M.S. THESIS

Hard Sample Handling using Attention Ensemble Network for Single Image Super-Resolution

단일 영상 초해상도 복원을 위한 어텐션 앙상블 네트워크 기반의 하드 샘플 처리 기법

BY

Kwak Jun-Hyung

FEBRUARY 2019

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

M.S. THESIS

Hard Sample Handling using Attention Ensemble Network for Single Image Super-Resolution

단일 영상 초해상도 복원을 위한 어텐션 앙상블 네트워크 기반의 하드 샘플 처리 기법

BY

Kwak Jun-Hyung

FEBRUARY 2019

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Hard Sample Handling using Attention Ensemble Network for Single Image Super-Resolution

단일 영상 초해상도 복원을 위한 어텐션 앙상블 네트워크 기반의 하드 샘플 처리 기법

지도교수 이 경 무

이 논문을 공학석사 학위논문으로 제출함

2019년 2월

서울대학교 대학원

전기 정보 공학부

곽 준 형

곽준형의 공학석사 학위 논문을 인준함

2019년 2월

위 원 장: ______________________

부위원장: ______________________

위 원: ______________________

Abstract

Single image super-resolution is developing very quickly, but it has been focused on simply improving performance for all pixels. However, in this problem, the degree of error varies depending on the characteristics of each pixel position. Therefore, it may be effective to focus on particular pixels with large error. An additional system is needed to handle hard pixels. Because we can get another type of hard pixels from the system, it is ideal system to complement each other's hard pixels.

An ensemble is a way to easily increase the performance of the system and is also used in machine learning. An ensemble consists of the systems which complement each other. Therefore, the above problem is solved by ensemble method, so that each network can complement each other's hard pixels. Also, unlike normal ensemble methods, the difficulty of each pixel is different, so the weight for a pixel must be different. Also, information about which pixels to focus on should be communicated. Since the attention network is well suited to this role, the attention method is additionally applied to the ensemble. We achieved higher performance than the existing algorithm through the proposed method. We also defined the hard pixel and measure their accuracy. It was shown that they are complementary when applying the attention network, and the whole system is particularly robust to hard pixels.

keywords: Single Image Super-resolution, Hard sample handling, Deep learning, Ensemble, Attention

student number: 2016-20864

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

SISR is an old problem that restoring high resolution image from low resolution image by downsampling such as bicubic. This problem is also an ill-posed problem the solution is not unique. It is usually applied to many computer vision problems because it improves quality of an image.

The basis works of SISR studies are a bicubic interpolation and resampling [6]. SR progresses to example-based methods[18], which make a dictionary of low/high patch pair, sparse-coding-based methods [17], which code an image patch. For restoring a high resolution image, the system should exploit the dictionary.

Recently, deep learning succeeds in resolving computer vision problems. For example, Resnet even outperforms human's ability in image classification. On the way, single image super-resolution has also been improved through deep learning. SRCNN [7] is the first work to apply deep learning to super-resolution. But the work has a limitation that the receptive field of SRCNN is small because they fail to stack deeper than three layers. VDSR solve the problem by adopting global skip connection. One step further, SRResnet [10] uses local skip connection as well as global skip connection by applying the residual block to their network. EDSR [11] shows a more efficient structure for super-resolution.

As shown in figure 1.1, most errors of an super resolution image occur only in certain pixels. However, most of works about super-resolution have focused on not a specific region but the accuracy of an image. It may be effective to concentrate on hard pixels, where most errors are occurred. So we approach to improve performance by hard pixel handling. If we introduce a new system to handle the hard pixel of the original system, there will be another hard pixel by the new system. The most appropriate solution is to allow the original system to handle the new hard pixel. Figure 1.1 shows this idea.

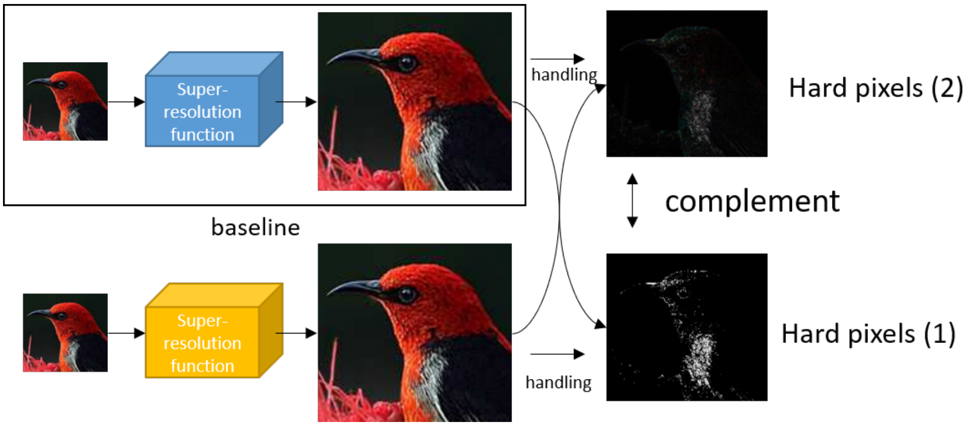


Figure 1.1: Example of error map, easy pixel, hard pixel. Large error pixels are hard pixel, and conversely small error pixels are easy pixel.

