Mixer-based Local Residual Network for Lightweight Image Super-resolution

Garas Gendy[1], Nabil Sabor[2], Jingchao Hou[1], Guanghui He[1,*]
[1]Micro-Nano Electronics Department, Shanghai Jiao Tong University, Shanghai 200240, China.
[2]Electrical Engineering Department, Faculty of Engineering, Assiut University, Assiut 71516, Egypt
garasgaras@yahoo.com,nabil_sabor@aun.edu.eg,{jingchaohou,guanghui.he}@sjtu.edu.cn

Abstract

Recently, the single image super-resolution (SISR) based on deep learning algorithm has taken more attention from the research community. There are many methods that are developed to solve this task using CNNs methods. However, most of these methods need large computational resources and consume more runtime. Due to the fact that the runtime is essential for some applications, we propose a mixer-based local residual network (MLRN) for lightweight image super-resolution (SR). The idea of the MLRN model is based on mixing channel and spatial features and mixing low and high-frequency information. This is done by designing a mixer local residual block (MLRB) to be the backbone of our model. Moreover, the bilinear up-sampling is utilized to transfer and mix low-frequency information with extracted high-frequency information. Finally, the GELU activation is used in the main model, proving its efficiency for the SR task. The experimental results show the effectiveness of the model against other state-of-the-art lightweight models. Finally, we took part in the Efficient Super-Resolution 2023 Challenge and achieved good results.

1. Introduction

The image super-resolution (SR) task is one of the computer vision sub-topics. This SR task focuses on converting the input image with low resolution to another one with high-resolution and clearer details [2, 11, 41]. This task is considered an ill-posed task because there a variety of the output images that can be generated from one input image. Also, this task can be divided into two main types SISR and multiple image super-resolution (MISR) [18]. In practice, the SISR has taken much attention based on the application of this task [2, 11, 41].

Recently, many methods are proposed to solve this task using convolution neural networks (CNNs) [2, 11, 41]. These models can be divided into efficient [28, 42, 49] and classical models [5, 25]. efficient models are needed in applications that focus on computational resource and memory consumption. So, to push the progress for the efficient SR model. Many efficient, super-resolution competitions are developed based on using advanced deep-learning algorithms [23, 46, 47].

The first competition that focuses on designing an efficient model is AIM 2019 challenge [47], which based on finding the model that optimizes five factors of (runtime, parameters, FLOPs, activations, and depths). The winning model of this competition is the information multi-distillation network (IMDN) [16] which is based on cascaded information multi-distillation blocks (IMDB) for the feature extraction module. After that, AIM 2020 challenge is organized to find a more efficient model [46] using the same evaluation criteria as the previous one. The winning solution of this challenge is the residual feature distillation network (RFDN) [26], which further improves the IMDN based on introducing residual learning in the main block. This RFDN achieved more efficient performance compared to the previous one in the evaluation criteria. Then, NTIRE 2022 challenge [23] is made to find a more efficient model than the previous two competitions. The winning solution of this challenge is the residual local feature network (RLFN) [20] which achieves a much faster runtime compared to the previous competition.

Lastly, the NTIRE 2023 Efficient Super-Resolution challenge [24] is organized as an extension of previous challenges. In this challenge, we aimed further to improve the winning solution of the previous challenge RLFN [20] to decrease its weight and reduce the runtime. To do that, one 3×3 layer of the RLFN main lock is replaced with a convolution mixer block. This ConvMix block is based on using depthwise convolution and pointwise convolution for channel and spatial feature mixing. Based on this ConvMix, the mixer local residual block (MLRB) is designed as the main block for our model. Using this MLRB, the mixer local residual network (MLRN) is built for the SR task. In addition, the GELU activation function is used instead of the original LeakyReLU, which proof is more helpful for the SR task. Finally, inspired by the previous methods, bilinear up-sampling added to transfer low-frequency information

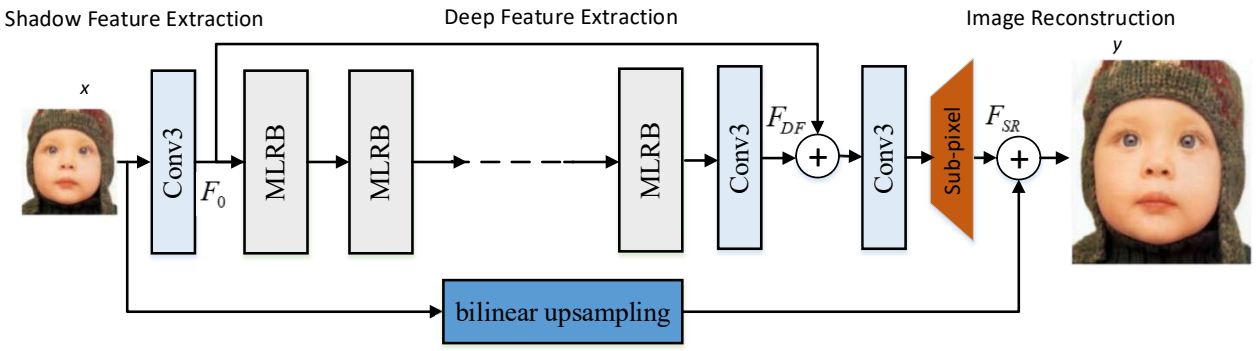

Figure 1. The structure of the proposed MLRN

of the input image to the final layer.

The paper contribution is summarised as the following :

- We propose MLRB as a main backbone block of the SR network based on using ConvMix layer, which represents a low computations bock,
- Based on the designed MLRB and the bilinear upsampling, an efficient MLRN model is proposed for the SISR.
- The developed model achieved state-of-the-art on the SR benchmark with a good balance between performance and computation. Then, an ablation study is made, aiming to show the impact of each network component.

2. Related Work

The related work of this paper will include CNNs-based efficient SR, attention-based efficient SR, and extreme efficient-based SR.

2.1. CNNs-based Efficient SR

In [28], Luo et al. developed a multi-scale receptive field fusion network (MRFN) based on utilizing a multi-scale receptive field fusion block for the fusion of different scales of spatial information and increasing the receptive field. After that, in [42], the authors introduced a progressive representation recalibration network (PRRN) [42] based on the learning of full and effective feature representations using a progressive representation recalibration block (PRRB). Then, an asymmetric information distillation network (AIDN) [49] is suggested by basing on the idea that repeating the process of distilled information can help the ability of key information extraction.

A lightweight super-resolution network named DDistill-SR [39] is developed based on the idea to capture and reuse more supportive information in a static-dynamic feature distillation manner. Moreover, a multi-stage residual distillation network (MRDN) [44] is introduced based on using a multi-stage residual distillation block (MRDB) that decreases the parameters using a combination of the channel separation and the skip connection. Also, a hierarchical residual feature network (HRFFN) [32] is developed based on using multiple mixed attention blocks (MABs) as basic blocks for boosting the model representative ability.

