RESIDUAL ATTENTION NETWORK FOR WAVELET DOMAIN SUPER-RESOLUTION

Jing Liu, Yuan Xie, Haichuan Song, Wang Yuan, Lizhuang Ma

School of Software Engineering, East China Normal University, China
Emails:{51174500035,51184501076}@stu.ecnu.edu.cn,
{hcsong,lzma}@sei.ecnu.edu.cn,xieyuan8589@foxmail.com

ABSTRACT

Single-image super-resolution plays an important role in computer vision area. However, previous works using convolutional neural networks perform badly when reconstructing high frequency details, result in over-smooth and lacking of textural information in the output. At the same time, super-resolution computation always relays on convolutional neural networks with huge depth, which is super tricky to train and use. In this paper, we propose a novel network with better textural details in wavelet domain, which is composed of a feature extract layer, residual channel attention groups (RCAG) and a residual up-sampling layer based on inverse discrete wavelet transform. Channel attention and spatial attention layers are inserted into residual channel and spatial attention blocks (RCSAB), enhancing the learning of high frequency information with attention maps. Composed of a chain of RCSAB and a channel attention layer with short skip connection, RCAG is good at catching long-term high frequency information. Then the feature mapping component is composed of a chain of RCAG. Experiment shows that our method performs better than state-of-the-art methods on benchmark datasets in different scales.

Index Terms— Single Image Super Resolution, Image Processing, Convolutional Neural Networks, Spatial Attention, Inverse Discrete Wavelet Transformation

1. INTRODUCTION

Single image super resolution (SISR) is a classical low-level vision task, aiming to reconstruct the corresponding high-resolution (HR) image from a low-resolution (LR) image, which is widely applied in security and surveillance imaging [1], medical imaging [2] and HDTV in recent years.

CNN-based networks[3] have been developed to dramatically improve the super resolution quality and recover the details of the images, owning to its its powerful learning ability and efficient feature expressive ability. The simplest way to boost network performance is to stack more or wider convolution layers (Conv) for feature mapping. For example, EDSR [4] introduces a residual base block architecture for SR which is widely-used in following papers, and widens the network to 256 feature maps with 64 3x3 Conv. However, simply stacking more residual blocks can hardly obtain better improvements, while computation complexity and memory storage requirements grows rapidly. In our proposed WRAN, a channel attention grouping strategy with residual scale factor is used in the base group - RCAG (Residual Channel Attention Group), with which the stability in training process could be increased, the network could be reached much deeper compared with previous CNN-based methods, long-term high frequency details could be kept.

Attention mechanism is proposed to recalibrate the allocation of avalibale computational resources towards the most informative components, which is useful across many tasks such as image generation[5], image captioning[6] and image restoration[7][8]. SENet [9] firstly employs a lightweight module to treat channel-wise feature differently according to the respective weight responses. [10] introduced a Convolutional Block Attention Module which emphasize meaningful features along channel and spatial axes. In SR field, most previous networks, such as EDSR [4] and RDN [11], deal with different types of information in the same way, which limits the representation ability and fitting capacity of deep networks. RCAN [7] firstly introduced channel attention mechanism in SR task, making full use of channel independent information, but ignored the spatial information in the same channel. To overcome this limitation, we propose a base block, residual channel and spatial attention block (RCSAB), for deep super resolution network, which can pay more attention to more important channel and more important region in each channel.

Up-sampling method in CNN based SISR network. There are three sub-nets in recently CNN-based SISR network: feature extraction part, feature mapping part and method for up-sampling. Interpolation up-sampling was first used in SRCNN[3], following SRCNN using pre-interpolation, VDSR[12], DRCN[13] and MemNet[14] used different feture extraction module. However, this pre-processing step causes additional calculations, especially in high magnifaication SR tasks. Deconvolution proposed in [15] is then used as up-sampling layer in the tail of SR network. Recently SR network such as FSRCNN[16], LapSRN[17], DBPN[18] and etc used deconvolution as the up-sampling module for better performence. At the same time, sub-pixel convolution used in ESPCN[19], EDSR[4], SRMD[20] and RCAN[7] is proposed to avoid checkerboard artifacts caused by deconvolution as a up-sampling module by reshaping the output feature map. Unlike the above methods, instead of using pre-interpolation, we also propose a residual up-sampling strategy based on inverse discrete wavelet transformation (IDWT) to catch the easily-lost high frequency textural details, which weakens the problem of blurring and over-smoothing of reconstructed images.

