4.

Since more feature information is needed to fuse the largest scale feature maps, the three scales of encoding information are fused to generate the largest scale feature maps in AFF. Other scales of images do not require so much feature information. To reduce computational time, ISB is used to generate feature maps at other scales. Specifically, the formulae for AFF and ISB are as follows:

$$ISB_i^{out} = ISB_i \left((EB_1^{out})^\downarrow; EB_i^{out} \right), \tag{5}$$

$$AFF_i^{out} = AFF_i \left(EB_1^{out}; (EB_2^{out})^\uparrow; (EB_3^{out})^\uparrow \right), \tag{6}$$

where ISB_i^{out} and AFF_i^{out} denote the output of ISB and AFF at layer i , respectively. \uparrow and \downarrow denote upsampling and downsampling, respectively.

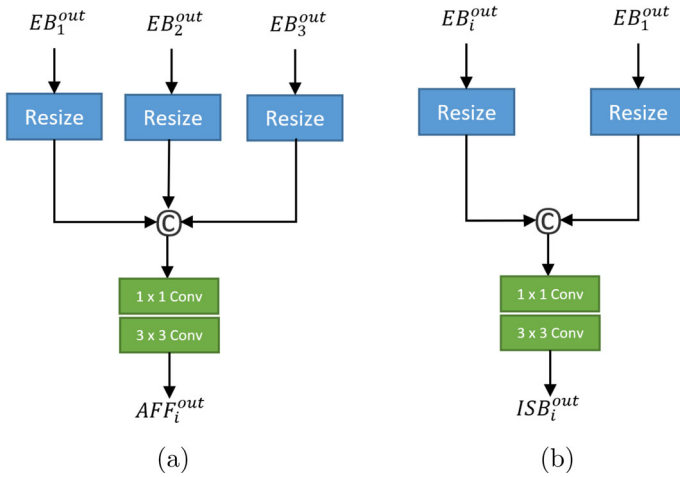
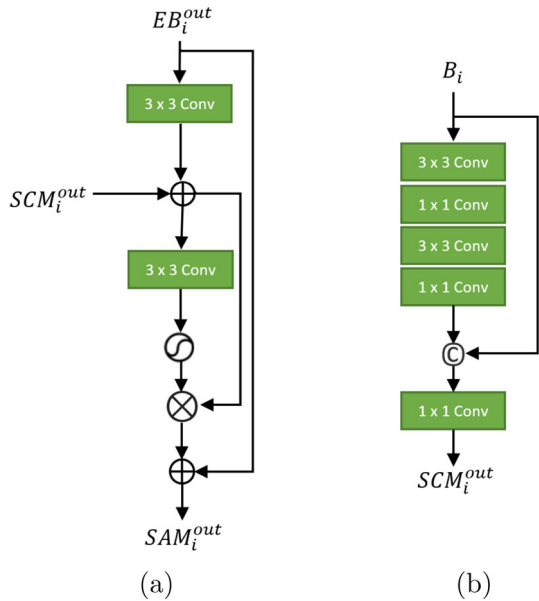


Fig. 4 The structures of sub-modules: **a** information supplement block (ISB), **b** asymmetric feature fusion (AFF)

Fig. 5 The structures of sub-modules: supervised attention module (SAM), **b** shallow convolution module (SCM)



3.3 Supervised Attention Module (SAM)

Recently, supervised attention modules have achieved good results in multi-stage image recovery tasks [26, 27]. Therefore, we introduce SAM to our network for obtaining important feature information. This module was proposed by Zamir et al. [28]. We improve the module and the improved structure is shown in Fig. 5a.

In this module, the feature maps obtained from the encoding blocks are turned into residual maps by using convolution. The residual maps are added to the initial features obtained from the shallow convolution module (structure shown in Fig. 5b). Mask maps are generated using

a 3×3 convolution and a Sigmoid function. The mask maps and the input features are calibrated. The calibrated information is added to the encoded information. Finally, a set of features with weights is obtained and passed to the next stage. By introducing this module, the weight of useful information is increased and the weight of useless information is decreased.

3.4 Loss Function

To better recover the images, we propose loss functions in the spatial and frequency domains respectively. The multi-scale content reconstruction loss function is used in the spatial domain in order to reduce the structural differences between the generated image and the real sharp image. In the frequency domain, the multi-scale frequency reconstruction loss function is used to reduce the difference between the generated image and the real sharp image in the high frequency region and the low frequency region. The total loss function of the training network is expressed as follows:

$$\mathcal{L} = \mathcal{L}_{MSCR} + \mathcal{L}_{MSFR}. \quad (7)$$

Since reducing the difference between the multi-scale image and the real sharp image helps to improve the sharpness of the final latent image, we use both the multi-scale image and the final latent image for the loss calculation. The specific definition of each loss function is as follows:

Multi-scale content reconstruction loss function. The multi-scale content reconstruction loss function is used to calculate the distance between the generated latent image and the real sharp image. The loss function is formulated as follows:

$$\mathcal{L}_{MSCR} = \sum_{i=1}^I \frac{\|\hat{S}_i^{out} - S_i\|_1}{N_i} + \alpha \frac{\|\hat{S}^{out} - S_1\|_1}{N_1}, \quad (8)$$

where I denotes the total number of layers. S_i denotes the real sharp image corresponding to the i -th layer. N_i denotes the total number of elements in the i -th layer. The hyperparameter α is set to 0.5.