2.2. Attention-based Efficient SR

In [33], an attentive residual refinement network (ARRFN) is developed based on using a stack of attentive residual refinement blocks (ARRFB) for improving the performance based on the attentive residual mechanism. After that, a non-linear perceptual multi-scale network (NLPM-SNet) [43] is suggested by considering the fusion of the multi-scale image information in a non-linearly. In [9], a large kernel attention super-resolution network (LKASR) is introduced based on using multiple cascaded visual attention modules (VAM) for to extract global and local features iteratively. In [50], the authors introduced self-calibrated efficient Transformer (SCET) [50] using the notion of the pixel attention mechanism is a self-calibrated module for effectively extracting image features. In [12], Gendy et al. suggested a balanced spatial feature distillation and pyramid attention (BSPAN) based on balancing different attention types.

In addition, a lightweight local-global attention network (LGAN) [35] is developed based on the idea of improving both the local features and global features. A vast-receptive-field attention model is introduced [48] based on enhancing the pixel attention block to improve the performance and decrease the number of parameters. Lastly, a wavelet-based Transformer for image super-resolution (WTSR) [34] is developed based on implicitly mining the self-similarity of image patches on the wavelet domain based on using Transformer. The attention-based methods can achieve good performance, however, these models need large memory and use a large computational resource.

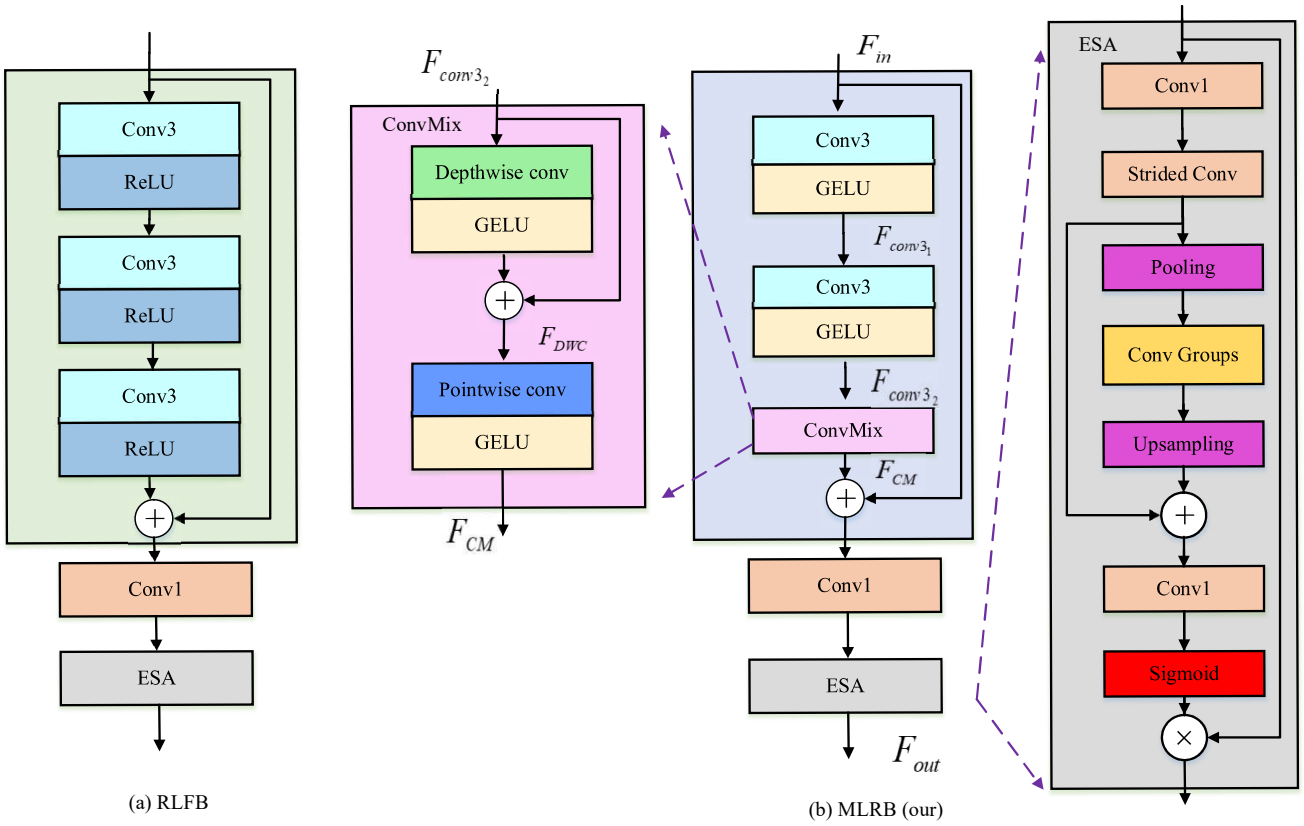

Figure 2. The structure of the proposed MLRB

2.3. Extreme Efficient-based SR

A very lightweight and efficient image super-resolution network (VLESR) is developed by Gao et al. [10] based on using a frequency grouping fusion block (FGFB) for the fusion of the high-frequency and low-frequency feature information. Then, in [36], the ShuffleMixer is developed for lightweight image SR based on exploring large convolution and channel split-shuffle operation. In [38], an edge-enhanced feature distillation network (EFDN) is suggested based on preserving high-frequency information under constrained resources. Moreover, a channel mixing Net (CDFM-Mobile) [13] is introduced based on the idea of channel mixing by utilizing a pointwise convolution layer. A fast nearest convolution module (NCNet) [29] is introduced based on nearest convolution, which is much fast than nearest upsampling and achieves similar performance. Lastly, a content-aware dynamic quantization (CADyQ) [14] model is suggested based on adaptively allocating optimal bits to local regions and layers.

3. Proposed Model.

The proposed MLRN is built using three stages, including shadow and deep feature extraction, and ends with one module to reconstruct the image. The shallow feature extraction is a basic convolution layer for transforming the image space to feature space. Then, the deep feature stage is based on a mixer-based local residual block, further modifying the original RLFN [20]. The MLRB is built by adding a convolution mixer [37] to the RLFN for spatial and channel feature mixing. In addition, GELU activation is used instead of the LeakyReLU in the RLFN. Finally, we used bilinear up-sampling [8, 13] to further improve the model performance with any extract parameter cost. We will discuss the detail of the MLRB as follows:

3.1. Mixer-based Local Residual Block (MLRB)

The MLRB further improves the previous RLFB [20] for reducing the parameters and flops. This is done by replacing one 3×3 convolution with convolution mixer (ConvMix) [37]. Also, similar to some previous methods, ESA [27] is used at the end of the block. Given the input feature F_{in}, the MLRB block function is be represented as:

$$F_{conv3_1} = H_{GELU}(H_{Conv3}(F_{in})), \quad (1)$$

$$F_{conv3_2} = H_{GELU}(H_{Conv3}(F_{conv3_1})), \quad (2)$$

Table 1. The Results on Benchmark Datasets to Evaluate Quantitatively . Best is indicated in **Bold** and Second Best is indicated in Underline. We Averaged the Time in (ms) on DIV2K Validation Dataset.