2. WAVELET RESIDUAL ATTENTION NETWORK

2.1. Network Architecture

The overall architecture of our WRAN is illustrated in Fig. 1, which mainly consists of three parts: a shallow feature extraction layer, Feature Mapping Net (FMN), and an up-sampling module. Let's denote I_{LR} and I_{SR} as the input LR image and the corresponding output HR image of WRAN respectively. Same as [8], we only use one convolutional layer (Conv) to extract the shallow feature F_0 from the input:

$$F_{SF} = f_0(I_{LR}), \tag{1}$$

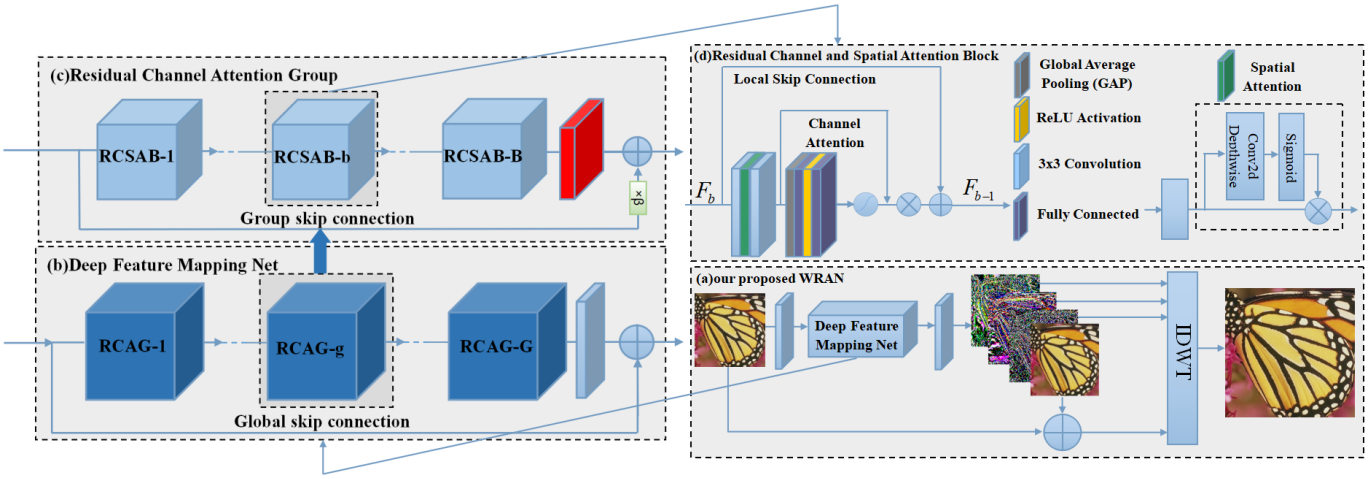

Fig. 1. (a) Network architecture of our proposed Wavelet Residual Attention Network(WRAN), (b) Deep Feature Mapping Net, (c) Residual Channel Attention Group and (d) Residual Channel and Spatial Attention Block.

Where $f_0(\cdot)$ denotes the first convolution. F_{SF} is then used as input features of FMN. F_{SF} is updated through G Residual Channel Attention Groups (RCAGs) and one convolution, and then the updated feature maps are added to F_{SF} by using global skip connection:

$$\begin{aligned} F_{DF} &= H_{FMN}(F_{SF}) \\ &= F_{SF} + f_{df}(G_G(G_{G-1}(...(G_1(F_{SF}))...))), \end{aligned} \tag{2}$$

where $G_g(\cdot)$ denotes the g-th RCAG, $f_{df}(\cdot)$ denotes the last convolution of feature mapping part and H_{FMN} denotes our proposed FMN with two stage skip connection structure. FMN achieves large depth and receptive field size, same as RCAN [7], and feature extracted by FMN noted as F_{DF}. F_{DF} is then fed into an up-sampling module with I_{LR} to reconstruct the final SR image:

$$I_{SR} = H_{UP}(F_{DF}, I_{LR}) = H_{WRAN}(I_{LR}), \tag{3}$$

where $H_{UP}(\cdot)$ and $H_{WRAN}(\cdot)$ denote the up-sampling module and our whole super resolution network, respectively.