To solve this problem, the ensemble methods are employed in this thesis. Ensemble method complements the weaknesses of weak learners in the ensemble to get better performance. So we introduce a system that assembles multiple baseline networks to complement each other's hard pixels. But some issues still remain. An ordinary ensemble method does not consider which pixel is a hard pixel. For example, weight average ensemble is a general ensemble form of an ensemble but it can not be applied in a pixelwise manner. In addition, it can not convey information about difficulty of a pixel. We apply attention network to generate weight mask. By doing this, the weight mask determines how hard a pixel is and the hardness is applied pixelwise.

In this thesis, an ensemble network system using attention networks is proposed. In the system, the attention network emphasizes the hard pixel of each network in the ensemble. As a result, we can tackle the hard pixel which the network can not deal with and outperform weight average ensemble. We also show the proposed method actually complements the hard pixels of each network.

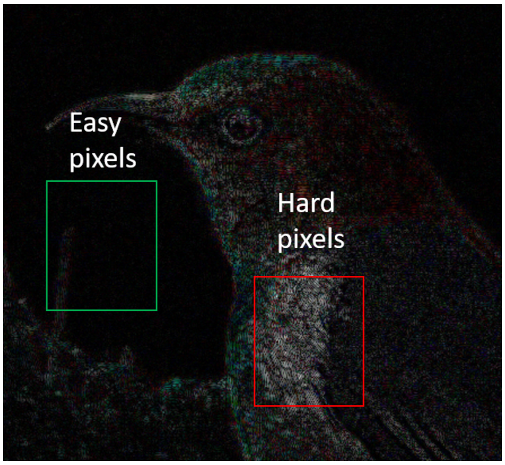


Figure 1.2: Concept of hard pixel handling for super-resolution. Each super-resolution system handle each other system's hard pixels.

1.2 Outline

In Chapter 2, studies on the super-resolution, ensemble and attention techniques are described. Recent researches on hard example handling are also described. In Chapter 3, the structure and the training method of the proposed attention ensemble network (AttEnsNet) are introduced. In Chapter 4, the experimental results of the proposed methods are shown. The quantitative and quallitative comparison of the performance are reported. The robustness to hard pixels is also studied Finally, in Chapter 5, we conclude this thesis.

1.3 Contributions

In this thesis, there are three contributions. First, this is a new approach to super-resolution in terms of hard sample handling. Hard sample handling is the key to improving performance. The quantitative experiments show that the proposed method achieves better performance against baseline results. Second, the proposed method is not an ordinary ensemble method but a pixelwise ensemble. In super-resolution, pixelwise ensemble is more appropriate. Last, we propose methods to evaluate performance in hard pixels. It is possible to show how effective ensemble is by simply showing the performance evaluation in hard pixels as well as the hard pixel evaluation method in ensemble. The hard pixel complementary experiment proves that the proposed method is effective in hard pixels and efficient ensemble.

Chapter 2

Related Works

2.1 Single Image Super-Resolution

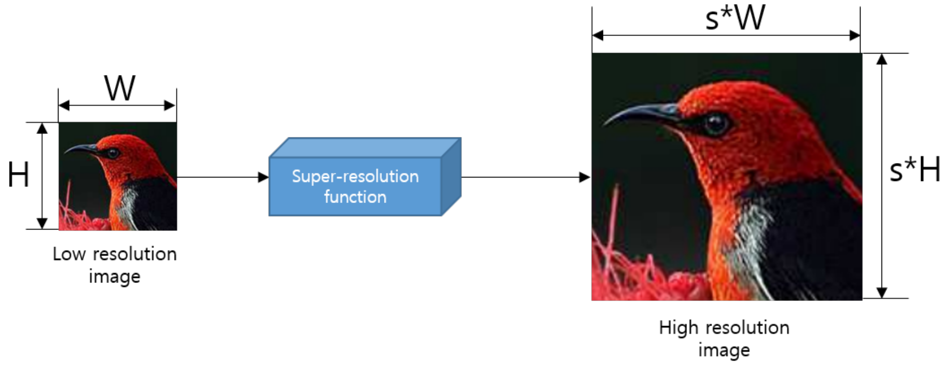


Figure 2.1: Super-resolution example

Single image super-resolution is an old and low-level computer vision problem. This helps in satellite images, medical images, microscopic images, and astronomical images. Previous studies have attempted to approach example-based [18] and sparse-coding-based [17]. An example-based method builds a low / high resolution patch pair in a dictionary and finds the nearest high resolution patch in the dictionary by using a method like as nearest neighbor (NN). The sparse coding method encodes a low resolution image to a sparse coefficient and reconstructs it using a dictionary.