Method	Scale	#Params	#Mult-Adds	Time	Set5		Set14		B100		Urban100		Manga109	
					$PSNR$	$SSIM$	$PSNR$	$SSIM$	$PSNR$	$SSIM$	$PSNR$	$SSIM$	$PSNR$	$SSIM$
Bicubic	2	-	-	-	33.66	0.9299	30.24	0.8688	29.56	0.8431	26.88	0.8403	30.80	0.9339
SRCNN [6]	2	8K	52.7G	23	36.66	0.9542	32.45	0.9067	31.36	0.8879	29.50	0.8946	35.60	0.9663
FSRCNN [7]	2	12k	6.0G	15	37.00	0.9558	32.63	0.9088	31.53	0.8920	29.88	0.9020	36.67	0.9710
CARN [21]	2	1,592K	222.8G	210	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
LapSRN [22]	2	251K	29.9G	320	37.63	0.9588	33.04	0.9118	31.85	0.8942	30.75	0.9133	37.55	0.9732
IDN [17]	2	553K	174.1G	230	37.83	0.9600	33.30	0.9148	32.08	0.8985	31.27	0.9196	38.01	0.9749
IMDN [16]	2	694K	158.8G	150	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283	**38.88**	**0.9774**
RFDN [26]	2	534K	102.7G	140	38.05	0.9606	33.68	0.9184	32.16	0.8994	32.12	0.9278	**38.88**	0.9773
RLFN [20]	2	527K	97.9G	121	**38.07**	**0.9607**	33.72	**0.9187**	**32.22**	0.9000	**32.33**	**0.9299**	-	-
MLRN (Our)	2	488k	90.4G	130	**38.07**	**0.9607**	33.59	0.9180	32.21	**0.9000**	32.28	0.9297	38.76	0.9773
Bicubic	3	-			30.39	0.8682	27.55	0.7742	27.21	0.7385	24.46	0.7349	26.95	0.8556
SRCNN [6]	3	8K	52.7G	14	32.75	0.9090	29.30	0.8215	28.41	0.7863	26.24	0.7989	30.48	0.9117
FSRCNN [7]	3	12 k	5.0G	9	33.18	0.9140	29.37	0.8240	28.53	0.7910	26.43	0.8080	31.10	0.9210
CARN [21]	3	1,592K	118.8G	100	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
IDN [17]	3	553K	105.6G	170	34.12	0.9254	30.04	0.8382	28.97	0.8025	27.57	0.8398	33.00	0.9403
IMDN [16]	3	703K	71.5G	67	34.36	0.9270	30.32	0.8417	29.09	0.8046	28.17	0.8519	33.61	0.9445
RFDN [26]	3	541K	52.1G	64	34.41	**0.9280**	30.34	0.8420	29.09	0.8050	**28.21**	0.8525	**33.67**	0.9449
MLRN (Our)	3	496K	40.9G	62	**34.46**	0.9267	**30.35**	**0.8426**	**29.10**	**0.8054**	28.20	**0.8533**	33.66	**0.9450**
Bicubic	4	-	-		28.42	0.8104	26.00	0.7027	25.96	0.6675	23.14	0.6577	24.89	0.7866
SRCNN [6]	4	8K	52.7G	10	30.48	0.8626	27.50	0.7513	26.90	0.7101	24.52	0.7221	27.58	0.8555
FSRCNN [7]	4	12 k	4.6G	8	30.72	0.8660	27.61	0.7550	26.98	0.7150	24.62	0.7280	27.90	0.8610
CARN [21]	4	1,592K	90.9G	79	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
LapSRN [22]	4	502K	149.4G	163	31.54	0.8850	28.19	0.7720	27.32	0.7270	25.21	0.7560	29.09	0.8900
IDN [17]	4	553K	81.87G	145	31.82	0.8903	28.25	0.7730	27.41	0.7297	25.41	0.7632	29.41	0.8942
IMDN [16]	4	715K	40.9G	41	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838	30.45	0.9075
RFDN [26]	4	550K	26.5G	38	32.24	0.8952	28.61	0.7819	27.57	0.7360	26.11	0.7858	**30.58**	0.9089
RLFN [20]	4	543K	25.3G	35	32.24	0.8952	**28.62**	0.7813	**27.60**	0.7364	**26.17**	**0.7877**	-	-
MLRN (Our)	4	507K	23.5G	38	**32.30**	**0.8956**	**28.62**	**0.7824**	27.57	**0.7365**	26.10	0.7867	30.56	**0.9092**

where F_{conv3_1}, F_{conv3_2} is the output of the first and second convolution layers of MLRB. Also, H_{GELU}, H_{conv3} are the function of the GELU activation and the 3×3 convolution. Then, the depthwise convolution with residual learning are applied to F_{conv3_2} as:

$$F_{DWC} = H_{GELU}(H_{DWC}(F_{conv3_2})) + F_{conv3_2}, \quad (3)$$

where H_{DWC} is the function of the depthwise convolution layer and F_{DWC} is the output of depthwise convolution layer. Afterward, the pointwise convolution layer (H_{PWV}) with GELU activation are added as fellows:

$$F_{CM} = H_{GELU}(H_{PWV}(F_{DWC})) \quad (4)$$

where F_{CM}, is the output of ConvMix layer. After that, both the second pointwise convolution with ESA are added.

$$F_{out} = H_{ESA}(H_{conv1}(F_{CM} + F_{in})), \quad (5)$$

where F_{out} is the output of MLRB. Also, H_{ESA}, H_{conv1} are the functions of the f ESA layer, and the second pointwise convolution layer.

