IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 31, 2019, accepted January 14, 2020, date of publication January 20, 2020, date of current version January 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2967823

Scale-Iterative Upscaling Network for Image Deblurring

MINYUAN YE, DONG LYU, AND GENGSHENG CHEN
State Key Laboratory of ASIC and System, Fudan University, Shanghai 201203, China
Corresponding author: Gengsheng Chen (gschen@ fudan.edu.cn)

ABSTRACT Machine learning based methods for blind deblurring are efficient to handle real-world blurred images, whose blur may be caused by various combined distortions. However, existing multi-level architectures fail to fit images of various scenarios. In this paper, we propose a scale-iterative upscaling network (SIUN) that restores sharp images in an iterative manner. It is not only able to preserve the advantages of weights sharing across scales but also more flexible when training and predicting with different iterations to fit different images. Specifically, we bring in the super-resolution structure instead of the upsampling layer between two consecutive scales to restore a detailed image. Besides, we explore different curriculum learning strategies for both training and prediction of the network and introduce a widely applicable strategy to make SIUN compatible with different scenarios, including text and face. Experimental results on both benchmark datasets and real blurred images show that our method can produce better results than state-of-the-art methods. Code is available at https://github.com/minyuanye/SIUN.

INDEX TERMS Blind deblurring, curriculum learning, scale-iterative, upscaling network.

I. INTRODUCTION

Image deblurring, aiming to recover a sharp image from its blurred source, has long been a challenging and fundamental problem in computer vision and image processing. Single image deblurring is highly ill-posed. Existing deblurring methods on this issue can be classified into two major categories: traditional iterative optimization algorithms and learning-based methods.

Traditional methods simplify this problem by applying various constraints to characterize the image blur (e.g. uniform/non-uniform/depth-aware) and imposing different image priors [1]–[6] to regularize the solution space, which highly relies on the assumption model. However, in most real scenarios (e.g. nature/text/face), image blurs are caused by various distortions which are far more complex than the assumption model.

To pattern complicated natural blurs, learning-based methods are introduced. They are first proposed for non-blind deconvolution: Schuler *et al.* [7] follow a traditional two-step procedure but learn the second step on a large dataset of natural images through a neural network; Zhang *et al.* [8] train a fully convolutional network to deconvolve the blurred images iteratively in a multi-stage framework; Xu *et al.* [9] develop a deep convolutional neural network with two sub-modules. However, the above methods are still under the traditional framework.

With more advanced and complicated network models brought in, the end-to-end deep learning methods are proposed for blind deblurring. Nah *et al.* [10] propose an adversarial network for deblurring with a multi-scale generator. The finer scale image deblurring is aided by coarser scale features so that the latent image is obtained step by step from the $1/2^k (k = 2)$ scale to the original scale. However, using the coarse-to-fine mechanism directly via a scale-cascaded structure leads to excessive network size and depth. To improve it, Tao *et al.* [11] propose a scale-recurrent network (SRN) with long-short term memory (LSTM) for the network to share weights across scales, which significantly reduces the parameters of the network. As the multi-scale and scale-recurrent models cost expensive runtime, Zhang *et al.* [12] propose a deep multi-patch hierarchical network (DMPHN) which uses a multi-patch hierarchy as input and refines the whole image by the consecutive upper levels. All these state-of-the-art methods adopt the multi-scale framework with fixed levels, for both training and prediction. Therefore, their proposals show some inadequacies when applied to various real-world blurred images.

The associate editor coordinating the review of this manuscript and approving it for publication was Tomasz Trzcinski.

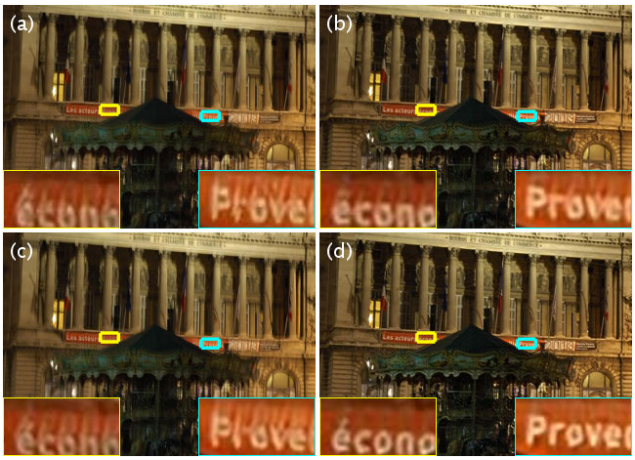


FIGURE 1. One real example. (a) Result of Nah *et al.* [10]. (b) Result of Tao *et al.* [11]. (c) Result of Zhang *et al.* [12]. (d) Our result.