Multi-scale frequency reconstruction loss function. To reduce the difference between the generated image and the real sharp image in the frequency domain, the image is transformed to the frequency domain to calculate the loss. This loss function is defined as follows:

$$\mathcal{L}_{MSFR} = \beta \sum_{i=1}^I \frac{\|\mathcal{F}(\hat{S}_i^{out}) - \mathcal{F}(S_i)\|_1}{N_i} + \gamma \frac{\|\mathcal{F}(\hat{S}^{out}) - \mathcal{F}(S_1)\|_1}{N_1}, \quad (9)$$

where \mathcal{F} denotes the fast Fourier transform of the image. The hyperparameters β and γ are set to 0.3 and 0.1 respectively

4 Experiment

In this section, we conduct several experiments to demonstrate the effectiveness of our network in the image deblurring task. Our method is compared with popular image deblurring methods in recent years and ablation experiments are performed to verify the effectiveness of our proposed module. Next, datasets with different resolutions are experimented to ensure

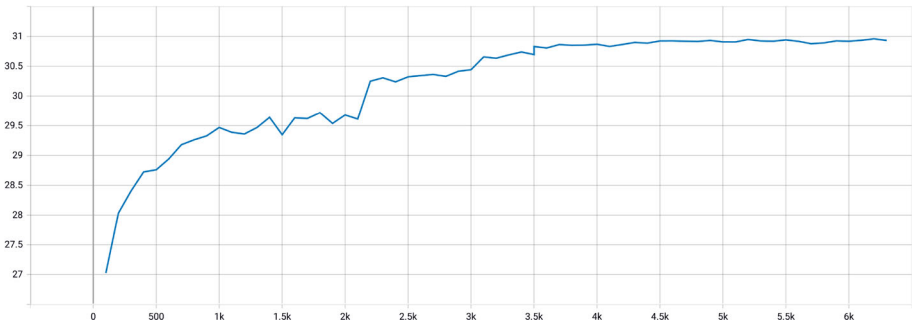


Fig. 6 PSNR value calculated every 100 epochs (x-axis represents the number of epochs, and y-axis represents the PSNR values)

that our proposed method has good temporal stability. Finally, in order to verify the applications of our proposed method, we conduct experiments on the Real image dataset, the LFW dataset and the VOC2007 dataset.

4.1 Implementation Details

In each iteration of training, we will randomly sample four images and crop them randomly into images of size 256×256 . In each patch, we set a probability of 0.5 for horizontal flipping. Adam optimizer is used to train the model, where parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is set to 1×10^{-4} for the first 3000 epochs and is decayed by a factor of 0.5 every 1000 epochs. The validation curve is shown in Fig. 6. The experiments show that 6000 epochs are sufficient for convergence. All of our experiments are run with an Intel(R) Xeon(R) CPU E5-2680 v3 CPU and an NVIDIA TITAN X (Pascal) GPU.

4.2 Dataset

GoPro dataset [18]. To train the network using more realistic blurred images, the GoPro dataset proposed by Nah et al. [18] is used to train the network. The blurred images in the GoPro dataset are obtained using a high-speed camera that acquires a sequence of clear images. These short intervals of clear images are averaged to obtain blurred images. The GoPro dataset is one of the most popular datasets in the field of deblurring images, as the blurred images obtained are particularly realistic. This dataset contains 3214 pairs of sharp and blurred images with a resolution of 1280×720 , as shown in Fig. 7, with 2103 pairs in the training set (we use 1472 pairs for training and 631 pairs for validation) and 1111 pairs in the test set. This dataset is used to train the model and to conduct qualitative and quantitative comparisons.

Real image dataset [29]. Real image dataset is used for testing, in order to verify that our proposed method can remove real-world blur well. As shown in Fig. 8, this dataset consists of 100 images in different scenes. All these blurred images are captured in the real-world scenarios from different cameras (e.g., consumer cameras, DSLR, or cellphone cameras), different settings (e.g., exposure time, aperture size, ISO), and different users.