Two-stage skip connection structure is used to stabilize the training of very deep network. In the first stage of FMN, the component consists of several RCAGs and one convolution layer with global skip connection(as shown in Fig. 1b). In the second stage in RCAG (as shown in Fig. 1c), the component consists of several Residual Channel and Spatial Attention Blocks (RCSABs), one channel attention layer with a scaling Group Skip Connection. The global skip connection is the longest skip connection in feature mapping stage. By learning residual information in a coarse level, the global skip connection eases the flow of information across RCAGs, which makes more abundant global low-frequency information easier bypassed in the training process. The network would pay more attention on high-frequency information, which was more difficult to learn. Group skip connection is the second stage skip connection in our two-stage skip connection structure. We find that networks straightly stacked with several feature mapping base blocks with a group skip connection will not converge. As investigated in [21], in a very deep CNN, only after a few thousands of iterations, the last layer would start to produce zeros, resulting in the network "die" early in training process, which is inevitable even by reducing learning rate or adding extra normalization layer. Thus, we scale the input by β=0.2 and then use a residual connection in the output, to avoid scaling the input to residual network too much. Because different types of long-term information have different weights, we insert channel attention layer at the end of RCAG. The g-th RCAG is formulated as:

$$\begin{aligned} F_g &= H_g(F_{g-1}) \\ &= \beta * F_{g-1} + C_g(B_{g,b}(B_{g,b-1}(...B_{g,1}(F_{g-1})...))), \end{aligned} \tag{4}$$

where $B_{g,b}$ denotes the b-th RCSAB in g-th RCAG, C_g denotes the channel attention layer in the g-th RCAG, H_g denotes the g-th RCAG and F_{g-1} and F_g are the input and output of the g-th RCAG.

2.2. Attention Mechanism

The features extracted by a deep network contain different types of information across channels and spatial regions which contribute different importance for high-frequency details recovery. Inspired by this, feature map and spatial region that contribute more in recovering high-frequency details will be payed more attention to learn the most import feature. Residual Channel and Spatial Attention Block (RCSAB) is produced as our base block, the structure of which is illustrated in Fig. 1d. The b-th RCSAB is formulated as:

$$F_b = B_b(F_{b-1}) = F_{b-1} + H_{CA}(W_b^2 \delta H_{SA}(W_b^1)), \tag{5}$$

where H_{CA} denotes the channel attention layer, H_{SA} denotes the spatial attention layer, B_b denotes the b-th RCSAB, F_b and F_{b-1} are the input and output of the b-th RCSAB, W_b^1 and W_b^2 are weight sets of the two convolution layers in the b-th RCSAB.

Channel attention (CA) has two advantages in SR task. On one hand, each filter in the base block Conv layer uses local receptive field, so the output after convolution cannot utilize the context information outside the local region, while Global Average Pooling (GAP) in Channel attention incorporates channel-wise global spatial information. On the other hand, different channels in feature maps play different roles such as extracting low frequency or high frequency components. In SR tasks that are difficult to reconstruct high-frequency details, channel attention layer is often helpful to assign higher weight to the channel for extracting high-frequency components. The remaining low frequency component extraction channels are assigned lower weights.

As shown in Fig. 1(c) and (d), we follow the structure of channel attention in SENet [9] and place the channel attention layer at the end of the base block and RCAG. We use X_c to denote the input of the channel attention layer, which consists of c feature maps, then a GAP layer is used to shrink X through spatial dimensions, the output is the channel-wise summary statistic (mean values) $\mathbf{z}$, as the input of two fully connected layers and two activation layers:

$$s = \sigma(W_2 \delta(W_1 \mathbf{z})), \tag{6}$$

where $\delta(\cdot)$ and $\sigma(\cdot)$ denote ReLU function and the sigmoid gating, respectively. $W_1 \in R_{\frac{c}{r}\times c}$ and $W_2 \in R_{c\times\frac{c}{r}}$ are the weight sets of fully connected layers, which play the roles of channel compression and channel expansion, respectively, r is the channel reduction ratio. In our experiments, we set r to 16, which is the same as RCAN. The sigmoid gating is to adapt the values between 0 and 1. The final channel statistics s is used to rescale the input x:

$$\hat{\mathbf{x}} = H_{CA}(\mathbf{X}) = \mathbf{s}\mathbf{x}, \tag{7}$$

where H_{CA} denotes the channel attention layer, s and x are the scaling factors and feature maps before scaled. With the above process, the base block is adaptively rescaled according to the channel-wise statistics of input to boost the channel-wise feature discriminability.