Recently, there have been many studies on deep learning bases. SRCNN [7], which is the first deep learning-based super-resolution work, show that deep learning is effective in restoring super-resolution by stacking 3 convolution layers. They interpret the effectiveness as sparse-coding-based method. Each layer of SRCNN corresponds to patch extraction and representation, non-linear mapping, and reconstruction. The performance is better than non-deep-learning-based method. However, they fail to obtain good performance when the network has more than 3 layers. VDSR [8] adopt global skip connections to deeper SRCNN structure and it allows to stack deeper layers. The global skip can easily recover low-frequency components. The correlation between the performance and the receptive field is shown in this work. The residual block of Resnet [31] is effective in image classification. SRResnet [10] showed that it can improve performance by applying the residual block to SR as with Resnet. More recently, papers have been published that use more and more connections, such as EDSR [11] and RCAN [14]. EDSR proposed an effective residual block in the wider and deeper network to achieve improved performance. RCAN proposed a Residual-in-Residual structure for SR and channel attention block to exploit channel-specific information from a feature map.

Most studies focus on improving quantitative performance such as PSNR. They use traditional low-level loss such as L1, L2, but it is not suitable for reconstructing structured images. It results in blurry imaging results. To handle this issue, SRGAN [10] tries a perceptual approach with adversarial loss and contextual loss in order to obtain a more natural image. The adversarial loss makes the network to generate realistic patches, and contextual loss makes the network to generate the patch with the same context. SRFEAT [16] further improved SRGAN by applying an adversarial loss on the feature map.

2.2 Ensemble

An ensemble method is a simple and effective way to increase the performance of the system. A specific learning techniques might yield not optimal models. So the idea of ensemble method would be to create a more perfect system by combining multiple systems that is not perfect, in other words, weak learner. Because of that effect, the machine learning model usually uses an ensemble method to improve performance in practical cases. Examples of using ensemble technique are logistic regression, random forest, and decision trees. Common methods of ensemble methpd include average weight, voting, bagging, boosting. The average method is to obtain the ensemble output by multiplying the result value and the weight, and voting is a method of selecting the ensemble output among the result values. Bagging is a method of dividing a sample into several samples, learning each model, and aggregating the results. Boosting makes several weak learners iterative with boot-trapped test data. The i-th learner learns by boosting the mispredicted data of the $i - 1$th learner. Finally, we predict using the last generated learner. In super-resolution, [12] applied the ensemble to SR with weighted average method which is commonly used. It did not leads to a significant performance improvement. In this thesis, we use an attention as a weight mask to boost the performance gap.

2.3 Attention

Attention helps emphasize important parts of the input processing. So that the model can focus on the important information. Generally, the final output is obtained with a gating function such as sigmoid. There are several works using attention in field of speech recognition [26], machine translation [27], computer vision [28, 29, 30]. Wang et al [28] propose trunk-and-mask caution mechanism for image classification. Yao et al. [30] is a study of video description and it introduces the temporal attention structure to focus on the most relevant temporal segments. The general form of attention is shown in Figure 2.2. The created attention map is usually multiplied by some feature map and added to another feature map in the mainstream. For example, $output = x + ay$. In this paper, we apply slightly modified form $output = ax + (1 - a)y$ to the ensemble, where x and y are outputs of two different systems, respectively, and a is an attention mask.