In this paper, we propose a scale-iterative upscaling network (SIUN). It starts from a down-sampled scale and works in an iterative way. For each iteration, the output is up-scaled until a full resolution image gets restored. There are several advantages of this new framework: (1) Like scale-recurrent structure, it has far fewer trainable parameters through weights sharing than scale-cascaded structure; (2) Compared with previous fixed-level architecture, it is more flexible when training and predicting with variable iterations to fit different images; (3) Instead of the upsampling layer used in [10] and [11], it adapts residual dense network (RDN) [13], a super-resolution architecture for upscaling so that more details of the image can be restored; (4) It is more compatible with diverse scenarios of real-world images, including text and face, with the curriculum learning strategy designed for both training and prediction. A visual example compared with [10]–[12] is shown in Fig. 1. Our major contributions are as follows:

1) We propose a novel scale-iterative architecture for image deblurring. Compared with previous fixed-level architectures, our network is more flexible by applying different iterations for training and prediction, with shared weights across scales.
2) We propose an upscaling network that adapts super-resolution structure instead of the upsampling layer used in previous works so that the essential features of the down-sampled deblurring image can be preserved while upscaling between iterations to restore more detailed images.
3) In both training and prediction, an appropriate increment of iterations can decrease recovery difficulty by reducing blur magnitude between consecutive iterations. We explore different curriculum-based strategies for training and prediction, then adopt the best 3-iterations strategy for training and the widely applicable 4-iterations strategy for prediction. Experimental results show that a well-designed curriculum learning strategy can bring with a faster convergence and is more efficient to deal with images of various scenarios, including text and face.

II. RELATED WORK

A. ITERATIVE & MULTI-LEVEL APPROACH FOR IMAGE DEBLURRING

Iterative approach is widely used in traditional methods. Fergus *et al.* [1] and Shan *et al.* [2] introduce a coarse-to-fine strategy and the typical MAP (Maximum a Posteriori) framework in traditional deblurring. With their pioneer works, almost all the traditional energy-optimization-based methods [3]–[6], [14]–[17] deal with dynamic deblurring iteratively, which proves to be efficient: optimizing from a down-sampled scale, upscaling gradually between iterations until reaching the original scale. Early learning-based methods use neural networks in the deblurring process, such as blur kernel prediction [18]–[20] and non-blind deconvolution [7]–[9], or just as an image prior [21], [22]. However, these methods are still under the traditional iterative-optimization framework, which highly relies on the assumption model.

More recent learning-based works restore sharp images in an end-to-end manner, where the multi-level approach is very efficient. Hradiš *et al.* [23] propose a CNN focused on the restoration of text documents and Kupyn *et al.* [24] train a generative adversarial network (GAN) for image deblurring. Their results are suboptimal due to flat architectures. Nah *et al.* [10] propose a deep CNN based on multi-scale architecture (as shown in Fig. 2(a)) for dynamic scene deblurring. However, using coarse-to-fine mechanism directly via scale-cascaded structure leads to excessive network size and depth. Thus, Tao *et al.* [11] improve this architecture by sharing weights across scales with a long short-term memory (LSTM)-based scale-recurrent network (Fig. 2(b)). As the multi-scale and scale-recurrent models cost expensive runtime, Zhang *et al.* [12] propose a real-time model named deep hierarchical multi-patch network (DMPHN) (Fig. 2(c)). These methods all use multi-level architecture and start deblurring from a tiny scale. However, the multi-level architecture is not flexible enough as their levels are not configurable after the networks designed, thus cannot be applied to images of various scenes or different blur magnitudes. To solve this issue, we propose a scale-iterative upscaling architecture (Fig. 2(d)) that combines multi-level with iterative approach.

B. IMAGE STRUCTURES & DETAILS RESTORATION

1) ENCODER-DECODER NETWORK

The encoder-decoder architecture is a neural network design pattern. It is prevalent in neural machine translation and sequence-to-sequence prediction [25]. Recently, it also shows success in various computer vision tasks [26]–[29]. As its name suggests, it is partitioned into two symmetric parts: the encoder and the decoder. In an image-to-image task,

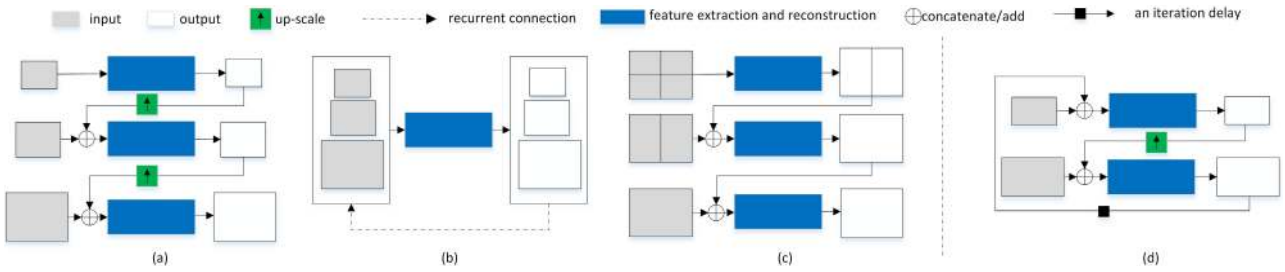


FIGURE 2. Networks architecture comparison. (a) Scale-cascaded architecture of Nah *et al.* [10]. (b) Scale-recurrent architecture of Tao *et al.* [11]. (c) Hierarchical multi-patch architecture of Zhang *et al.* [12]. (d) Scale-iterative architecture of our network.