3.2. The Mixer-based Local Residual Network (MLRN) Framework

As indicated in Figure 1, the MLRN model is designed based on using the stages of shallow feature extraction, deep feature extraction, and image reconstruction. For the shallow feature extraction stage, we used 3×3 convolution (H_{conv3}), which is able to change from the image domain to feature domains. Then, the stage function can be represented as:

$$F_0 = H_{conv3}(x) \quad (6)$$

After that, we built the deep feature extraction stage based on utilizing m layers of MLRB (H_{MLRB}).

$$F_i = H_{MLRB_i}(F_{i-1}), \quad i = 1, 2, .., m \quad (7)$$

where F_i represents the m layer of output of the the MLRB. Afterward, 3×3 is used for smoothing the aggregated features as:

$$F_{DF} = H_{Conv3}(F_m), \quad (8)$$

where H_{Conv3} is the function of 3×3 convolution layer. In addition, we generated the output SR image based on using reconstruction modules as follows:

$$F_{SR} = H_{recont}(F_{DF} + F_0), \quad (9)$$

where H_{recont} deontes the reconstruction function represented by both 3×3 convolution and Sup-pixel upsampling and, F_{SR} denotes the SR model output. Finally, the bilinear up-sampling is made for the input image x and added to the image.

$$y = H_{BU}(x) + F_{SR}, \quad (10)$$

where y is the model output and H_{BU} is the function of the bilinear up-sampling module.

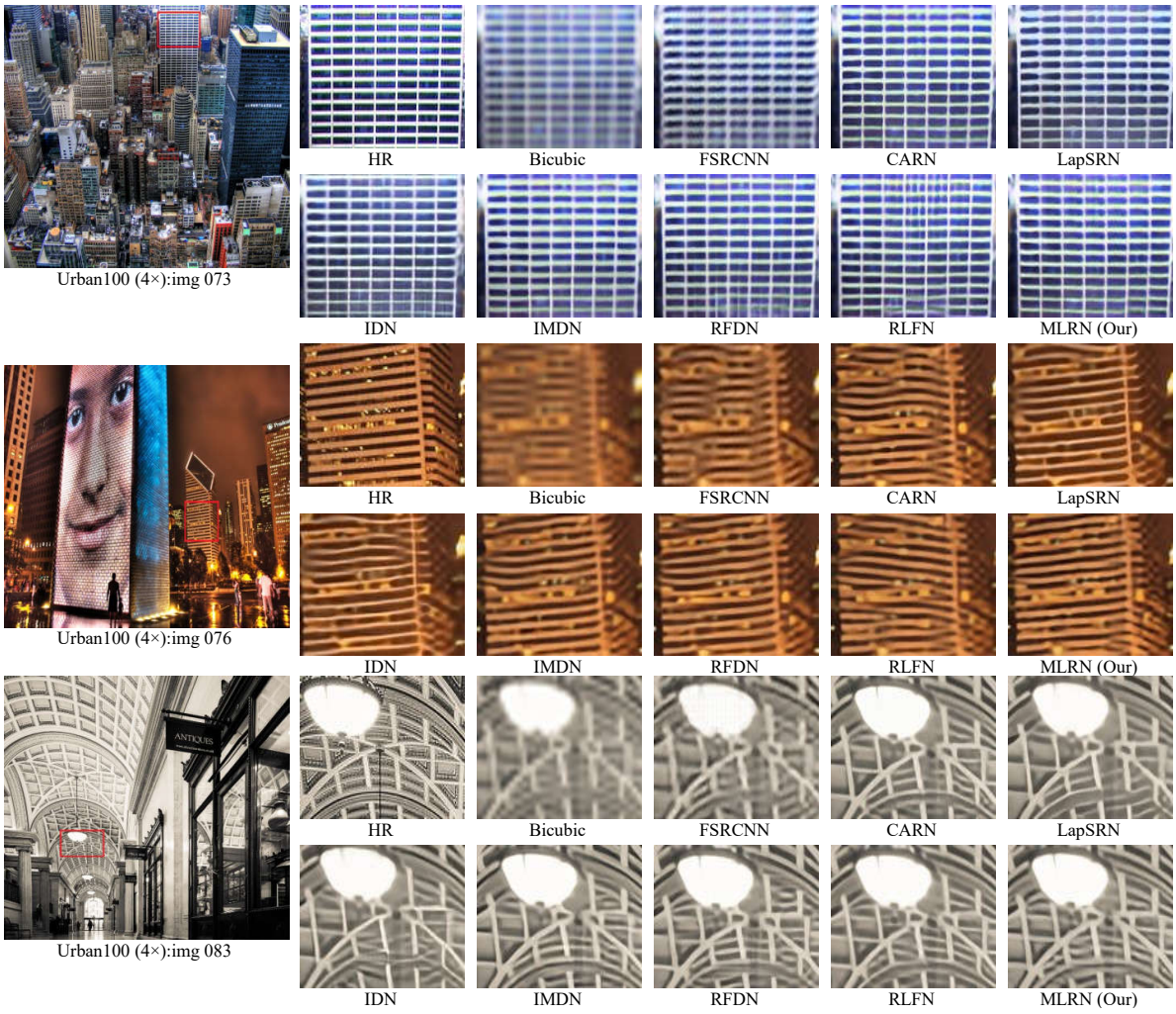

Figure 3. Urban100 dataset Visual Comparison at $\times$ 4 SR.

4. Experiment

4.1. Benchmarks

As for the training section, the DIV2K [1] dataset is used for training our method, and we used the bicubic downsampling method to create the LR image from the HR image by downsampling it using the bicubic downsampling method. There are five datasets that were used to test the model, including Set5 (5 images) [4], Set14 (14 images) [45], B100 (100 images) [3], Urban100 (100 images) [15], and Manga109 (109 images) [30]. Finally, in order to evaluate the model based on the Y channel, the PSNR and structural similarity index (SSIM) [40] were used.

4.2. Implementation Details

The patch size is set to 192 $\times$ 192, 192 $\times$ 192, 256 $\times$ 356 for scales of $\times$ 2, $\times$ 3, and $\times$ 4, respectively. In addition, batch sizes is fixed to 64 to train the MLRN models. Moreover, 90, 180, and 270 degrees of random rotation and horizontal flipping are used to augment the input images. Afterward, the number of MLRB is set to 6. In addition, the number of features is fixed at 58. The ADAM optimizer [19] is applied with β_1 = 0.9, β_2 = 0.99 and ϵ= $1e^{-8}$. Additionally, the learning rate starts with 5×10^{-4} and half every 200 epochs. We used the L_1 loss function for training the model for 1000 epochs. Also, the warm-start strategy [20] is utilized for the MLRN, and it is not used for the ablation study model for saving time. Finally, we designed the model based on the PyTorch [31] framework and trained using Nvidia 2080 Ti GPUs.