Spatial attention (SA). [22] proposed a new activation unit called xUnit, which is particularly suitable for image restoration problems. In contrast to the widespread per-pixel activation units like ReLUs and sigmoids, this unit implements a learnable nonlinear function with spatial connections. Inspired by xUnit, we simplified the xUnit structure for the SR problem, by removing the BN layer that is not suitable for the SR problem, and placing the spatial attention part after the ReLU layer. As shown in Fig. 1d, the spatial attention weight map is constructed by passing the input through a depth-wise convolution and an element-wise gating function ([22] used a Gaussian, we use a sigmoid) which maps the dynamic range to $[0, 1]$:

$$\hat{\mathbf{x}} = H_{SA}(\mathbf{x}) = \delta(W_{depth-wise}\mathbf{x})\mathbf{x}, \tag{8}$$

where H_{SA} denotes the spatial attention layer, $X = [X_1, \cdots, X_c, \cdots, X_C] \in R^{H\times W\times C}$ denotes the feature maps before spatial attention layer, $W_{depth-wise}$ denotes the weight of the depth-wise convolution layer, $\delta(\cdot)$ denotes the sigmoid gating, and $\hat{\mathbf{x}}$ is the feature maps after spatial attention layer.

Different channels have different importance, for example, in the feature map with complex edges and texture feature, the complex filter is more important. However, for different regions in the same channel, if we can focus on regions with more high-frequency details, it will help the recovery of the final high-frequency details. The proposed SA layer convolve each channel on themselves separately, rather than performing a squeeze between channels. On one hand, it helps to maintain channel-specific characteristics. On the other hand, depth-wise convolution has very low computational complexity, which does not increase the complexity of the model.

2.3. Up-sampling module based on IDWT

For transformations from low-scale to high-scale, aiming at the problem that SR texture details are difficult to restore, an up-sampling module based on inverse wavelet transform is proposed, as shown in Fig.1a. Each time Haar IDWT generates a $2\times$ resolution map by 4 small resolution feature maps, then 4^J feature maps can be reconstructed to $2^J\times$ resolution image using multiple Haar IDWT.

$$F_0, F_1, ..., F_{2^J-1} = f_{tail}(F_{DF}), \tag{9}$$

$$\begin{aligned} I_{SR} &= H_{UP}(F_{DF}, I_{LR}) \\ &= IDWT((2^J-1)I_{LR} + F_0, F_1, ..., F_{2^J-1}), \end{aligned} \tag{10}$$

where F_0 is the low-frequency of the SR image, and other feature maps are the high-frequency of the SR image, f_{tail} denotes the last convolution of the whole network to mapping the deep features to wavelet coefficients of each high-frequency sub-bands and the residual of low-frequency and LR image input, $2^J\times$ denotes the magnification factor.

The mean value of the low-frequency sub-band of the J-level wavelet transform is 2^J times of the mean value of the bicubic $2^J\times$ sub-sampled LR image. Therefore, we can learn the high-frequency sub-band of the SR image by learning the residual of LR image by Eq. (10), and the residual between the mean and sub-band share the same low-frequency as input.

3. EXPERIMENTS

3.1. Settings

Datsets and Metric. Following [4] [20], we trained our networks using 800 images from DIV2K. DIV2K consists of 800 training images, 100 validation images, 100 test images and their corresponding LR training images for $\times2$,$\times3$, and $\times4$ downscaling factors. Similar to EDSR, we also adopt self-ensemble strategy to further improve our WRAN and denote the self-ensembled WRAN as WRAN+. The SR results are evaluated with PSNR and SSIM [23] on Y channel after converted to YCbCr space. In ablation studies, we test each epoch using the first ten images of the validation sets during the training period. Four SR standard benchmarks are used to test our final model: Set14[24], B100[25], Urban100 [26], and Manga109 [27].