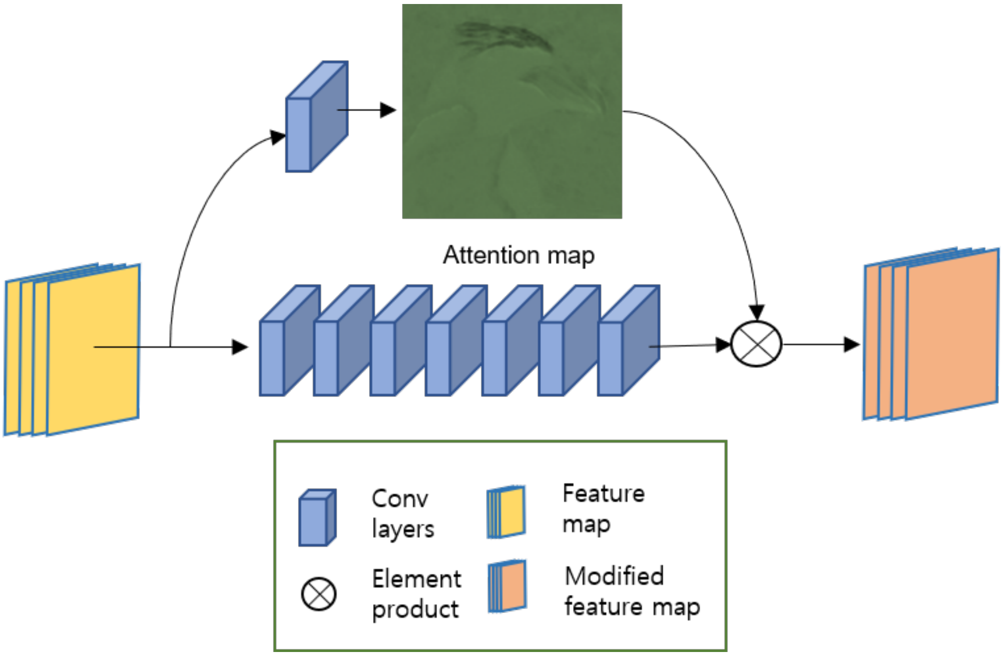


Figure 2.2: Example of attention

2.4 Hard Sample Handling in Computer Vision

Hard sample handling is an approach that improves performance by doing considerations to samples that are not easily handled by algorithms. It is widely used in detection [23]. It is important to focus on hard sample because there are a number of proposals in the bounding box in Detection [25]. Since the number of negative samples is overwhelmingly larger than the number of positive samples, the performance of the detection model can be improved by mining hard negative samples. Other highlevel computer vision applications include segmentations [21]. They focus on the fact that all pixels of the segmentation mask do not have the same degree of difficulty. In low-level vision, there are studies which focus on easy samples. There are studies that define the robust loss to eliminate outliers [20], use the EM method [22], or apply curriculum learning [19]. However, there are not many studies to treat hard samples preferentially at low-level. In this thesis, we focus on the hard sample rather than the easy sample.

Chapter 3

Proposed Method

In this chapter, we describe the background of designing a network structure from hard pixel handling and propose an Attention ensemble structure based on the background.

3.1 Backgroud of Proposed Method

The ensemble compiles the results of each system and produces the final result. If we apply a general ensemble method such as a weighted average ensemble to the super-resolution, each output of the network is multiplied by constant, and it is the final result. However, a network in super-resolution is not effective because the restoration difficulty differs for each pixel. We have to find a way to achieve ensemble in pixelwise. However, there is no way to know in advance which pixels are difficult and which pixels are easy. Pixel difficulty mask is also based on learning. To do so, we introduce attention to create a mask that emphasizes hard pixels.

3.2 Attention Ensemble Network

The overall structure of the attention ensemble network consists of attention networks and baseline networks. The baseline network used in this thesis is EDSR [11]. Af-

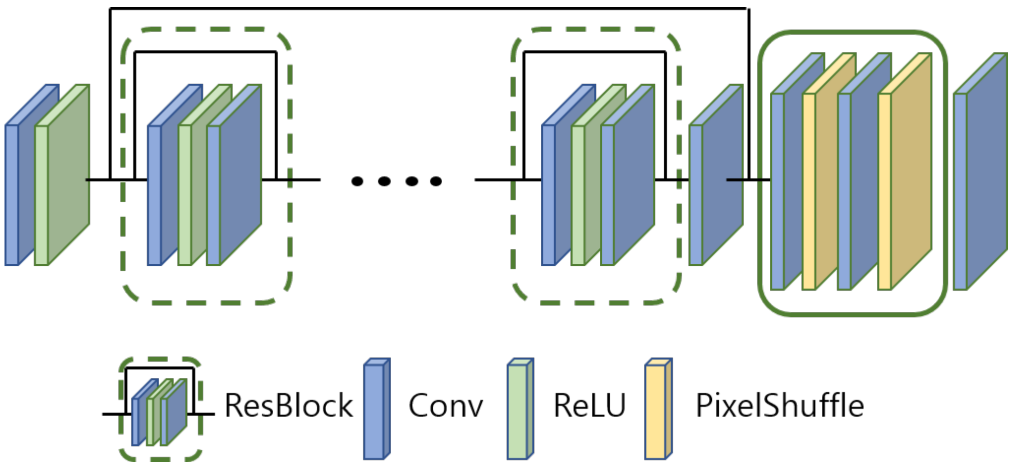


Figure 3.1: Baseline network structure

ter describing the structure of the attention network, we describe the overall system structure.

Unlike general attention adding the product of attention mask and feature map to the main feature map, we use the attention mask as a weight mask to ensemble the network outputs. To express the average weighted sum, we set the output of attention to between 0 and 1. Therefore, Sigmoid is used as the activation function of attention mask. We created masks for all 3-channel colors to get a more accurate mask. The structure of the attention network is shown in Table 3.1. The network takes two images as input and concatenate those. The layer column contains information about the convolution type. The activation function differs only to the last sigmoid. The number indicates how many times the layer has been stacked.

Another attempt has been to add a batch normalization layer to the structure and replace the attention network with a residual network. However, both methods have similar or slightly lower performance. In the case of depth, there is a problem that the performance is lowered. This seems to be caused by gradient vanishing. Since the receptive field of the attention network is significantly smaller than the baseline network, it can be a problem of the proposed attention network.