4.3. Comparison with State-of-the-art SR models

Our methods are compared with 8 state-of-the-art methods for SR of lightweight images that have been developed

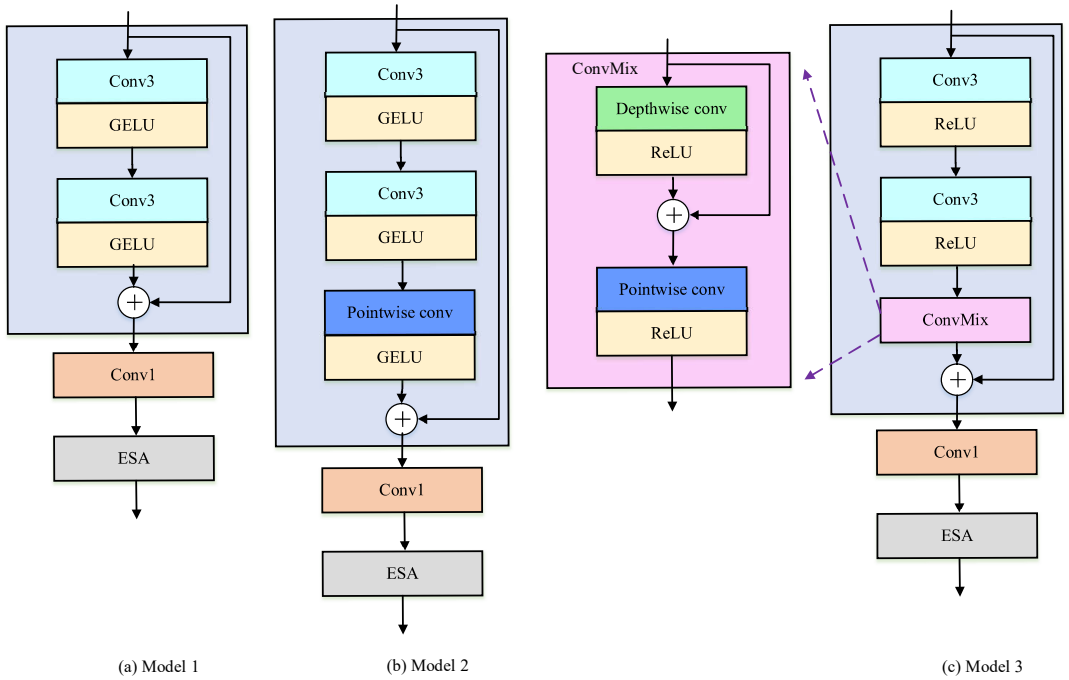

Figure 4. The Models Used in Ablation Study

in recent years. The methods can be categorized to SRCNN [6], FSRCNN [7], CARN [21], LapSRN [22], IDN [17], IMDN [16], RFDN [26], and RLFN [20]. In order to compare our model to these methods, we analyzed them quantitatively, qualitatively, and in relation to their model size.

4.4. Comparison with State-of-the-art SR models

4.4.1 Quantitative Evaluations

In this section, there five datasets are utilized to make this quantitative comparison with other state-of-the-art methods, as indicated in Table 1. The table shows that our model has better performance in cases of the PNSR and SSIM for most of the cases. For example, at the 3 $\times$ scale, our model has better PSNR than RFDN [26] for the datasets of Set5, Set14, and B100 by 0.05 dB, 0.01 dB, and 0.01 dB. Also, for the case of SSIM, our model performs better than the RLFN [20] for datasets of Set5, Set14, and B100 by improving the values from 0.8952, 0.7813, and 0.7364 to 0.8956, 0.7824, and 0.7365, respectively. These results show that our model can benefit from the ConvMix block to make the mixing of channel and spatial features.

4.4.2 Qualitative Evaluations

In the part of the result, we aimed to compare our model with other state-of-the-art models in the case of qualitative results, as shown in Figure 3. The figure shows that our model has better visual image quality than the other different methods, including the RFDN [26] and RLFN [20]. For instance, for the img 073, the window details are much clearer in our case than in the other methods. Moreover, in the case of img 076, the horizontal lines of the building in our method are much straight compared to the different models. The results indicated that our model is able to extract more features using ConvMix, which helps it achieve good performance.

4.4.3 Model Size Analysis

Here, these three factors parameters, multi-Adds, and runtime, are used to make the model size analysis compared to the winner for the previous three challenges, as shown in Table 1. Firstly for the parameters, our model has few parameters compared to RLFN [20] by 39K at the scale of 2 $\times$ with comparable good performance. In addition, for the Mult-Adds, our model has 11.2G fewer Mult-Adds than RFDN [26] at the scale of $\times$ 3. Moreover, for the runtime, it is clear that our model has a slight increase in the runtime of about 8% than RLFN at the scale of $\times$ 4, but our model has fewer parameters, multi-Adds, and better performance. So, our model can balance the performance and computational cost.

Table 2. The Ablation Study Results at the Scale $\times$ 2

Method	#Params	#Mult-Adds	Set5		Set14		B100		Urban100		Manga109	
			$PSNR$	$SSIM$	$PSNR$	$SSIM$	$PSNR$	$SSIM$	$PSNR$	$SSIM$	$PSNR$	$SSIM$
MLRN	488k	90.4G	38.05	0.9606	33.59	0.9179	32.19	0.8999	32.18	0.9288	38.67	0.9771
MLRN W/O ConvMix	484.3K	85.6G	38.05	0.9606	33.49	0.9172	32.17	0.8996	32.08	0.9278	38.61	0.9770
MLRN W/O depthwise conv	484.4k	89.7G	38.04	0.9606	33.59	0.9179	32.18	0.8997	32.13	0.9283	38.65	0.9771
MLRN W LeakyReLU	488k	90.4G	37.99	0.9604	33.52	0.9173	32.17	0.8999	32.05	0.9229	38.58	0.9770
MLRN W/O Bilinear Up-sampling	488k	90.4G	38.03	0.9606	33.60	0.9177	32.19	0.8998	32.14	0.9284	38.63	0.9770
MLRN W Warm-Start	488k	90.4G	38.07	0.9607	33.59	0.9180	32.21	0.9000	32.28	0.9297	38.76	0.9773

4.5. Ablation Study

During our ablation, the same model settings for the different scales were used, but we used the model before the warm-start strategy to save time. This ablation studies important factors used in the designing of the model, such as the impact of including the ConvMix block in the model, the impact of including the depthwise convolution layer in the model, the impact of using GELU activation instead of LeakyReLU activation which is found on the original model, the effect of using the bilinear up-sampling, and at the end, we show the impact of using the warm-start strategy.

4.5.1 Ablation Study in Using the ConvMix Block

In this part of ablation, we aimed to show the impact of the ConvMix layer, so we used the same model as MLRN and compared it with the model when the ConvMix is not included in the main model, as shown in Figure 4 (Model 1). The result is shown in Table 2 (the second row vs. the first row). It is clear that the ConvMix impacts the result with a slight increase of the parameters and Mult-Adds. For instance, the PSNR dropped by 0.1 dB and 0.1 dB for Set14 and Urban100 datasets, respectively. Based on this result, we can conclude that ConvMix can help the model by making channel and spatial mixing with sight increase the computational cost.