Training settings. Data augmentation is performed on the 800 training images, which are randomly augmented by mirror transformation and rotation. Our models are optimized using the Adam optimizer with β_1 = 0.9, β_2 = 0.999, and $\epsilon = 10^{-8}$ to minimize the L1 loss function:

$$\frac{1}{N}\sum_{i=1}^{n}\|x_i^{gt} - H_{WSAN}(x_i^{lr})\|_1, \tag{11}$$

where x_i^{gt}, x_i^{lr} denote the i-th ground truth sample and LR sample. The initial leaning rate is set to 10^{-4} and then decreases to half in 150^{th}, 200^{th} and 225^{th} epoch. Images are randomly cropped to 48×48, and 16 LR color patches are extracted as inputs when training. PyTorch is used to implement our models with one GTX1080Ti.

Model implementation details. We set the size and number of filters as 3×3 and 64 in all convolution layers except those for the upscaling part, depth-wise convolution in SA layer, and the first convolution to do feature extraction, this is the same as the RCAN. Our baseline model is tested on 2x factor within 10G (RCAGs) and 10B (RCSABs). The final model's configuration is 10G and 20B.

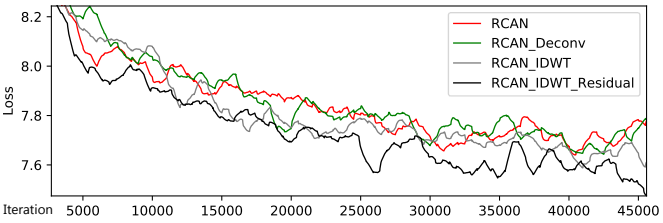

Fig. 2. Loss-Iteration figure of different up-sampling layer.

3.2. Model analysis

Structure of RCAG and CA in RCSAB. RCAN is our baseline to research the structure of RCAG - whether the deletion of the convolution layer and the addition channel attention at the end of the group is effective. Based on the no_Conv_t_GCA(setting D) RCAG structure, we further insert the channel attention layer in different locations in the baseline base block. Results and details are shown in table 1. The number is the location of the channel attention layer. 0 represents the position before the ReLU active layer, 1 represents the position after the ReLU active layer, and 2 represents the position after the second convolution layer. Two digits represent the insertion of channel attention layer at two different locations. As illustrated in Table 1, the best location of CA is after the second convolution, the performance of other locations are relatively backward, and if we add channel attention in both 0 and 2 positions, the network will converge to the local optimum and get poor performance.

Because the tested PSNR after each epoch during the training period will be slightly unstable jitter, the PSNR point in the graph is the average index of 7 epoches adjacent to it. Note that we have not yet introduced IDWT-based up-sampling now.

Effectiveness of IDWT-based up-sampling. We find that the proposed up-sampling method converges faster than previous ones. We

Settings	A	B	C	D	E	F	G	H	I	J
Up-Sampling	SP	SP	SP	SP	SP	SP	SP	SP	IR	IR
Tail Conv	1	1	0	0	0	0	0	0	1	0
Group CA	0	1	0	1	1	1	1	1	0	1
CA Location	2	2	2	2	0	1	2	12	2	2
PSNR	35.892	35.965	35.996	36.008	35.921	35.960	15.567	35.950	35.998	36.019

Table 1. Performance of different structures in optimal setting (B=20). "SP" is the Sub-Pixel Conv and IR is the IDWT_Residual. Setting A is the baseline model RCAN. Setting D has the best group setting and CA block setting. Setting I has the best up-sampling setting, and Setting J is our final model.

take RCAN as baseline. A 3×3 conv and a Sub-Pixel [19] layer are used as the up-sampling layer in RCAN. We also replace the Pixel Shuffle layer with a deconv layer for up-sampling layer as comparison to prove the superiority of our method. In addition, we remove the residual connections in our method for a set of experiments, which prove the necessity to learn the residual of LR image and low-frequency to control the mean of features of low-frequency the same as LR. As illustrated in Fig. 2, our IDWT-based up-sampling method makes convergence fastest. Experimental results under different structures (Table 1) demonstrates that the effectiveness of IDWT-based up-sampling.