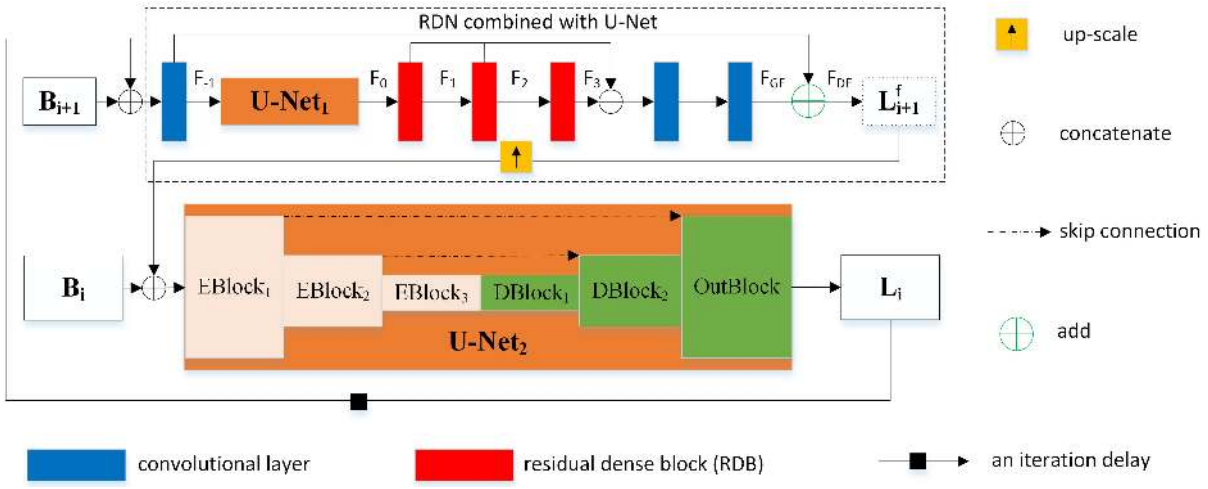


FIGURE 3. Details of our proposed scale-iterative upscaling network (SIUN) architecture.

the encoder maps an input image to a feature space, subsequently the decoder takes this feature map as an input and maps it to an output image. Ronneberger *et al.* [30] add skip connections between corresponding feature maps in encoder and decoder to improve its regression ability, which is called U-Net. The encoder-decoder networks perform well for image deblurring: Kupyn *et al.* [24] present an encoder-decoder network as the generator based on conditional GAN and Tao *et al.* [11] use the U-Net structure with ResBlocks [10] in a scale-recurrent way. Similarly, we adopt the encoder-decoder network for image structures restoration.

2) SUPER-RESOLUTION (SR) NETWORK

It is well known that deblurring at a down-sampled scale is much easier as the blur magnitude decrease. However, image details are also lost during downsampling. When it comes to upscaling, previous works [10], [11] use a simple upsampling layer, which is not enough to preserve and recover details information. To solve this problem, we use the structure of SR to replace the upsampling layer, aiming to reconstruct a high-resolution (HR) image from its low-resolution (LR) counterpart so that the high frequency information can be well mapped to the next scale level. In this paper, we select to use RDN for details restoration, as it makes full use of both global and local hierarchical features.

C. CURRICULUM LEARNING

Curriculum learning means learning by gradually increasing the difficulty of the tasks. Elman [31] and Bengio *et al.* [32] have shown the positive effects of curriculum learning for network performance in several tasks. More recent researches [33], [34] show that its effectiveness is highly sensitive with respect to the modality of progression through the tasks. Different from [10]–[12], which directly reconstruct latent images through the multi-level network, our SIUN reconstructs latent images starting from down-sampled scale and restore the images to its original size through several iterations gradually. Thus, a well-designed curriculum learning strategy is of a good help to our network. We study different strategies with various iterations for both training and prediction when designing SIUN, and finally adopt the most suitable curriculum strategy to fit diverse scenes of blurred images. The curriculum containing easy-to-hard decisions can be settled for one query to gradually restore the corrupted blurry image.

III. NETWORK ARCHITECTURE

Fig. 3 gives the architecture of our proposed scale-iterative upscaling network (SIUN). The SIUN has mainly two levels. The first (upper) level is constructed by using a modified RDN, combined with a U-Net (U-Net$_1$). The first level

performs a typical deblurring operation at a relatively smaller image scale. Then the output of the first level is up-scaled before fed to the second (lower) level. In such a design, the image scale obtains a double in size while at the same time the details of the deblurring results retain preserved. We use another U-Net (U-Net$_2$) to construct the second level. Although the operating scale in the second level is larger with a higher blur magnitude, with the aid of the first level, the augment of its complexity is not salutatory but gradual. With these two levels working together, a sharp image can be obtained.

By iteratively repeating the above processing, the produced sharp image has its size kept on being up-scaled until reaching full resolution. This full resolution image is taken as the final output. There are two indispensable parts to enforce our curriculum learning strategy: (1) **upscaling network** (to learn a series of tasks with gradually increasing difficulty as the blur magnitude decrease when downsampling) and (2) **scale-iterative structure** (to achieve iterative process).

A. UPSCALING NETWORK (UN)

The two-level upscaling network consists of two main parts: the encoder-decoder network for image-structure restoration and the upscaling structure for image-detail restoration. We use the Encoder ResBlocks (EBlocks) and Decoder ResBlock (DBlocks) that proposed by [11] to construct our U-Net, as shown in Fig. 3. U-Net$_1$ has the same architecture of U-Net$_2$, except that it does not contain OutBlock. This is because U-Net$_1$ is designed for features transformations and U-Net$_2$ is for image transformations. They can be expressed as follows:

$$\begin{aligned} F_0 &= UNet_1(F_{-1}) \\ L_i &= UNet_2([B_i, L_{i+1}^{\uparrow}]) \end{aligned} \tag{1}$$

where F_{-1} is the shallow feature extracted from the first-level's input image and F_0 is the feature reconstructed by U-Net$_1$. At the second level, U-Net$_2$ takes the pyramid image B_i and the up-scaled image L_{i+1} from the first level as its input to reconstruct the second level's output image L_i. These two U-Nets promise the restoration of the image structures.