4.5.2 Ablation Study in Using the depthwise convolution Layer

In this part of ablation, we aimed to show the impact of the depthwise convolution, so the same model as MLRN is used and compared it with the model when the depthwise convolution was not included in the main model. The result is shown in Table 2 (the third row vs. the first row). It is clear that the depthwise convolution impacts the result with a slight increase of the parameters and Mult-Adds. For instance, the PSNR dropped by 0.05 dB and 0.02 dB for Urban100 and Manga109 datasets, respectively. Based on this result, we can conclude that the depthwise convolution can help the model by making spatial mixing with sight increase the computational cost.

4.5.3 Ablation Study in Using the GELU Activation

In this part of ablation, we aimed to show the impact of the GELU activation, so we used the same model as MLRN and compared it with the model when the LeakyReLU activation was used in the main model, as shown in Figure 4 (Model 3). The result is shown in Table 2 (the fourth row vs. the first row). It is clear that the GELU activation greatly impacts the result; for example, the PSNR decreased by 0.06 dB and 0.13 dB for Set5 and Urban100 datasets, respectively. Based on this result, we can conclude that GELU activation is more suitable for SR task than LeakyReLU activation.

4.5.4 Ablation Study in Using the Bilinear Up-sampling

In this part of ablation, we aimed to show the impact of the bilinear up-sampling, so we used the same model as MLRN and compared it with the model when the bilinear up-sampling is not included in the main model. The result is indicated in Table 2 (the fifth row vs. the first row). It obvious the bilinear up-sampling impacts the result without an increase of the parameters and Mult-Adds. For example, the PSNR dropped by 0.03 dB, 0.04 dB, and 0.04 dB for Set5, Urban100, and Manga109 datasets, respectively. Based on this result, we can conclude that bilinear up-sampling can help by transferring the low-frequency information to the final layer.

4.5.5 The Ablation Study on Warm-Start Strategy

In this task, the warm-start strategy [20] is utilized to make retraining the model again, beginning from the pre-train model on the same scale. We made the comparison in Table 2 (the last row compared to the first row). The results indicate that the warm-start strategy can improve performance. For example, in Urban100 and Manga109 datasets, the model performance is case of PSNR enhanced from 32.18 dB and 38.67 dB to 3.28 dB and 38.86 dB, respectively. So, these results show that the warm-start strategy can enhance the performance without any additional parameters and Multi-Adds cost.

Table 3. Results of NTIRE 2023 Efficient SR Challenge.