LFW dataset [30] and **VOC2007 dataset** [31]. We conduct experiments on the LFW dataset and the VOC2007 dataset to verify that our proposed method can improve the accuracy of different vision tasks. A special function is used to generate blur of different sizes and



Fig. 7 Image pairs in the GoPro dataset



Fig. 8 Images from the real image dataset



Fig. 9 Blurred images generated on different datasets

orientations on the image, as shown in Fig. 9. We use the LFW (Labeled Faces in the Wild) dataset for our experiments on face recognition. This dataset contains the face data of 5,749 people from different countries, ages and genders. 4,952 images from the VOC2007 dataset, of which this dataset contains 20 species, are selected for object detection.

4.3 Performance Comparison

Quantitative Evaluation. Our proposed method is compared with existing methods on the GoPro dataset. PSNR (Peak Signal-to-Noise Ratio), SSIM (Structure Similarity Index

Table 1 Quantitative comparison with existing methods on the GoPro test set

Method	PSNR	SSIM	Runtime
Sun et al. [15]	24.64	0.843	20 (min)
Gong et al. [32]	26.06	0.863	25 (min)
Nah et al. [18]	28.49	0.917	15 (s)
Zhang et al. [33]	29.19	0.931	1.4 (s)
Li et al. [34]	29.73	0.902	1.2 (s)
SRN [19]	30.10	0.932	1.6 (s)
DBGAN [35]	30.10	0.942	1.3 (s)
ESTRNN [36]	31.07	0.902	0.206 (s)
Tang et al. [16]	31.13	0.951	0.088 (s)
MT-RNN [37]	31.15	0.945	0.063 (s)
Wan et al. [38]	31.24	0.945	0.42 (s)
MIMO-UNET [17]	31.36	0.947	0.016 (s)
Ours	32.09	0.955	0.023 (s)

Measure) and runtime are selected as evaluation metrics. The experimental results are shown in Table 1. Our proposed model achieves good performance in all three metrics. Although it only underperforms MIMO-UNET by 0.007 s in terms of runtime, we achieve optimal results in terms of PSNR and SSIM.

Qualitative Evaluation. We select images from different scenes on the GoPro dataset for qualitative comparison. As shown in Fig. 10, through qualitative comparison, we can find that our proposed method outperforms other deblurring methods in different scenes.

4.4 Ablation Study

In order to select the optimal model, we perform an ablation study of the component modules in the model. In addition to this, we compare the performance of the trained models for different values of the hyperparameters (α , β and γ). The GoPro test set is used for this set of studies and for the metrics we choose PSNR and SSIM.

4.4.1 Ablation Study of Component Modules

We conduct experiments to verify the contribution of our proposed modules to the network, which include SAM, ISB and multi-scale image restorer. SAM is replaced by element-wise sum. The image is recovered in the decoder. The results are shown in Table 2. Our proposed module can steadily improve the quality of image recovery. In particular, the PSNR is improved by 0.67 when the restorer is used in the network.

4.4.2 Ablation Study of the Hyperparameters of the Loss Function

The loss function consists of a multi-scale content reconstruction loss function and a multi-scale frequency reconstruction loss function, where three hyperparameters are used to control the weighting of each component. α is set to 0.4, 0.5 and 0.6, while β and γ are set to 0.1, 0.2 and 0.3. The results of the experiments are shown in the Table 3. The model performs best when α , β and γ are set to 0.5, 0.3 and 0.1 respectively.