Figure 3.2 is the default form of the proposed attention ensemble network. The yellow box is the attention network, and the blue box is the baseline network. A green box means ensemble the output of the output of connected boxes. The EDSR baseline [11] is used as the base network, as shown in Figure 3.1. In this thesis, we extended the ensemble size to three and four in Figure 3.3. There is no experiment of three baseline case. When three or more baseline networks are used, the ensemble structure of the previous version is combined with new baseline networks. The formula is expressed as

$$y_2 = \alpha * x_1 + (1 - \alpha) * x_2$$

$$y_3 = \beta * y_2 + (1 - \beta) * x_3$$

$$y_4 = \gamma * y_2 + (1 - \gamma) * y_2'$$

x_n means the output of the $n - th$ baseline network, α, β, γ means the output of $n - th$ attention network respectively, and y_n means ensemble with n networks. If we use softmax, we can achieve ensemble by one attention network but in that case, we get a little lower performance than the proposed method.

Table 3.1: Structure of Attention Network

Layer	Activation	Number
Conv(6,64), k3s1	ReLU	1
Conv(64,64), k3s1	ReLU	12
Conv(64,3), k3s1	Sigmoid	1

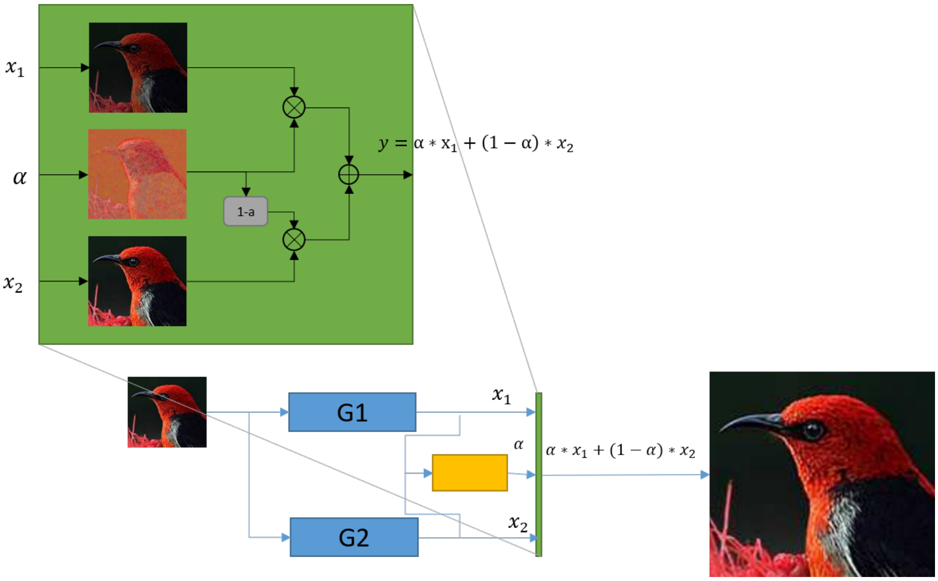

x_1
α
x_2
1-a
$y = \alpha * x_1 + (1-\alpha) * x_2$
G1
G2
x_1
α
x_2
$\alpha * x_1 + (1-\alpha) * x_2$

Figure 3.2: Structure of ensemble with two networks

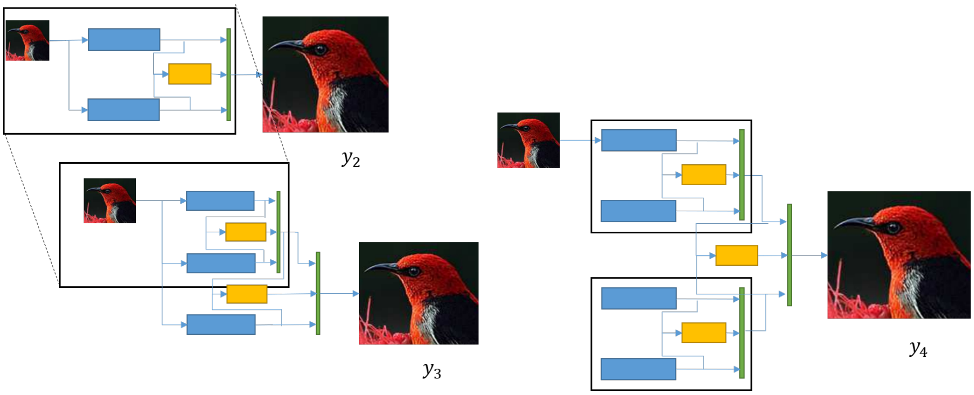


Figure 3.3: Illurstration of how to expand ensemble

Chapter 4

Experiment

In this chapter, we first discuss the training and test data and describe how we have trained the network. Next, the experimental results of the proposed method, the baseline and previous methods are compared quantitatively and qualitatively. Finally, we have added an experiment on how the proposed method contributes to hard pixels.