Team	Time [ms]			PSNR [dB]		#Params	FLOPs	#Acts	GPU Mem.	Model	Overall	#Conv
	Ave.	Val.	Test	Val.	Test	[M]	[G]	[M]	[M]	Comp.		
MegSR	$18.30_{(1)}$	21.26	15.33	29.04	26.95	$0.243_{(12)}$	$14.90_{(11)}$	$72.97_{(6)}$	$495.91_{(19)}$	$23_{(11)}$	$49_{(2)}$	39
Zapdos	$18.59_{(2)}$	21.69	15.48	28.96	27.03	$0.352_{(25)}$	$21.97_{(25)}$	$63.01_{(2)}$	$420.50_{(13)}$	$50_{(25)}$	$67_{(10)}$	26
DFCDN	$18.71_{(3)}$	21.91	15.51	29.00	27.08	$0.245_{(13)}$	$15.49_{(14)}$	$82.76_{(13)}$	$376.99_{(12)}$	$27_{(14)}$	$55_{(4)}$	39
KaiBai Group	$20.49_{(4)}$	23.94	17.05	28.95	27.01	$0.272_{(17)}$	$16.76_{(17)}$	$65.10_{(3)}$	$296.45_{(7)}$	$34_{(17)}$	$48_{(1)}$	35
R.I.P. ShopeeVideo	$20.65_{(5)}$	24.34	16.96	28.97	27.04	$0.255_{(15)}$	$16.16_{(16)}$	$74.97_{(7)}$	$439.60_{(14)}$	$31_{(15)}$	$57_{(5)}$	35
Antins_cv	$20.92_{(6)}$	24.45	17.39	29.00	26.95	$0.315_{(24)}$	$20.07_{(24)}$	$70.82_{(5)}$	$488.61_{(17)}$	$48_{(24)}$	$76_{(14)}$	29
Young	$22.09_{(7)}$	25.86	18.33	28.97	27.00	$0.543_{(30)}$	$33.38_{(30)}$	$61.87_{(1)}$	$293.05_{(6)}$	$60_{(30)}$	$74_{(13)}$	23
NTU607_ESR	$22.71_{(8)}$	26.73	18.68	29.00	27.07	$0.281_{(19)}$	$17.31_{(19)}$	$76.11_{(9)}$	$364.24_{(11)}$	$38_{(19)}$	$66_{(8)}$	39
CMVG	$24.42_{(9)}$	28.51	20.33	29.01	27.08	$0.307_{(21)}$	$18.98_{(21)}$	$81.55_{(11)}$	$454.51_{(15)}$	$42_{(21)}$	$77_{(15)}$	41
Touch_Fish	$25.61_{(10)}$	30.09	21.12	29.00	27.09	$0.415_{(27)}$	$27.16_{(27)}$	$75.50_{(8)}$	$769.56_{(27)}$	$54_{(27)}$	$99_{(26)}$	20
CUC_SR	$25.97_{(11)}$	30.58	21.37	28.99	27.05	$0.402_{(26)}$	$25.23_{(26)}$	$81.88_{(12)}$	$344.51_{(9)}$	$52_{(26)}$	$84_{(20)}$	39
SeaOuter	$26.26_{(12)}$	30.93	21.59	28.95	27.05	$0.285_{(20)}$	$18.63_{(20)}$	$80.48_{(10)}$	$218.97_{(3)}$	$40_{(20)}$	$65_{(7)}$	44
NoahTerminalCV B	$27.83_{(13)}$	32.73	22.94	28.96	27.03	$0.209_{(10)}$	$13.34_{(10)}$	$118.71_{(18)}$	$188.21_{(1)}$	$20_{(10)}$	$52_{(3)}$	49
NJUST_R	$28.63_{(14)}$	33.59	23.66	28.99	27.07	$0.237_{(11)}$	$15.40_{(13)}$	$86.11_{(14)}$	$303.06_{(8)}$	$24_{(12)}$	$60_{(6)}$	58
NoahTerminalCV A	$28.71_{(15)}$	33.74	23.69	28.99	27.06	$0.310_{(23)}$	$19.99_{(23)}$	$68.38_{(4)}$	$188.60_{(2)}$	$46_{(23)}$	$67_{(9)}$	25
Sissie_Lab	$30.34_{(16)}$	34.62	26.07	29.00	27.00	$0.461_{(28)}$	$28.85_{(28)}$	$107.07_{(16)}$	$628.94_{(25)}$	$56_{(28)}$	$113_{(29)}$	48
GarasSjtu (Our)	$32.30_{(17)}$	37.99	26.62	28.91	26.99	$0.275_{(18)}$	$16.85_{(18)}$	$97.87_{(15)}$	$556.22_{(22)}$	$36_{(18)}$	$90_{(22)}$	43
USTC_ESR	$34.16_{(18)}$	40.11	28.20	29.03	27.09	$0.503_{(29)}$	$31.56_{(29)}$	$112.57_{(17)}$	$489.33_{(18)}$	$58_{(29)}$	$111_{(28)}$	48
SEU_CNII	$40.84_{(19)}$	48.35	33.33	28.99	27.08	$0.616_{(31)}$	$38.63_{(31)}$	$133.57_{(19)}$	$944.91_{(29)}$	$62_{(31)}$	$129_{(31)}$	64
AVC2_CMHI_SR	$43.46_{(20)}$	51.30	35.61	29.01	27.06	$0.262_{(16)}$	$15.52_{(15)}$	$154.19_{(20)}$	$821.45_{(28)}$	$31_{(16)}$	$99_{(25)}$	84
NJUST_M	$68.11_{(21)}$	79.54	56.68	28.96	27.05	$0.104_{(3)}$	$6.56_{(3)}$	$199.35_{(23)}$	$503.49_{(20)}$	$6_{(3)}$	$70_{(12)}$	66
TelunXupt	$75.89_{(22)}$	88.10	63.68	29.00	27.09	$0.095_{(1)}$	$5.58_{(1)}$	$220.88_{(25)}$	$517.14_{(21)}$	$2_{(1)}$	$70_{(11)}$	317
Set5 Baby	$99.79_{(23)}$	117.33	82.25	29.01	27.08	$0.129_{(6)}$	$8.29_{(5)}$	$202.70_{(24)}$	$652.41_{(26)}$	$11_{(5)}$	$84_{(19)}$	86
NJUST_E	$106.61_{(24)}$	125.02	88.20	28.97	27.04	$0.099_{(2)}$	$6.02_{(2)}$	$242.96_{(26)}$	$606.38_{(24)}$	$4_{(2)}$	$78_{(16)}$	66
LVGroup_HFUT	$112.68_{(25)}$	132.10	93.26	28.98	27.05	$3.426_{(32)}$	$224.19_{(32)}$	$335.28_{(30)}$	$590.58_{(23)}$	$64_{(32)}$	$142_{(32)}$	94
FRL Team 4	$124.13_{(26)}$	145.18	103.07	28.95	27.02	$0.173_{(7)}$	$10.60_{(7)}$	$187.32_{(22)}$	$1266.92_{(31)}$	$14_{(7)}$	$93_{(23)}$	198
Dase-IDEALab	$130.73_{(27)}$	153.30	108.17	29.00	27.07	$0.118_{(5)}$	$9.06_{(6)}$	$332.39_{(29)}$	$1114.77_{(30)}$	$11_{(6)}$	$97_{(24)}$	122
FRL Team 1	$186.02_{(28)}$	218.82	153.23	29.01	27.03	$0.200_{(9)}$	$12.76_{(9)}$	$243.20_{(27)}$	$265.25_{(5)}$	$18_{(9)}$	$78_{(18)}$	100
FRL Team 0	$196.64_{(29)}$	230.14	163.14	29.01	26.98	$0.115_{(4)}$	$7.38_{(4)}$	$170.26_{(21)}$	$2028.66_{(32)}$	$8_{(4)}$	$90_{(21)}$	58
FRL Team 3	$201.34_{(30)}$	237.73	164.94	29.00	27.09	$0.179_{(8)}$	$11.54_{(8)}$	$285.41_{(28)}$	$262.67_{(4)}$	$16_{(8)}$	$78_{(17)}$	112
AIIA-SR	$224.45_{(31)}$	264.99	183.91	29.00	27.07	$0.307_{(22)}$	$19.53_{(22)}$	$355.47_{(31)}$	$482.70_{(16)}$	$44_{(22)}$	$122_{(30)}$	89
FRL Team 2	$282.42_{(32)}$	331.55	233.29	29.02	27.02	$0.245_{(14)}$	$15.37_{(12)}$	$422.90_{(32)}$	$355.94_{(10)}$	$26_{(13)}$	$100_{(27)}$	158

4.6. MLRN for NTIRE 2023 Challenge

We took part in NTIRE 2023 Challenge on Efficient Super-Resolution [24], and our model achieved a good result, as shown in Table 3. Our MLRN-entire model is slightly changed from the MLRN model in the paper; it contains 4 MLRB blocks, each containing two 3×3 convolution with one ConvMix block each with GELU activation, in which the number of feature maps is set to 50. Also, the channel number of the ESA is set to 16, similar to [20], and we set the RGB range to 255, not to 1, as in the paper. In our training, DIV2K, and LSDIR [22] are used to train the model. After that, the model is trained in the following steps. At the starting stage, we trained the model from scratch using the DIV2K and LSDIR [22] datasets, with a patch size of 256×256 and a batch size of 64. In this training, the L_1 loss function is used with the Adam optimizer. This stage is trained for 800 epochs with an initial learning rate 5×10^{-4} reduced by half every 200 epochs. After the previous stage, the model starts its training from the previous pre-trained weights using the DIV2K and Flickr2K datasets with an initial learning rate 5×10^{-4}that drops by 50% at every 200 epochs for 1000 epochs using L_1 loss

5. Conclusion

This paper proposes a mixer-based local residual network (MLRN) for single image super-resolution (SISR). The MLRN was built by further improving the original RLFN using the convolutional mixer (ConvMix). The ConvMix is built based on using depthwise and pointwise to mix channel and spatial features. Based on the ConvMix, the mixer residual local block (MLRB) is built as the backbone for our model. In addition, both GELU activation and bilinear up-sampling are used to improve the model further. Based on the result, our model achieved good results against other state-of-the-art models in multiple benchmarks. In addition, our model takes part in Efficient Super-Resolution Challenge 2023 and shows amenable performance.