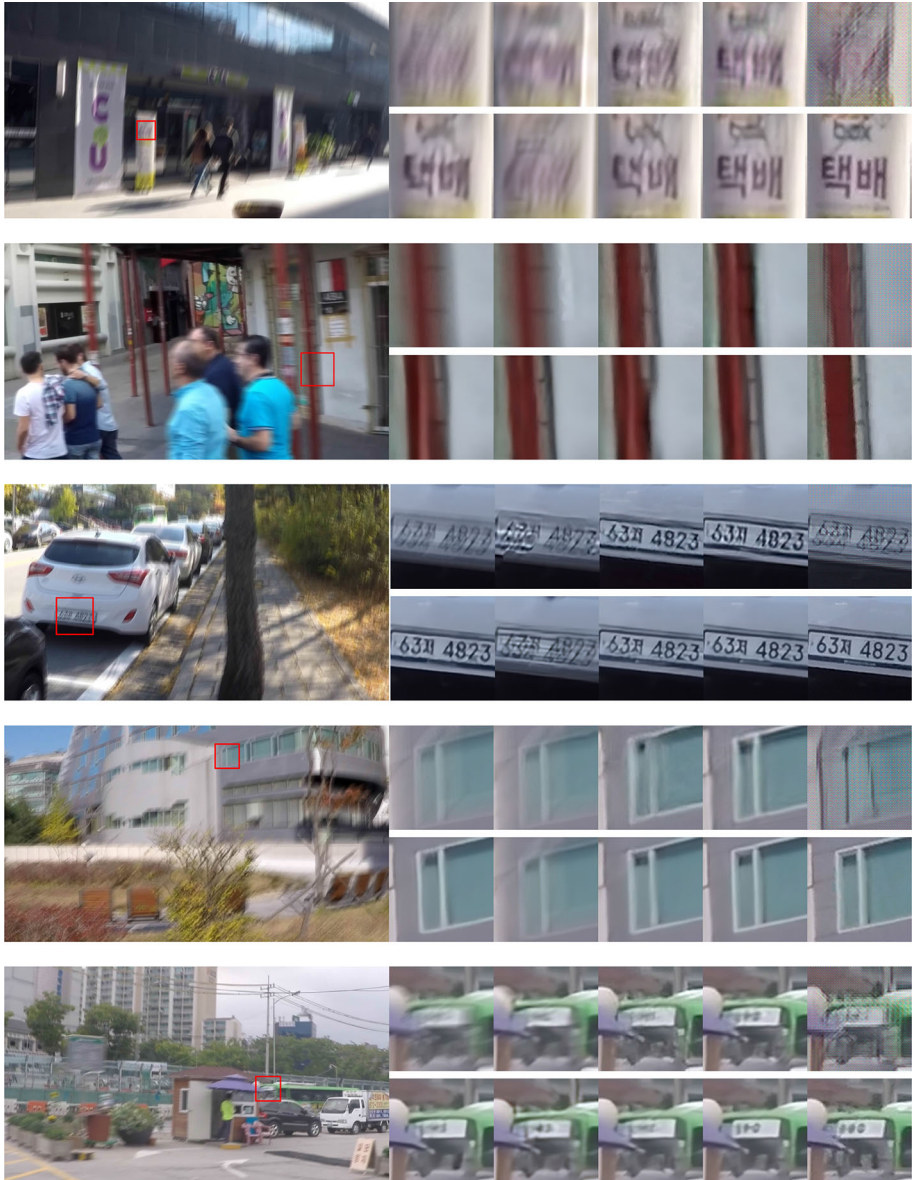


Fig. 10 Qualitative comparison of different methods in different scenarios on GoPro test sets. For clarity, the magnified parts of the resultant images are displayed. From left-top to right-bottom: Blurry images, Gong et al., Nah et al., Zhang et al., DBGAN, MT-RNN, ESTRNN, MIMO-UNET, Ours, Ground-truth images, respectively

4.5 Computational Time

In current society, there is an increasing demand for higher resolution images, so when performing deblurring tasks images of different resolutions are encountered. In order to

Table 2 Ablation study of component modules

SAM	ISB	Restorer	PSNR	SSIM
×	×	×	30.13	0.931
×	✓	×	31.34	0.944
✓	×	×	31.22	0.945
✓	✓	×	31.42	0.948
✓	✓	✓	32.09	0.955

Validate the effectiveness of our proposed different components on the GoPro test set

Table 3 Ablation study

α	β	γ	PSNR	SSIM
0.4	0.1	0.1	32.062	0.954
0.5	0.1	0.1	32.075	0.954
0.6	0.1	0.1	32.052	0.954
0.5	0.2	0.1	32.075	0.955
0.5	0.3	0.1	32.094	0.955
0.5	0.3	0.2	32.090	0.955
0.5	0.3	0.3	32.037	0.954

Performance of the model on the GoPro test set with different hyperparameters

Table 4 Computational time comparisons on datasets of different resolutions

	256×256	512×512	1024×1024
Computational time	0.0226	0.0238	0.0238

verify the impact of different resolutions of images on the computational time, three different resolutions of datasets are used to compare. The experimental results are shown in Table 4. We can see that the resolution of the image has very little effect on our model.

4.6 Applications

Our method is applied to the Real image dataset, the LFW dataset and the VOC2007 dataset. The purpose of this is to verify that our method can remove blur well in the real world, and to verify the effectiveness of our proposed method in different vision tasks.

Removing blur from the real world. To validate that our proposed method can address blurred images in the real world, our method is applied to the Real image dataset. The results are shown in the Fig. 11. Although this dataset contains images of different scenes, at different scales, the latent clear images can still be recovered well.

Face Recognition. In this experiment, the face recognition algorithm (Facenet) proposed by Google as is used as the main network [39]. During training, Inception-ResNet-v1 is used as the feature extraction network and the dataset used is CASIA-WebFace. Blurred and deblurred datasets are used for the experiments (as shown in Fig. 12). In addition, three metrics, Accuracy, Validation rate and ROC curve, are selected for comparison. The performance on different datasets is shown in Table 5 and Fig. 13.