Effectiveness and efficiency of spatial attention. Our proposed spatial attention(SA) layer allows us put a few parameters to get the improvement of performance. A network with weaker performance increase more obviously, but the proportion of parameters of spatial attention layers is higher. Different size of the convolution kernel of the Depth-Wise convolution in the SA layer will affect the proportion of SA layer parameters and the final performance of the network, which is a trade-off. We use our proposed WRAN model to conduct a control experiment with different sizes of convolution kernels. The experimental results are shown in Table 2. The input image size of memory and FLOPS is $3 \times 16 \times 16$, PSNR is the average test value of on DIV2K 801-810. Our experimental results show that the introduction of the SA layer whose DepthWise Conv kernel size larger than 5×5 leads to higher PSNR and brings about 1.66 times of memory usage.

We further observe the changes in feature maps before and after the SA layer to analyze its role. An example feature map before Spatial Attention and the corresponding SA weights is shown in Fig. 3. We can clearly see from 3c that our spatial attention layer can give higher weight to the detailed high frequency region, which proves the effectiveness of our proposed spatial attention layer.

Fig. 3. (a) Baby from "Set5" (b) Example feature map before SA (c) Corresponding Spatial Attention weights of (b).

3.3. Comparison with the-state-of-the-arts

Several state-of-the-art methods are compared in terms of quantitative evaluation (Tab. 3) and visual quality to prove the effectiveness of WRAN. Since WRAN introduces IDWT, *our method does not support 3× SR, which is one limitation of our method.*

Quantitative evaluation. We show quantitative comparisons for 2× and 4× SR in Table 3. Without a self-ensemble strategy, our WRAN outperforms the existing methods in 2× and 4× scale. Using self-ensemble, WRAN+ can achieve better results. Our method has much better performance in Manga109 dataset, for images in Manga109 dataset consist of a large number of flat areas and edges, which benefit by our SA layer greatly. RCAN and WRAN performance much better than other methods that is not very deep also indicate that study about how to stack a deeper model in SR task is meaningful.

Table 2. Parameters, memory usage, calculation amount and performance data w/o and with different SA layer in baseline setting (B=10). DWKS refers to the kernel size of DWConv in SA layer.

Methods-DWKS	Parameters	Memory	FLOPS	PSNR
RCAN	8.001K	20.09MB	4.064G	35.756
WRAN	7.495K	18.94MB	3.799G	35.768
WRAN-3×3	7.559K	31.44MB	3.829G	35.750
WRAN-5×5	7.661K	31.44MB	3.881G	35.774
WRAN-7×7	7.815K	31.44MB	3.960G	35.780
WRAN-9×9	8.019K	31.44MB	4.065G	35.787

Table 3. Quantitative evaluation of state-of-the-art SR algorithms: average PSNR/SSIM for scale factors 2× and 4×. Best results are in red and the second best results are in blue.

Methods	Scale	Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×2	30.38	0.868	29.56	0.843	26.88	0.840	30.81	0.934
SRCNN[3]	×2	32.45	0.906	31.36	0.887	29.51	0.894	35.64	0.966
VDSR[12]	×2	33.05	0.912	31.90	0.896	30.77	0.914	37.33	0.974
LapSRN[17]	×2	33.00	0.912	31.52	0.895	30.41	0.910	37.27	0.974
MemNet[14]	×2	33.28	0.914	32.08	0.897	31.31	0.919	37.81	0.974
EDSR[4]	×2	33.92	0.919	32.32	0.901	32.93	0.935	39.13	0.977
SRMD[20]	×2	33.32	0.915	32.05	0.898	31.33	0.920	38.07	0.976
DBPN[18]	×2	33.84	0.919	32.27	0.900	32.55	0.932	38.89	0.977
RDN[7]	×2	34.14	0.921	32.34	0.901	32.89	0.935	39.18	0.978
RCAN[7]	×2	34.12	0.921	32.41	0.902	33.34	0.938	39.44	0.978
WRAN	×2	34.24	0.923	32.43	0.902	33.38	0.938	39.66	0.979
WRAN+	×2	34.35	0.924	32.47	0.903	33.57	0.940	39.83	0.979
Bicubic	×4	26.00	0.702	25.96	0.667	23.14	0.657	24.89	0.787
SRCNN[3]	×4	27.50	0.751	26.90	0.712	24.52	0.722	27.56	0.855
VDSR[12]	×4	28.02	0.767	27.29	0.712	25.18	0.753	28.85	0.886
LapSRN[17]	×4	28.15	0.769	27.32	0.726	25.21	0.755	29.09	0.889
MemNet[14]	×4	28.26	0.773	27.40	0.728	25.50	0.763	29.42	0.895
EDSR[4]	×4	28.80	0.787	27.71	0.742	26.64	0.803	31.03	0.915
SRMD[20]	×4	28.35	0.778	27.49	0.733	25.68	0.773	30.09	0.902
DBPN[18]	×4	28.80	0.786	27.71	0.739	26.38	0.794	30.91	0.913
RDN[7]	×4	28.81	0.787	27.72	0.741	26.61	0.802	31.00	0.915
RCAN[7]	×4	28.87	0.788	27.77	0.743	26.82	0.808	31.22	0.917
WRAN	×4	28.95	0.790	27.79	0.743	26.89	0.808	31.46	0.919
WRAN+	×4	29.00	0.791	27.83	0.744	27.04	0.812	31.71	0.921