As for upscaling, [10] and [11] use an upsampling layer to scale up the image generated by the previous level. Differently, we apply an RDN-based super-resolution architecture for a better reserving of the features from previous level. After reconstructing features F_0 by U-Net$_1$, we further extract hierarchical features with a set of RDBs (residual dense blocks), then conduct a dense feature fusion to get F_{DF}. The first-level network can be expressed as follows:

$$L_{i+1}^{\uparrow} = H_{RDN}([B_{i+1}, L_{i+1}]) \tag{2}$$

where B_{i+1} is the pyramid image and L_{i+1} is the output image of the previous iteration. Thus, the two-level upscaling network can be represented as follows:

$$\begin{aligned} L_i &= UN([[B_{i+1}, L_{i+1}], B_i]) \\ &= UNet_2([B_i, H_{RDN}([B_{i+1}, L_{i+1}])]) \end{aligned} \tag{3}$$

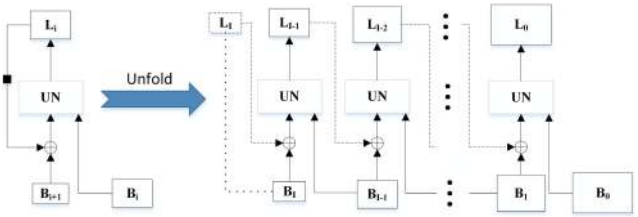


FIGURE 4. Unfolding our SIUN to express an I iterations deblurring processing.

where L_i is the output image at i-th iteration. The RDN promises the restoration of the image details. Therefore, by such a specific design, the whole two-level upscaling network are able to restore both the main structures and details of the blurred image.

B. SCALE-ITERATIVE STRUCTURE

We use a scale-iterative structure in our proposed SIUN network. For an I-iterations deblurring processing, as shown in Fig. 4, the SIUN can be unfolded to a very deep network. We first generate a set of pyramid blurry images $B_i (i = I, I-1, \ldots, 1, 0)$, with B_i denoting a $1/2^i$ down-sampled image, and then start the deblurring processing iteratively from the smallest scale $i = I$. At the very beginning, we assume $L_I = B_I$ as the 0^{th} iteration's output. In the first iteration, we use L_I together with B_I as the first level's inputs, and use B_{I-1} as the second level's input.

In the second iteration, L_{I-1} (the first iteration's output) and B_{I-1} are used together as its first level's inputs, and B_{I-2} is used as its second level's input to produce its output L_{I-2}. By repeating this processing with I iterations, we are able to get the final output L_0 of the original size, and we denote the final restored image L_0 as L^I. We use the subscript to describe the intermediate results of different scales and use the superscript to denote the final full resolution result. Thus, for any i-th iteration, the network can be described as follows:

$$L_i = UN([[B_{i+1}, L_{i+1}], B_i]), \quad i = I-1,\ I-2, \ldots, 1, 0 \tag{4}$$

where we assume $L_I = B_I$ for the first iteration. Such kind of the scale-iterative structure promises an easy-to-hard restoration and meanwhile extend the compatibility of our SIUN.

C. CURRICULUM LEARNING STRATEGY

According to (4), for a blurry image B, we can divide the deblurring task into I sub-tasks. Each sub-task is targeted to obtain a sharp image (L_i) from the pyramid blurry images (B_{i+1}, B_i) and the predicted image of previous sub-task (L_{i+1}). The difficulty of the sub-tasks increases gradually, because blur magnitude decreases when the blurry image is down-sampled.

With regard to training, previous works tend to use the multi-scale architectures [10], [11] and take the whole deblurring as one single task. They take the output of each level into

consideration and sum up them with a weight factor:

$$Loss = \sum_{i=1}^{3} w_i \cdot f_i(B, G)$$

where w_i is weight factors, $f_i(B, G)$ is the i-th level loss, B is the blurry image, and G is the ground truth. Different from their methods, our scale-iterative architecture takes each sub-task as an independent component within one unified task. In our design, the output image L_i of i-th iteration is a restored image at a certain scale, thus our training process can end at any iteration while still having an effective trained network. Therefore, each iteration is considered independently in our loss function. We choose L_1 loss to optimize our proposed network:

$$Loss = \frac{\|L_i - G_i\|}{N_i}$$

where L_i is the output of the network, G_i is the ground truth, and N_i is the number of elements in L_i to normalize at the i-th scale.

As for prediction, we can also apply the curriculum strategy although the network does not "learn" when predicting. The same as what we do in training, we divide the deblurring task into I sub-tasks, start the prediction from the I^{th} down-sampled scale, conduct the deblurring processing of the up-scaled image until the original scale image is predicted. Furthermore, the selection of the number I can be different for various images. Intuitively, we may use fewer iterations for an image with mild blur and use more iterations for an image with severe blur. We study the relationship between blur magnitude and the number of I for both training and prediction, to reach a more efficient and balanced curriculum learning strategy for SIUN. More detailed studies are given in Section IV-A and IV-B.