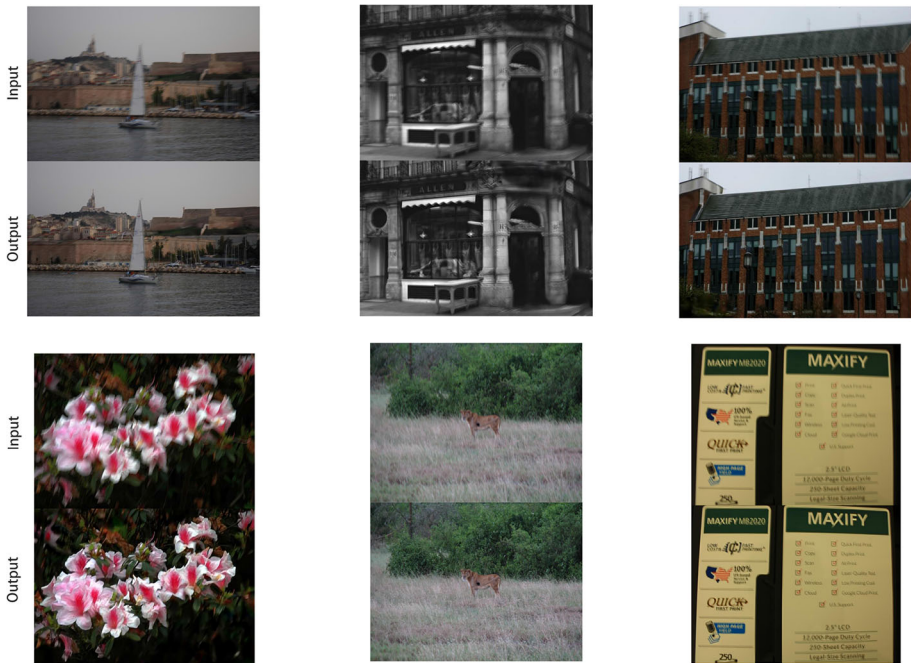


Fig. 11 Performance of our method on Real image datasets



Fig. 12 Deblurring of face images using our proposed methods

Table 5 Comparison of accuracy and verification rates on blurred and deblurred face datasets

	Blurred	Deblurred
Accuracy	59.45%	80.23%
Validation rate	2.03%	13.97%

Experimental results show that blurred images can severely degrade the performance of face recognition algorithms. However, the images processed by our method have a significant improvement in all metrics. Therefore, our method still performs well in removing face blur.

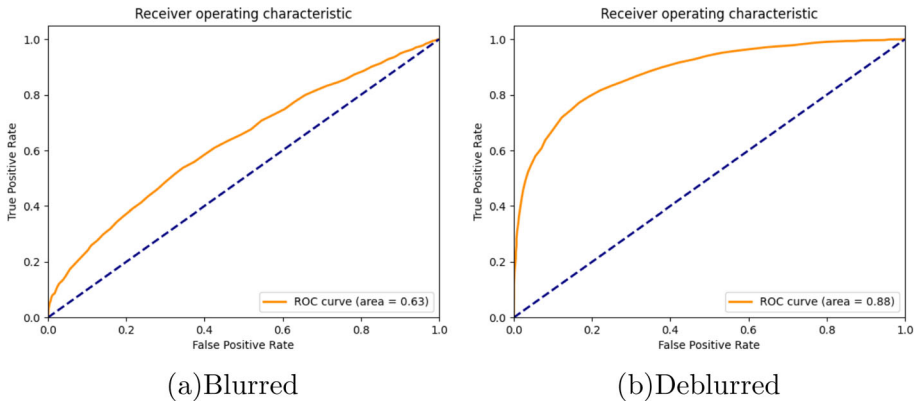


Fig. 13 Compute ROC curves on the blurred face dataset and the deblurred face dataset using our method

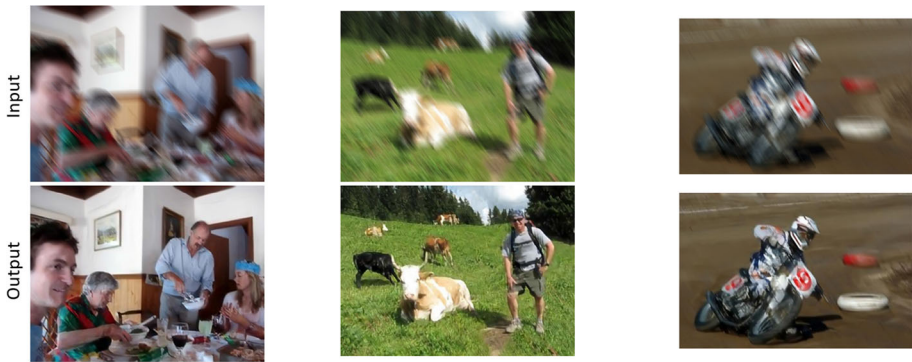


Fig. 14 Deblurring of the object detection dataset using our proposed method

Table 6 Yolo V3 performance on blurred and deblurred datasets

	mAP	mP	mR	mF1
Blurred	32.42%	77.43%	22.24%	0.35
Deblurred	65.16%	82.66%	53.94%	0.65

Object Detection. Our method is applied to object detection and to further validate the effectiveness of this method in the field of high-level vision. YOLO V3 is used in this experiment [40]. The training sets of VOC2012 and VOC2007 are used to train the model. We use the test set of VOC2007 for testing, which contains 20 object classes. Blur is added to the test set and the images are deblurred using our proposed method, as shown in Fig. 14. Mean Average Precision(mAP), mean Precision(mP), mean Recall(mR), mean F1 score(mF1) are used as evaluation metrics. We set the threshold value to 0.5. The experimental results are shown in the Table 6.