Visual Results. As shown in Fig. 4, for image "Wareware-HaOniDearu", since the texture is seriously lost by downsampled, the fuzzy details make the reconstruction very challenging. While other methods fail to recover the texture region on the background of this image, ours can reconstruct much closer details to the ground truth. For the text portion of the bottom right corner, it can be obviously seen that pre-interploation methods generate blurry results, methods starting from LR input recovers better, while results reconstructed by our methods recovered sharper.

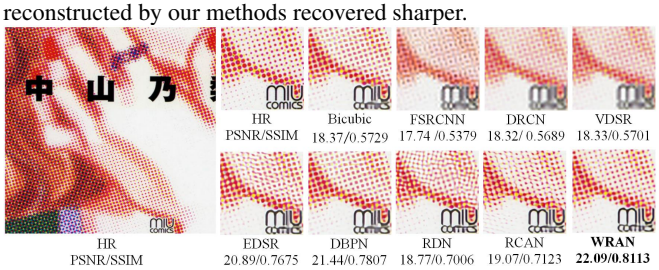

Fig. 4. Visual comparison for 4× SR in Manga109 dataset. The best results are **highlighted**.

4. CONCLUSION

In this paper, we propose a residual attention network for wavelet domain super-resolution. Channel attention and spatial attention are integrated in our base block. Furthermore, we propose an IDWT-based up-sampling module which can replace deconv or Sub-Pixel layer for up-sampling in SR tasks. Comprehensive evaluations on benchmark datasets demonstrate that the proposed network performs better than the state-of-the-art SR methods in terms of quantitative and qualitative measurements.

5. REFERENCES

[1] Wilman WW Zou and Pong C Yuen, "Very low resolution face recognition problem," *IEEE Transactions on image processing*, vol. 21, no. 1, pp. 327–340, 2011.

[2] Wenzhe Shi, Jose Caballero, Christian Ledig, Xiahai Zhuang, Wenjia Bai, Kanwal Bhatia, Antonio M Simoes Monteiro de Marvao, Tim Dawes, Declan O'Regan, and Daniel Rueckert, "Cardiac image super-resolution with global correspondence using multi-atlas patchmatch," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 9–16.

[3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[4] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 136–144.

[5] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov, "Generating images from captions with attention," *arXiv preprint arXiv:1511.02793*, 2015.

[6] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5659–5667.

[7] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 286–301.

[8] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 606–615.

[9] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[10] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[11] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481.

[12] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[14] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu, "Memnet: A persistent memory network for image restoration," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4539–4547.

[15] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[16] Chao Dong, Chen Change Loy, and Xiaoou Tang, "Accelerating the super-resolution convolutional neural network," in *European conference on computer vision*. Springer, 2016, pp. 391–407.

[17] 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*, 2017, pp. 624–632.

[18] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita, "Deep back-projection networks for super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1664–1673.

[19] Wenzhe Shi, Jose Caballero, and booktitle= Huszá"r, Ferenc and Totz, Johannes and Aitken, Andrew P and Bishop, Rob and Rueckert, Daniel and Wang, Zehan, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," .

[20] Kai Zhang, Wangmeng Zuo, and Lei Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3262–3271.

[21] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[22] Idan Kligvasser, Tamar Rott Shaham, and Tomer Michaeli, "xunit: learning a spatial activation function for efficient image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2433–2442.

[23] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[24] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[25] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *null*. IEEE, 2001, p. 416.

[26] 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*, 2015, pp. 5197–5206.

[27] 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*, vol. 76, no. 20, pp. 21811–21838, 2017.