IV. EXPERIMENTS

Training, Validation, And Testing Datasets: Nah *et al.* [10] propose a new large-scale dataset that provides blurry/sharp pairs of realistic images, named GOPRO. This dataset is captured by GOPRO4 Hero Black camera and is composed of 3214 pairs of blurry and sharp images at 1280 × 720 resolution. Different from early works [18], [19], [35] convolving sharp images with blur kernels, the blurry images in GOPRO are generated by averaging consecutive short-exposure frames to approximate long-exposure blurry frames. The same as [10]–[12], we use 2,103 pairs for training (training dataset) and the remaining 1,111 pairs for validation (validation dataset). Then we apply our trained model on Köhler dataset [36] (testing dataset), which is widely used by both traditional methods and learning-based methods, for further performance evaluation.

Training Details: For training, we first design a curriculum learning strategy for the purpose to determine the number of I (total number of iterations). Several pure-iteration strategies $I = 2$, $I = 3$, and $I = 4$ are conducted to study the performance influence of using different iterations in running our model. We further explore the mixed-iteration strategies $I = 2, 3$ (trained for $I = 2$ and $I = 3$ alternately between two epochs) and $I = 2, 3, 4$. We use Adam [37] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$ for all our experiments. In each iteration, we randomly crop a 256 × 256 patch from each original image as the training input with a batch-size of 16. The models for all the above strategies are trained with learning rate scheduler of $\{1e^{-4}, 3e^{-5}\}$ until convergence. Then the model with the best strategy (detail discussion is given in Section IV-A) is further trained with learning rate scheduler of $\{5e^{-6}, 1e^{-6}\}$. All these models are implemented with Keras framework (TensorFlow backend) and trained on NVIDIA TITAN Xp GPUs.

TABLE 1. Results of the models trained with different curriculum learning strategies.

Strategies	Pure-iteration			Mixed-iteration	
	I=2	I=3	I=4	I=2,3	I=2,3,4
PSNR	29.48	**30.12**	29.71	29.41	29.33

A. CURRICULUM LEARNING STRATEGY FOR TRAINING

We design five different curriculum learning strategies $I = 2$, $I = 3$, $I = 4$, $I = 2, 3$ and $I = 2, 3, 4$ for the training of our SIUN, comparing their restoration results on validation dataset, as shown in Table 1, where we make two main observations.

First, for pure-iteration strategies, an appropriate increment of iterations does help the model to obtain better restoration results ($I = 2$ vs. $I = 3$). We believe the reason is that, as the number of iterations increases, the difficulty of restoration in each iteration decreases, which makes the deblurring task easier to learn and helps the model to converge. However, increasing iteration excessively may lead to a worse performance ($I = 3$ vs. $I = 4$). It may be due to two reasons: (1) As iteration number increases, the pyramid images fed to the model will become too small to carry enough information and therefore to have little effect on optimizing the model; (2) More iterations may require a larger model and bigger patch size for training, and even a larger dataset.

Second, mixed-iteration strategies have obtained worse performances than pure-iteration ones. We can see that the performance of strategy $I = 2, 3$ is worse than those of $I = 2$ and $I = 3$. And strategy $I = 2, 3, 4$ has a similar case. In addition, the more different iterations got mixed for training, the worse the performance is ($I = 2, 3$ vs. $I = 2, 3, 4$). These experimental results indicate that changing iteration number during training can interfere with the model's convergence. In other words, a frequent change of the learning difficulty during training will impede the convergence of the model.

From the above, we can see that using the strategy $I = 3$ helps to reach the best training performance. Therefore, we select to use strategy $I = 3$ to further train our model, by a learning rate scheduler of $\{5e^{-6}, 1e^{-6}\}$, and use the trained model to study the curriculum strategy for prediction.