4.1 Dataset

The DIV2K dataset [1] is total of 1000 image dataset, consisting of 800 training dataset, 100 validation sets, and 100 test sets. This dataset comes with 2, 3, and 4x low resolution images, and includes a wide range of very high resolution images. Recently, it has been widely used as a training data set for the resolution restoration problem. In our experiments, we train the network with this training dataset.

Quantitative comparison is carried out by experimenting on Set5 [2], Set14 [3], BSD100 [4], and Urban100 [5] datasets, which are commonly used for super-resolution restoration problems.

4.2 Training Details

To achieve the maximum effect, ensemble networks and attention networks are trained together. The l1 objective function is also used because it gives a blurry result when the l2 objective function is used. The final objective function is as follows when the maximum of the pixel value is 255.

$$loss = \sum |x_{i,j,k}/255 - y_{i,j,k}/255|$$

x is ground truth and y is final output. y is output of attention ensemble or average ensemble.

The Adam optimizer function is used for training. Weight decay is set to 0 as in [11], and the learning rate is reduced by 0.5 times every 200,000 times until 1e6 iterations. Since our method uses already pretrained network, the initial learning rate is set to 4e-5. The patch size is 192 and batch size is 8. The applied augmentation is 90 ° rotation and horizontal and vertical flip. If three or more networks are used in an ensemble, they are trained from the pretrained previous ensemble version.

4.3 Results

The qualitative and quantitative results are shown in below. Experimental results on how well hard pixel is handled are also shown.

4.3.1 Quantative Results

This section shows quantitative results. There are many methods to compare, so we divide them into two tables. The first table shows the cases of the previous methods and the second table compares the baseline with the proposed method. Previous research methods to compare are bicubic interpolation, SRCNN [7], VDSR [8], and DRCN [9]. The scale factors in experiments are 2, 3, and 4. AvgEns, AttEns2, and AttEns4 means average ensemble, attention ensemble with 2 networks and attention ensemble with 4 networks, respectively. Bold indicates the best performance and underline indicates the second best performance. There are two measurements for comparing quantitative results in super-resolution: PSNR and SSIM. PSNR is an abbreviation of peak-signal-to-noise ratio and is an evaluation method for image quality loss information. PSNR can be expressed by the following equation.

$$PSNR = 10 \log_{10} \frac{MAX_I^2}{MSE}$$

SSIM is to measure the similarity of two images from brightness, contrast, and structure information. SSIM can be represented by the following formula.

$$\text{SSIM(x, y)} = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

For DIV2K dataset, we calculate PSNR and SSIM on RGB channel image, and in other cases, evaluate on Y channel image.

Table 4.1, 4.2 shows PSNR evaluation results and Table 4.3, 4.4 shows SSIM evaluation results.The proposed method with 2 and 4 baseline networks performs the second best and the best on all the datasets and all scaling factors. For Urban100, the PSNR gains of AttEns2 over the baseline are from 0.14 dB to 0.29 dB. Also, the increments of AttEns2 compared to AvgEns are from 0.06 dB to 0.07 dB. Therefore, the proposed method is effective.

Table 4.1: Quantative results in PSNR (dB)

Dataset	Scale	Bicubic	SRCNN	VDSR	DRCN
Set5	x2	33.66	36.66	37.53	37.63
	x3	30.39	32.75	33.66	33.82
	x4	28.42	30.48	31.35	31.53
Set14	x2	30.24	32.45	33.03	33.04
	x3	27.55	29.30	29.77	29.76
	x4	26.00	27.50	28.01	28.02
B100	x2	29.56	31.36	31.90	31.85
	x3	27.21	28.41	28.82	28.80
	x4	25.96	26.90	27.29	27.23
Urban100	x2	26.88	29.50	30.76	30.75
	x3	24.46	26.24	27.14	27.15
	x4	23.14	24.52	25.18	25.14
Manga109	x2	30.80	35.60	37.22	37.53
	x3	26.95	30.48	32.01	32.22
	x4	24.89	27.58	28.83	28.91
DIV2K val	x2	31.01	33.05	33.66	33.69
	x3	28.22	29.64	30.09	30.14
	x4	26.66	27.78	28.17	28.21

Table 4.2: Quantative results in PSNR (dB)

Dataset	Scale	EDSR	AvgEns	AttEns2	AttEns4
Set5	x2	37.99	38.03	38.05	**38.10**
	x3	34.37	34.43	34.47	**34.51**
	x4	32.15	32.14	32.21	**32.30**
Set14	x2	33.57	33.71	33.74	**33.79**
	x3	30.28	30.33	30.37	**30.41**
	x4	28.57	28.59	28.63	**28.66**
B100	x2	32.16	32.22	32.23	**32.26**
	x3	29.09	29.13	29.15	**29.17**
	x4	27.57	27.57	27.61	**27.63**
Urban100	x2	31.98	32.20	32.27	**32.38**
	x3	28.15	28.26	28.32	**28.38**
	x4	26.04	26.12	26.18	**26.20**
Manga109	x2	38.55	38.63	38.62	**38.70**
	x3	33.45	33.59	33.64	**33.72**
	x4	30.38	30.48	30.58	**30.64**
DIV2K val	x2	34.63	34.72	34.78	**34.85**
	x3	30.94	31.01	31.06	**31.10**
	x4	28.96	29.02	29.06	**29.09**