We can see that the model acquires higher values in the deblurred dataset on all types of metrics. In addition to this, we calculate the average precision (AP) for each class separately and the results are shown in Fig. 15. We can find that the precision of detecting each class of objects has been improved on the deblurred dataset.

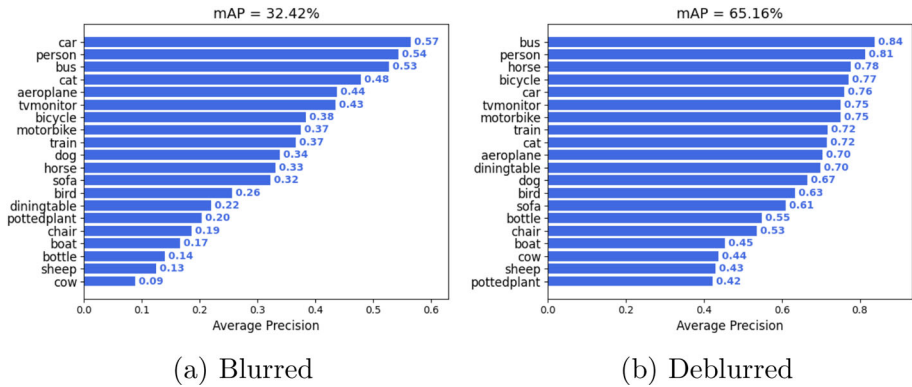


Fig. 15 Average Precision for each category in the dataset

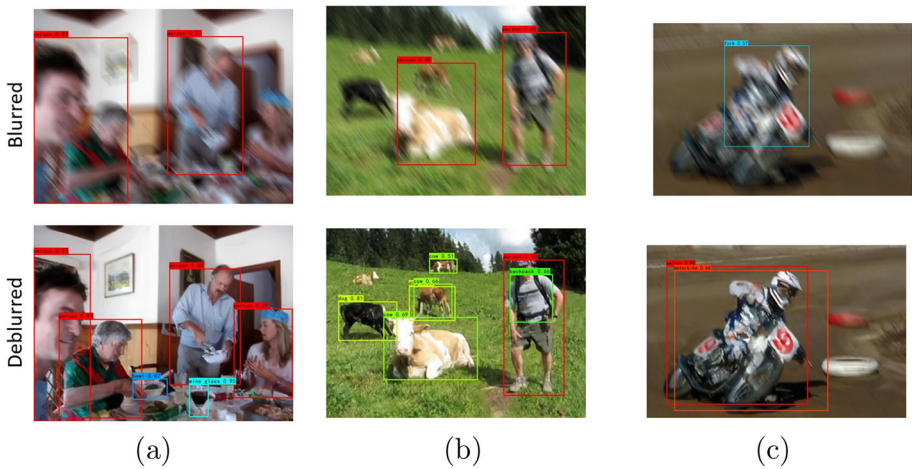


Fig. 16 Performance of blurred and deblurred images on YOLO V3

To further verify the effect of blurred images on object detection, we use the network to perform object detection on blurred images and deblurred images respectively. The experimental results are shown in Fig. 16.

We can find that in Fig. 16a the blurred image is detected for just two objects, while the deblurred image is able to be detected for six objects in the image. In Fig. 16b the blurred image is identified correctly for only one object, however the deblurred image is identified correctly for five objects. In Fig. 16c, the person and motorbike are not recognised in the blurred image. Finally, the threshold is changed to test whether our processed images still have good accuracy and recall. We set the threshold values to 0.3, 0.4, 0.5, and 0.6 respectively. The experimental results are shown in Table 7.

With different thresholds, the processed images can still achieve better results. Through the above experiments, we can see that our proposed deblurring method still works well even in the field of object detection.

Table 7 Performance of Yolo V3 on the deblurred dataset under different thresholds

	mP	mR	mF1
0.3	74.03%	59.51%	0.66
0.4	78.85%	56.72%	0.66
0.5	82.66%	53.94%	0.65
0.6	85.68%	50.59%	0.64

5 Discussion

According to the above experiments and the improvement in performance, the novelty and contribution of our method can be shown.