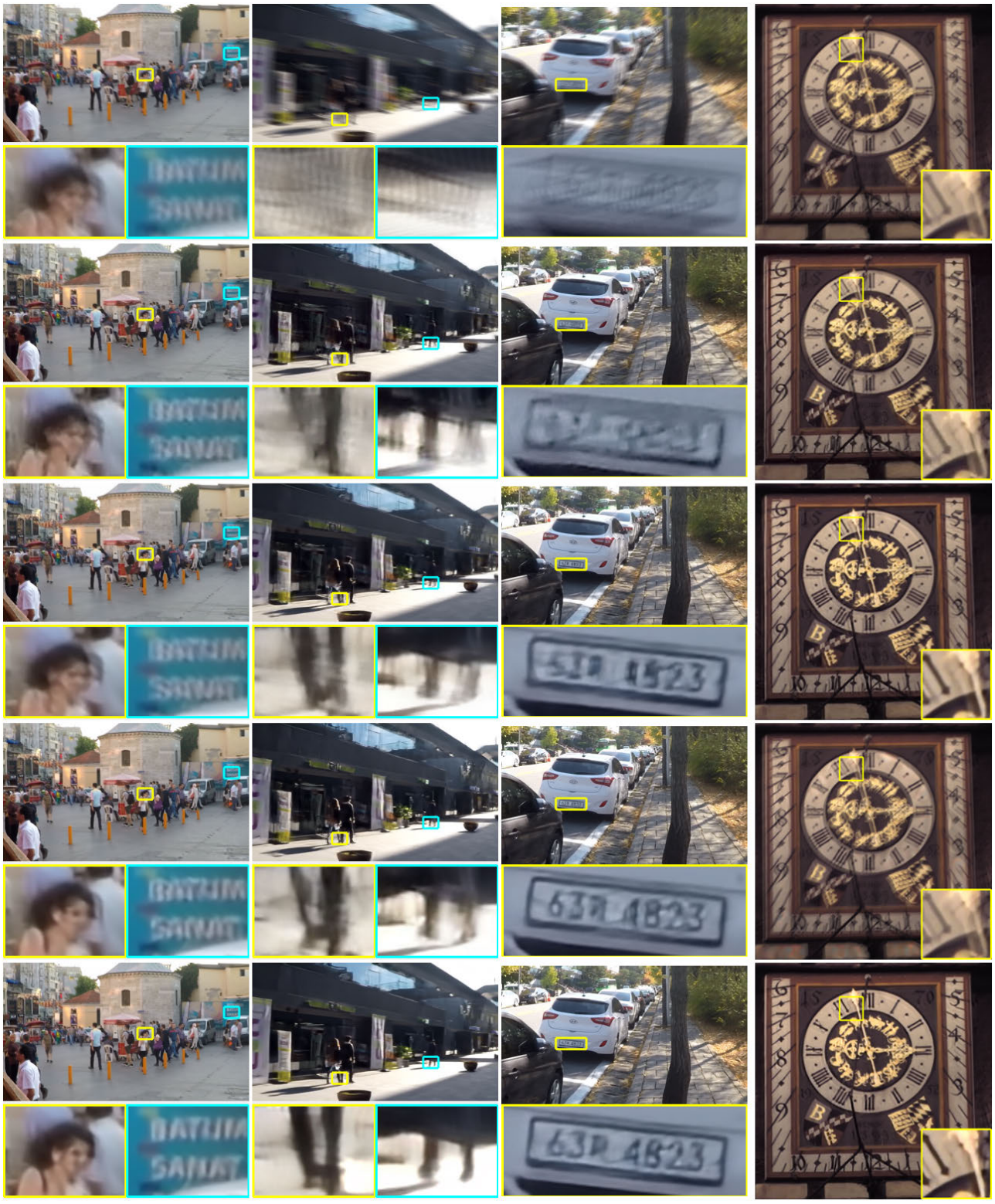


FIGURE 5. Visual comparison on benchmark datasets. From top to bottom are blurry input, deblurring results of Nah *et al.* [10], Tao *et al.* [11], Zhang *et al.* [12] and our results.

B. CURRICULUM STRATEGY FOR PREDICTION

On predicting a blurred image, by using our model, different latent image L^I can be obtained with different iteration number I. Thus, the curriculum strategies also affect the prediction results. We study the strategies $I = 1$, $I = 2$, $I = 3$, and $I = 4$ respectively on validation dataset, with the results shown in Table 2.

Note that for each blurry image, we can obtain four full resolution restored images L^i, $i = 1, 2, 3, 4$ with different iteration numbers. The "Best of all" is determined by

TABLE 2. Prediction results by applying different strategies.

Strategies	I=1	I=2	I=3	I=4	Best of all
Mean PSNR	29.21	30.08	30.21	30.22	30.28

TABLE 3. Statistics of blurry images which achieve the best results at different iterations.

Strategies	I=1	I=2	I=3	I=4
Numbers	102	224	343	442
Mean PSNR	29.89	26.82	25.01	24.56
Max PSNR	38.36	35.92	32.85	35.43
Min PSNR	23.86	20.32	18.33	18.70

calculating the highest PSNR of these four images. With different strategies, we can see that as iteration number increases, the mean PSNR value on the whole dataset also increases. However, different images may achieve their best quality with different iteration numbers. We try to find out the criteria that affects the selection of the iteration number. Table 3 presents the statistical results of our experiment.

First, we study two parameters in Table 3: "Numbers" and "Mean PSNR". By using strategy $I = 1$, there are 102 images achieving the best restoration results with mean PSNR of 29.89. Whereas using strategy $I = 4$, there are 442 images achieving the best restoration results with mean PSNR of 24.56. To achieve the best restoration, the required iteration number increases as the Mean PSNR decreases. In other words, more iterations are required for severely blurred images to achieve the best restorations.

Second, we study the parameters "Max PSNR" and "Min PSNR" in Table 3. We can see that, for all the 4 strategies, they each has a wide range of PSNR values. It implies that an image of poor quality may need just one iteration to achieve the best result, whereas an image of high quality may require multiple iterations. It shows that, for a specific image, there does not exist a certain relationship between the quality of the image and the required iteration numbers.

Finally, from Table 2 and Table 3, we can see that the 4-iterations (I = 4) strategy is the most applicable when using our iterative model for image deblurring. Note that our model is trained with 3-iterations (I = 3) strategy. So, benefited from our scale-iterative architecture, we can flexibly apply different iteration strategies for training and prediction.

C. COMPARISONS

We compare our method with state-of-the-art works on two benchmark datasets (GOPRO dataset and Köhler dataset) and on real blurred images (Lai dataset [38]) as well. For learning-based methods, we choose [10],[1] [11],[2] and [12][3] for comparisons, who have provided their source codes and models. For traditional methods, we choose [16][4] as a representation for our additional comparisons on real images, since it performs the best on real dataset among the state-of-the-art methods according to [38].