Table 4.3: Quantative results in SSIM

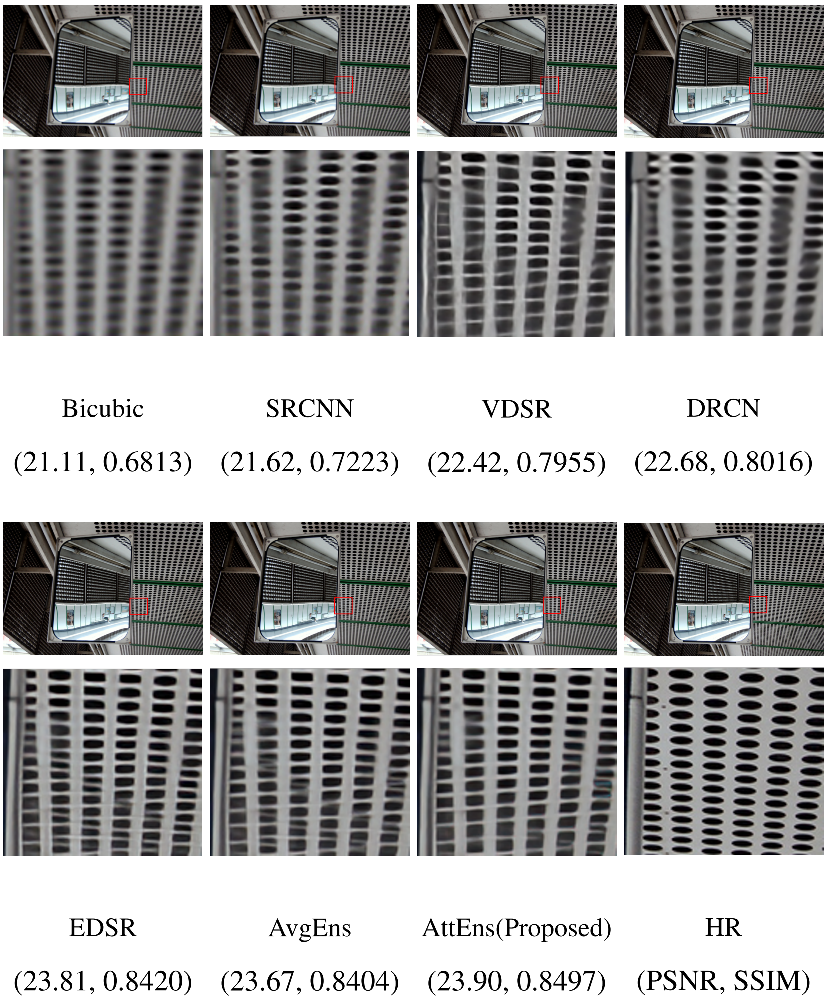
Dataset	Scale	Bicubic	SRCNN	VDSR	DRCN
Set5	x2	0.9299	0.9542	0.9587	0.9588
	x3	0.8682	0.9090	0.9213	0.9226
	x4	0.8104	0.8628	0.8828	0.8854
Set14	x2	0.8688	0.9067	0.9124	0.9118
	x3	0.7742	0.8215	0.8314	0.8311
	x4	0.7027	0.7513	0.7674	0.7670
B100	x2	0.8431	0.8863	0.8960	0.8942
	x3	0.7385	0.7863	0.7976	0.7963
	x4	0.6675	0.7101	0.7251	0.7233
Urban100	x2	0.8403	0.8946	0.9140	0.9133
	x3	0.7349	0.7989	0.8279	0.8276
	x4	0.6577	0.7221	0.7524	0.7510
Manga109	x2	0.9339	0.9663	0.9750	0.9731
	x3	0.8556	0.9117	0.9340	0.9339
	x4	0.7866	0.8555	0.8870	0.8848
DIV2K val	x2	0.9393	0.9581	0.9625	0.9619
	x3	0.8906	0.9138	0.9208	0.9204
	x4	0.8521	0.8753	0.8841	0.8835