First, the method tries to tackle the issue of image deblurring. We construct a new network by proposing the multi-scale image restorer. Based on the existing multiple input encoder, we improve and add the supervised attention module, this module can acquire effective information for image deblurring, thus enhancing the feature extraction capability of the model. In the decoder, the information supplement block is designed. This block combines information from different scales and helps to fuse features from different scales. Our proposed restorer can recover images at multiple scales and use them all in the calculation of losses. Since generating latent images at different scales can help remove different degrees of blur, these latent images are used in the calculation of the loss. Finally, we propose Multi-scale content reconstruction loss function and multi-scale frequency reconstruction loss function in the spatial and frequency domains for reducing the gap between latent images and true sharp images in terms of spatial structure and image edges, respectively.

Second, a large number of experiments has been conducted. The experimental results according to 4.3 show that our proposed model outperforms existing methods. The effectiveness of the proposed module and the optimality of the loss function parameters are verified by 4.4. The stability of the computational time on different resolution datasets is verified in 4.5, where the experimental results show that the image resolution has little influence on our method. In addition, in 4.6 we apply the method to two high-level computer vision tasks, namely face recognition and object detection, and we find that the method can effectively help improve the accuracy of high-level computer vision tasks.

In the future, we will focus on researching more lightweight and efficient networks and applying them to real-life scenarios.

6 Conclusion

In this paper, a novel network is proposed for removing motion blur from images. Although only a single-stage network is used, it is able to fuse multi-scale image features well for images recovery. In the input stage, different encoding blocks are fed with images of different scales. An improved supervised attention module is added to enhance the feature extraction capability of the encoder. In the decoding blocks, to enhance the feature fusion, we add an information supplement block. Finally, a restorer is proposed for the recovery of multi-scale images. In addition to this, we compute losses in the spatial and frequency domains. Experiments show that our method has good performance and can improve the accuracy of different vision tasks.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (Nos. 62173285 and 62103345), the Fujian Provincial Natural Science Foundation of China (Nos. 2021J011181 and 2020J02160), the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB2021C01) and Xiamen Youth Innovation Fund Project (Nos. 3502Z20206072 and 3502Z20206076).

Author Contributions YF wrote the main manuscript text and prepared figures. CH provide laboratory equipment and guide experiments. All authors reviewed the manuscript.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Whyte O, Sivic J, Zisserman A (2014) Deblurring shaken and partially saturated images. *Int J Comput Vis* 110(2, SI):185–201. <https://doi.org/10.1007/s11263-014-0727-3>
2. Pan J, Hu Z, Su Z, Yang M-H (2014) Deblurring text images via l0-regularized intensity and gradient prior. In: 2014 IEEE conference on computer vision and pattern recognition, pp 2901–2908. <https://doi.org/10.1109/CVPR.2014.371>
3. Babacan SD, Molina R, Do MN, Katsaggelos AK (2012) Bayesian blind deconvolution with general sparse image priors. Springer, Berlin, Heidelberg
4. Levin A, Weiss Y, Durand F, Freeman WT (2011) Efficient marginal likelihood optimization in blind deconvolution. In: The 24th IEEE conference on computer vision and pattern recognition, CVPR 2011, Colorado Springs, 20–25 June 2011
5. Cao X, Hao S, Xu L Single image deraining by fully exploiting contextual information. *Neural Process Lett* 1–15
6. Fu X, Huang J, Ding X, Liao Y, Paisley J (2017) Clearing the skies: a deep network architecture for single-image rain removal. *IEEE Trans Image Process* 26(6):2944–2956. <https://doi.org/10.1109/tip.2017.2691802>
7. Fernández-García M-E, Sancho-Gómez J-L, Ros-Ros A, Figueiras-Vidal AR Complete stacked denoising auto-encoders for regression. *Neural Process Lett* 1–11
8. Zhang K, Zuo W, Gu S, Zhang L (2017) Learning deep cnn denoiser prior for image restoration. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 2808–2817. <https://doi.org/10.1109/CVPR.2017.300>
9. Li Y, Yang Z, Hao T, Li Q, Liu W (2022) Pixel-level and perceptual-level regularized adversarial learning for joint motion deblurring and super-resolution. *Neural Process Lett*
10. Suin M, Purohit K, Rajagopalan AN (2020) Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3603–3612. <https://doi.org/10.1109/CVPR42600.2020.00366>
11. Huang W, Zhang L, Wang S, Wu H, Song A (2022) Deep ensemble learning for human activity recognition using wearable sensors via filter activation. *ACM Trans Embed Comput Syst*. <https://doi.org/10.1145/3551486>
12. Huang W, Zhang L, Wu H, Min F, Song A (2022) Channel-equalization-har: a light-weight convolutional neural network for wearable sensor based human activity recognition. *IEEE Trans Mobile Comput*. <https://doi.org/10.1109/TMC.2022.3174816>