Acknowledgment. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB2204500, in part by the National Natural Science Foundation of China under Grant. 62074097.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 5

[2] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020. 1

[3] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010. 5

[4] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 5

[5] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 1

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014. 4, 6

[7] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 391–407. Springer, 2016. 4, 6

[8] Zongcai Du, Jie Liu, Jie Tang, and Gangshan Wu. Anchor-based plain net for mobile image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2494–2502, 2021. 3

[9] Hao Feng, Liejun Wang, Yongming Li, and Anyu Du. Lkasr: Large kernel attention for lightweight image super-resolution. *Knowledge-Based Systems*, 252:109376, 2022. 2

[10] Dandan Gao and Dengwen Zhou. A very lightweight and efficient image super-resolution network. *Expert Systems with Applications*, 213:118898, 2023. 3

[11] Garas Gendy, Guanghui He, and Nabil Sabor. Lightweight image super-resolution based on deep learning: State-of-the-art and future directions. *Information Fusion*, 94:284–310, 2023. 1

[12] Garas Gendy, Nabil Sabor, Jingchao Hou, and Guanghui He. Balanced spatial feature distillation and pyramid attention network for lightweight image super-resolution. *Neurocomputing*, 509:157–166, 2022. 2

[13] Garas Gendy, Nabil Sabor, Jingchao Hou, and Guanghui He. Real-time channel mixing net for mobile image super-resolution. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 573–590. Springer, 2023. 3

[14] Cheeun Hong, Sungyong Baik, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Cadyq: Content-aware dynamic quantization for image super-resolution. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 367–383. Springer, 2022. 3

[15] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 5

[16] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th acm international conference on multimedia*, pages 2024–2032, 2019. 1, 4, 6

[17] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 723–731, 2018. 4, 6

[18] Michal Kawulok, Pawel Benecki, Szymon Piechaczek, Krzysztof Hrynczenko, Daniel Kostrzewa, and Jakub Nalepa. Deep learning for multiple-image super-resolution. *IEEE Geoscience and Remote Sensing Letters*, 17(6):1062–1066, 2019. 1

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[20] Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 766–776, 2022. 1, 3, 4, 5, 6, 7, 8

[21] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 4, 6

[22] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. 4, 6, 8

[23] Yawei Li, Kai Zhang, Radu Timofte, Luc Van Gool, Fangyuan Kong, Mingxi Li, Songwei Liu, Zongcai Du, Ding Liu, Chenhui Zhou, et al. Ntire 2022 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1062–1102, 2022. 1

[24] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. Ntire 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1, 8

[25] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF inter-*

national conference on computer vision, pages 1833–1844, 2021. 1

[26] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 41–55. Springer, 2020. 1, 4, 6

[27] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2359–2368, 2020. 3

[28] Jing Luo, Lin Zhao, Li Zhu, and Wenbing Tao. Multi-scale receptive field fusion network for lightweight image super-resolution. *Neurocomputing*, 493:314–326, 2022. 1, 2

[29] Ziwei Luo, Youwei Li, Lei Yu, Qi Wu, Zhihong Wen, Haoqiang Fan, and Shuaicheng Liu. Fast nearest convolution for real-time efficient image super-resolution. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 561–572. Springer, 2023. 3

[30] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 5

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[32] Jiayi Qin, Feiqiang Liu, Kai Liu, Gwanggil Jeon, and Xiaomin Yang. Lightweight hierarchical residual feature fusion network for single-image super-resolution. *Neurocomputing*, 478:104–123, 2022. 2

[33] Jinghui Qin and Rumin Zhang. Lightweight single image super-resolution with attentive residual refinement network. *Neurocomputing*, 500:846–855, 2022. 2

[34] Jinye Ran and Zili Zhang. Lightweight wavelet-based transformer for image super-resolution. In *PRICAI 2022: Trends in Artificial Intelligence: 19th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2022, Shanghai, China, November 10–13, 2022, Proceedings, Part III*, pages 368–382. Springer, 2022. 2

[35] Zijiang Song and Baojiang Zhong. A lightweight local-global attention network for single image super-resolution. In *Proceedings of the Asian Conference on Computer Vision*, pages 4395–4410, 2022. 2

[36] Long Sun, Jinshan Pan, and Jinhui Tang. Shufflemixer: An efficient convnet for image super-resolution. *arXiv preprint arXiv:2205.15175*, 2022. 3

[37] Asher Trockman and J Zico Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022. 3

[38] Yan Wang. Edge-enhanced feature distillation network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 777–785, 2022. 3

[39] Yan Wang, Tongtong Su, Yusen Li, Jiuwen Cao, Gang Wang, and Xiaoguang Liu. Ddistill-sr: Reparameterized dynamic distillation network for lightweight image super-resolution. *IEEE Transactions on Multimedia*, 2022. 2

[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[41] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3365–3387, 2020. 1

[42] Ruimian Wen, Zhijing Yang, Tianshui Chen, Hao Li, and Kai Li. Progressive representation recalibration for lightweight super-resolution. *Neurocomputing*, 504:240–250, 2022. 1, 2

[43] Aiping Yang, Leilei Li, Jinbin Wang, Zhong Ji, Yanwei Pang, Jiale Cao, and Zihao Wei. Non-linear perceptual multi-scale network for single image super-resolution. *Neural Networks*, 152:201–211, 2022. 2

[44] Xin Yang, Yingqing Guo, Zhiqiang Li, Dake Zhou, and Tao Li. Mrdn: A lightweight multi-stage residual distillation network for image super-resolution. *Expert Systems with Applications*, 204:117594, 2022. 2

[45] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. 5

[46] Kai Zhang, Martin Danelljan, Yawei Li, Radu Timofte, Jie Liu, Jie Tang, Gangshan Wu, Yu Zhu, Xiangyu He, Wenjie Xu, et al. Aim 2020 challenge on efficient super-resolution: Methods and results. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 5–40. Springer, 2020. 1

[47] Kai Zhang, Shuhang Gu, Radu Timofte, Zheng Hui, Xiumei Wang, Xinbo Gao, Dongliang Xiong, Shuai Liu, Ruipeng Gang, Nan Nan, et al. Aim 2019 challenge on constrained super-resolution: Methods and results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3565–3574. IEEE, 2019. 1

[48] Lin Zhou, Haoming Cai, Jinjin Gu, Zheyuan Li, Yingqi Liu, Xiangyu Chen, Yu Qiao, and Chao Dong. Efficient image super-resolution using vast-receptive-field attention. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 256–272. Springer, 2023. 2

[49] Zhikai Zong, Lin Zha, Jiande Jiang, and Xiaoxiao Liu. Asymmetric information distillation network for lightweight super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2022. 1, 2

[50] Wenbin Zou, Tian Ye, Weixin Zheng, Yunchen Zhang, Liang Chen, and Yi Wu. Self-calibrated efficient transformer for lightweight super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 930–939, 2022. 2