TABLE 4. Deblurring results on benchmark datasets. *Size* and *Runtime* are expressed in MB and millisecond. The best performance is shown in red and the second-best is in blue.

Method	GOPRO	Köhler	Size	Runtime[a]
	PSNR / SSIM	PSNR / MSSIM		
Nah *et al.* [10]	28.49 / 0.8543	25.44 / 0.7996	303.6	2704 ± 26
Tao *et al.* [11]	30.25 / 0.9030	26.57 / 0.8373	27.5	355.6 ± 4.5
Zhang *et al.* [12]	30.45 / 0.9057	24.21 / 0.7562	27.6	35.5 ± 7.5
SIUN-4I[b]	30.22 / 0.9041	26.90 / 0.8501	24.5	367.9 ± 11.2
SIUN-VI[c]	30.28 / 0.9046	26.99 / 0.8551		—

[a]Results of Nah *et al.* are tested with GeForce GTX 1070 GPU while others are tested with TITAN Xp GPU.
[b]SIUN with 4-iterations strategy.
[c]SIUN with variable-iterations strategy.

Our model produces a full resolution sharp image for each number of iteration, labeled as $\{L^i, i = 1, 2, 3, 4\}$. The image of the highest PSNR among them is selected and evaluated as the "variable-iterations" results on benchmark datasets. In other words, "variable-iterations" has the best restoration performance for this specific original blurred image. We also present the result, shown in Table 4, under the most applicable 4-iterations strategy. All the images shown for visual comparison are generated under this strategy as well.

1) BENCHMARK DATASETS

Table 4 shows our deblurring results on benchmark datasets. All the images evaluated in our experiment are in RGB mode unless otherwise stated. Thus, among the three models (lstm/gray/color) released by Tao *et al.* [11], we use their "color" model for comparison. Zhang *et al.* [12] release models with different hierarchies and we use their best (1-2-4-8) DMPHN model. The PSNR and SSIM metrics on GOPRO dataset are calculated by using MATLAB built-in function "psnr()" and "ssim()" based on the generated color results (SSIM is calculated in grayscale). The PSNR and MSSIM metrics on Köhler dataset are calculated by the executable provided by [36] .[5]

From Table 4, we can see that on GOPRO dataset, our model has a comparable result with Tao's but is better than Nah's. As on Köhler dataset, our outcome is better than that of both. Zhang's result on GOPRO dataset is the best, however, his result on Köhler dataset is much worse. We believe the reason is that his hierarchy approach is highly bounded to a specific image size of 1280×720, therefore not that applicable to images with different sizes. It can also explain why his model can reach 30fps@720p.

When we use another different dataset (Köhler dataset) for further evaluation, our model shows a much better

[1]https://github.com/SeungjunNah/DeepDeblur_release
[2]https://github.com/jiangsutx/SRN-Deblur
[3]https://github.com/HongguangZhang/DMPHN-cvpr19-master
[4]https://eng.ucmerced.edu/people/zhu/CVPR14_text_code_blind.zip
[5]http://people.kyb.tuebingen.mpg.de/rolfk/BenchmarkECCV2012/evaluation_code.zip

FIGURE 6. Deblurring results on real-world blurred images from Lai [38] dataset. From top to bottom are images restored by Pan *et al.* [16], Nah *et al.* [10], Tao *et al.* [11], Zhang *et al.* [12] and ours. As space limits, the original blurry images are omitted here. They can be viewed in Lai dataset with their names, from left to right: boy_statue, pietro, street4 and text1.

compatibility on different scenarios, which can be further proved by our experimental results on real blurred images (referred to our discussion in the next paragraph). In addition, as our model works in an iterative way, it shares weights across iterations to keep a small size. Although the most applicable strategy requires 4 iterations for deblurring an image, our model has a very close performance with that of Tao's in runtime, since it runs fast at small scales. All these learning-based methods can produce acceptable results of images' major structures, while our model can generate more explicit details and sharper structures. Visual examples are shown in Fig 5.

2) REAL BLURRED IMAGES

Although GOPRO dataset can simulate real-world blur well, it is synthesized from high-speed cameras. Köhler dataset is a real-world database, while it only contains four different scenes. Thus, we further test our model on the real dataset collected by Lai *et al.* [38], which contains 100 real-world blurred images of different scenes. We compare our method with that of Pan *et al.* [16], Nah *et al.* [10], Tao *et al.* [11], and Zhang *et al.* [12]. The visual comparisons are shown in Fig. 6, where we can see that our method can produce more clearly restored images for different scenarios, including face and text, with less artifact and more details. It surpasses the state-of-the-art works of both traditional and learning-based methods and shows wide compatibility to diverse scenes as well.

V. CONCLUSION

In this paper, we propose a scale-iterative upscaling network (SIUN) with weights sharing across iterations. In comparison with previous fixed-level structures for blind deblurring, our model is more flexible when applying to images of different sizes and scenarios by using variable iterations. The use of RDN for upscaling enables our network to restore blurred images with more sharp details. Also, we investigate the curriculum-based strategies for both training and prediction, then present the best strategy for training and the most applicable strategy for prediction, which extend the compatibility to deal with images of different scenarios, including text and face. Experimental results show that our method can produce better results on both benchmark datasets and real-world blurred images, compared with both traditional and learning-based methods.