13. Huang W, Zhang L, Teng Q, Song C, He J (2021) The convolutional neural networks training with channel-selectivity for human activity recognition based on sensors. *IEEE J Biomed Health Inform* 25(10):3834–3843. <https://doi.org/10.1109/JBHI.2021.3092396>
14. Couzinié-Devy F, Sun J, Alahari K, Ponce J (2013) Learning to estimate and remove non-uniform image blur. In: 2013 IEEE conference on computer vision and pattern recognition, pp 1075–1082. <https://doi.org/10.1109/CVPR.2013.143>
15. Sun J, Cao W, Xu Z, Ponce J (2015) Learning a convolutional neural network for non-uniform motion blur removal. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 769–777. IEEE Computer Society, Los Alamitos. <https://doi.org/10.1109/CVPR.2015.7298677>
16. Tang K, Xu D, Liu H, Zeng Z (2020) Context module based multi-patch hierarchical network for motion deblurring. *Neural Process Lett* 13:1–16
17. Cho S, Ji S, Hong J, Jung S, Ko S (2021) Rethinking coarse-to-fine approach in single image deblurring. [arXiv:2108.05054](https://arxiv.org/abs/2108.05054)
18. Nah S, Kim TH, Lee KM (2017) Deep multi-scale convolutional neural network for dynamic scene deblurring. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 257–265. <https://doi.org/10.1109/CVPR.2017.35>
19. Tao X, Gao H, Shen X, Wang J, Jia J (2018) Scale-recurrent network for deep image deblurring. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 8174–8182. <https://doi.org/10.1109/CVPR.2018.00853>
20. Wiener N (1964) *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications* (mit press)
21. Richardson WH (1972) Bayesian-based iterative method of image restoration*. *J Opt Soc Am* (1917–1983) 62(1):55–59
22. Boden AF, Redding DC, Hanisch RJ, Mo J (1996) Massively parallel spatially variant maximum-likelihood restoration of hubble space telescope imagery. *J Opt Soc Am A* 13(7):1537–1545
23. Yuan L, Sun J, Quan L, Shum HY (2008) Progressive inter-scale and intra-scale non-blind image deconvolution. *Acm Trans Graph* 27(3):1–10
24. Kaufman A, Fattal R (2020) Deblurring using analysis-synthesis networks pair. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 5810–5819. <https://doi.org/10.1109/CVPR42600.2020.00585>
25. Yuan Y, Su W, Ma D (2020) Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3552–3561. <https://doi.org/10.1109/CVPR42600.2020.00361>
26. Sun M, Purohit K, Rajagopalan AN (2020) Spatially-attentive patch-hierarchical network for adaptive motion deblurring. [arXiv:2004.05343](https://arxiv.org/abs/2004.05343)
27. Zhang H, Dai Y, Li H, Koniusz P (2019) Deep stacked hierarchical multi-patch network for image deblurring. *IEEE*
28. Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M-H, Shao L (2021) Multi-stage progressive image restoration. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 14816–14826. <https://doi.org/10.1109/CVPR46437.2021.01458>
29. Lai WS, Huang JB, Zhe H, Ahuja N, Yang MH (2016) A comparative study for single image blind deblurring. In: IEEE conference on computer vision and pattern recognition
30. Huang GB, Mattar M, Berg T, Learned-Miller E (2008) Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on faces in ‘real-life’ images: detection, alignment, and recognition
31. Everingham M, Eslami S, Van Gool L, Williams C, Winn J, Zisserman A (2015) The pascal visual object classes challenge: a retrospective. *Int J Comput Vis* 111(1):98–136. <https://doi.org/10.1007/s11263-014-0733-5>
32. Dong G, Jie Y, Liu L, Zhang Y, Shi Q (2017) From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR)
33. Zhang J, Pan J, Ren J, Song Y, Yang MH (2018) Dynamic scene deblurring using spatially variant recurrent neural networks. In: 2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR)
34. Li Y, Luo Y, Lu J (2021) Reference-guided deep deblurring via a selective attention network. *Appl Intell* 10:1–13
35. Zhang K, Luo W, Zhong Y, Ma L, Stenger B, Liu W, Li H (2020) Deblurring by realistic blurring. [arXiv](https://arxiv.org/abs/2004.05343)
36. Zhong Z, Gao Y, Zheng Y, Zheng B (2020) Efficient spatio-temporal recurrent neural network for video deblurring

37. Park D, Kang DU, Kim J, Chun SY (2020) Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In: European conference on computer vision
38. Wan S, Tang S, Xie X, Gu J, Huang R, Ma B, Luo L (2021) Deep convolutional-neural-network-based channel attention for single image dynamic scene blind deblurring. *IEEE Trans Circuits Syst Video Technol* (31-8)
39. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. [arXiv:1503.03832](https://arxiv.org/abs/1503.03832)
40. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. *CoRR* [arXiv:1804.02767](https://arxiv.org/abs/1804.02767)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.