REFERENCES

[1] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 787–794, 2006.

[2] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," *ACM Trans. Graph.*, vol. 27, no. 3, p. 73, 2008.

[3] S. Cho and S. Lee, "Fast motion deblurring," *ACM Trans. Graph.*, vol. 28, no. 5, p. 145, 2009.

[4] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce, "Non-uniform Deblurring for Shaken Images," *Int. J. Comput. Vis.*, vol. 98, no. 2, pp. 168–186, Jun. 2012.

[5] M. Hirsch, C. J. Schuler, S. Harmeling, and B. Schölkopf, "Fast removal of non-uniform camera shake," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 463–470.

[6] O. Whyte, J. Sivic, and A. Zisserman, "Deblurring shaken and partially saturated images," *Int. J. Comput. Vis.*, vol. 110, no. 2, pp. 185–201, Nov. 2014.

[7] C. J. Schuler, H. Christopher Burger, S. Harmeling, and B. Scholkopf, "A machine learning approach for non-blind image deconvolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1067–1074.

[8] J. Zhang, J. Pan, W.-S. Lai, R. W. Lau, and M.-H. Yang, "Learning fully convolutional networks for iterative non-blind deconvolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3817–3825.

[9] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1790–1798.

[10] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3883–3891.

[11] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8174–8182.

[12] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multi-patch network for image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5978–5986.

[13] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[14] L. Xu, S. Zheng, and J. Jia, "Unnatural L_0 sparse representation for natural image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1107–1114.

[15] D. Krishnan, T. Tay, and R. Fergus, "Blind deconvolution using a normalized sparsity measure," in *Proc. CVPR*, Jun. 2011, pp. 233–240.

[16] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, "Deblurring text images via L_0-regularized intensity and gradient prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2901–2908.

[17] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, "Deblurring face images with exemplars," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 47–62.

[18] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Scholkopf, "Learning to Deblur," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1439–1451, Jul. 2016.

[19] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 769–777.

[20] K. Schelten, S. Nowozin, J. Jancsary, C. Rother, and S. Roth, "Interleaved regression tree field cascades for blind image deconvolution," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 494–501.

[21] S. Sreehari, S. V. Venkatakrishnan, B. Wohlberg, G. T. Buzzard, L. F. Drummy, J. P. Simmons, and C. A. Bouman, "Plug-and-play priors for bright field electron tomography and sparse interpolation," *IEEE Trans. Comput. Imaging*, vol. 2, no. 4, pp. 408–423, 2016.

[22] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep cnn denoiser prior for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3929–3938.

[23] M. Hradiš, J. Kotera, P. Zemcık, and F. Šroubek, "Convolutional neural networks for direct text deblurring," in *Proc. BMVC*, 2015, vol. 10, p. 2.

[24] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8183–8192.

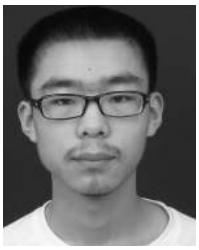
[25] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*. [Online]. Available: https://arxiv.org/abs/1406.1078

[26] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4463–4471.

[27] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1279–1288.

[28] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4472–4480.

[29] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[31] J. L. Elman, "Learning and development in neural networks: The importance of starting small," *Cognition*, vol. 48, no. 1, pp. 71–99, Jul. 1993.
[32] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. ACM 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.
[33] S. Reed and N. De Freitas, "Neural programmer-interpreters," 2015, *arXiv:1511.06279*. [Online]. Available: https://arxiv.org/abs/1511.06279
[34] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwi ska,, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, A. P. Badia, K. M. Hermann, Y. Zwols, G. Ostrovski, A. Cain, H. King, C. Summerfield, P. Blunsom, K. Kavukcuoglu, and D. Hassabis, "Hybrid computing using a neural network with dynamic external memory," *Nature*, vol. 538, no. 7626, pp. 471–476, Oct. 2016.
[35] A. Chakrabarti, "A neural approach to blind motion deblurring," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 221–235.
[36] R. Köhler, M. Hirsch, B. Mohler, B. Schölkopf, and S. Harmeling, "Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 27–40.
[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[38] W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, and M.-H. Yang, "A comparative study for single image blind deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1701–1709.



MINYUAN YE received the B.S. degree in microelectronic science and engineering from Fudan University, China, in 2014, where he is currently pursuing the Ph.D. degree with the State Key Laboratory of ASIC and System. His research interests include computer vision, machine learning, and embedded software.

DONG LYU received the B.S. degree in microelectronic science and engineering from Fudan University, China, in 2019, where he is currently pursuing the degree with the State Key Laboratory of ASIC and System. His research interests include computer vision, machine learning, and video codec.

GENGSHENG CHEN received the master's and Ph.D. degrees in electronic engineering from Fudan University, China. He was with Motorola Electronics, Nortel Networks, Siemens Technical Innovation Center, and Zarlink Semiconductor, from 1994 and 2005. He is currently a Senior Research Engineer with the State Key Laboratory of ASIC and System, Fudan University. He holds five Chinese patents. He has published over 30 articles in international academic conferences and journals. His major research interests include image processing, embedded systems, and FPGA circuits and systems.