Table 4.4: Quantative results in SSIM

Dataset	Scale	EDSR	AvgEns	AttEns2	AttEns4
Set5	x2	0.9604	0.9606	0.9606	**0.9608**
	x3	0.9270	0.9275	0.9275	**0.9279**
	x4	0.8943	0.8946	0.8953	**0.8961**
Set14	x2	0.9175	0.9188	0.9189	**0.9194**
	x3	0.8417	0.8422	0.8432	**0.8439**
	x4	0.7814	0.7822	0.7830	**0.7841**
B100	x2	0.8994	0.9002	0.9003	**0.9007**
	x3	0.8052	0.8060	0.8066	**0.8070**
	x4	0.7359	0.7368	0.7376	**0.7383**
Urban100	x2	0.9272	0.9295	0.9300	**0.9312**
	x3	0.8527	0.8547	0.8561	**0.8573**
	x4	0.7849	0.7872	0.7891	**0.7904**
Manga109	x2	0.9769	0.9771	0.9771	**0.9772**
	x3	0.9439	0.9450	0.9455	**0.9460**
	x4	0.9073	0.9085	0.9105	**0.9113**
DIV2K val	x2	0.9674	0.9680	0.9683	0.9682
	x3	0.9304	0.9312	0.9317	0.9322
	x4	0.8969	0.8977	0.8986	0.8991

4.3.2 Qualitative Results

Qualitative comparisons are shown in Figure 4.3 and 4.4 for scale x3, Figure 4.1 and 4.2 for scale x4. For comparison, the textures that are hard to restore cropped and enlarged. The textures have various shape. We can observe less blurry and less distorted structure in the result of the proposed method than others. For image “img004” and “img062”, ground truth is lattice structure. Usually, super-resolution output which takes a lattice structure as input creates lines in different directions. In the case of image “img004” and “img078”, the proposed method restores the lattice shape well while others generate lines that do not make sense. Image “img040” is a sprite pattern. In this case, there is an error that a line perpendicular to the pattern is generated upon restoration.



Bicubic (21.11, 0.6813)

SRCNN (21.62, 0.7223)

VDSR (22.42, 0.7955)

DRCN (22.68, 0.8016)

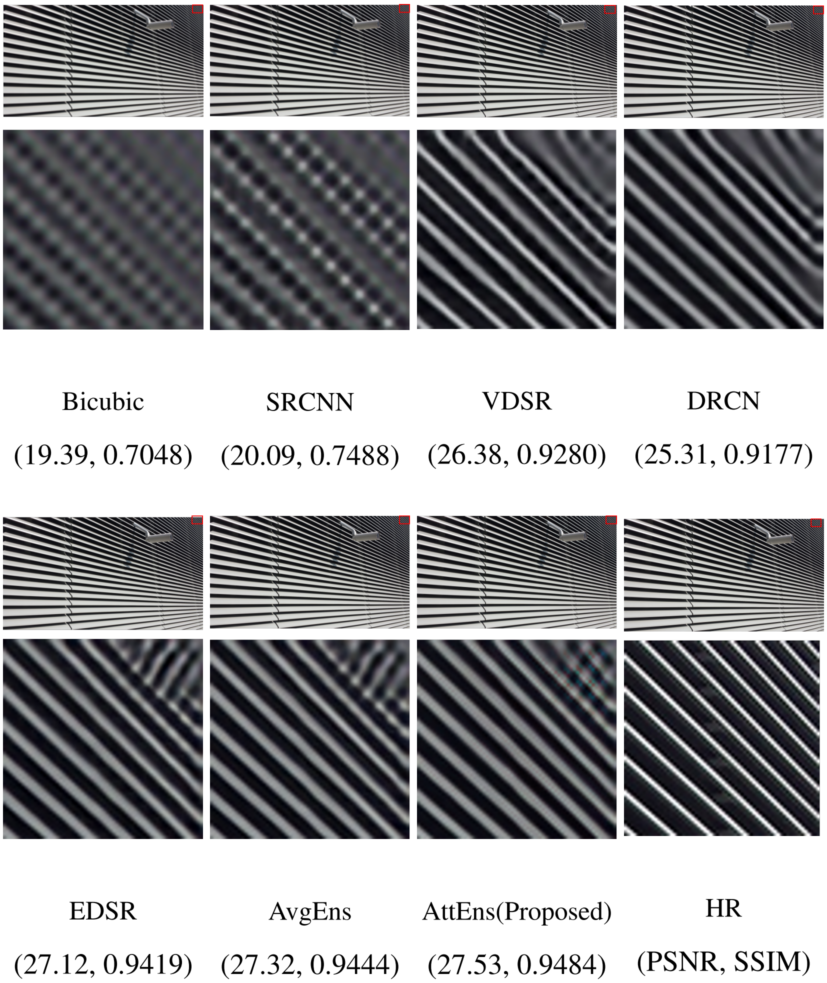
EDSR (23.81, 0.8420)

AvgEns (23.67, 0.8404)

AttEns(Proposed) (23.90, 0.8497)

HR (PSNR, SSIM)

Figure 4.1: Super-resolution results of ”img004” of Urban100 with scale factor x4.



Bicubic (19.39, 0.7048)

SRCNN (20.09, 0.7488)

VDSR (26.38, 0.9280)

DRCN (25.31, 0.9177)

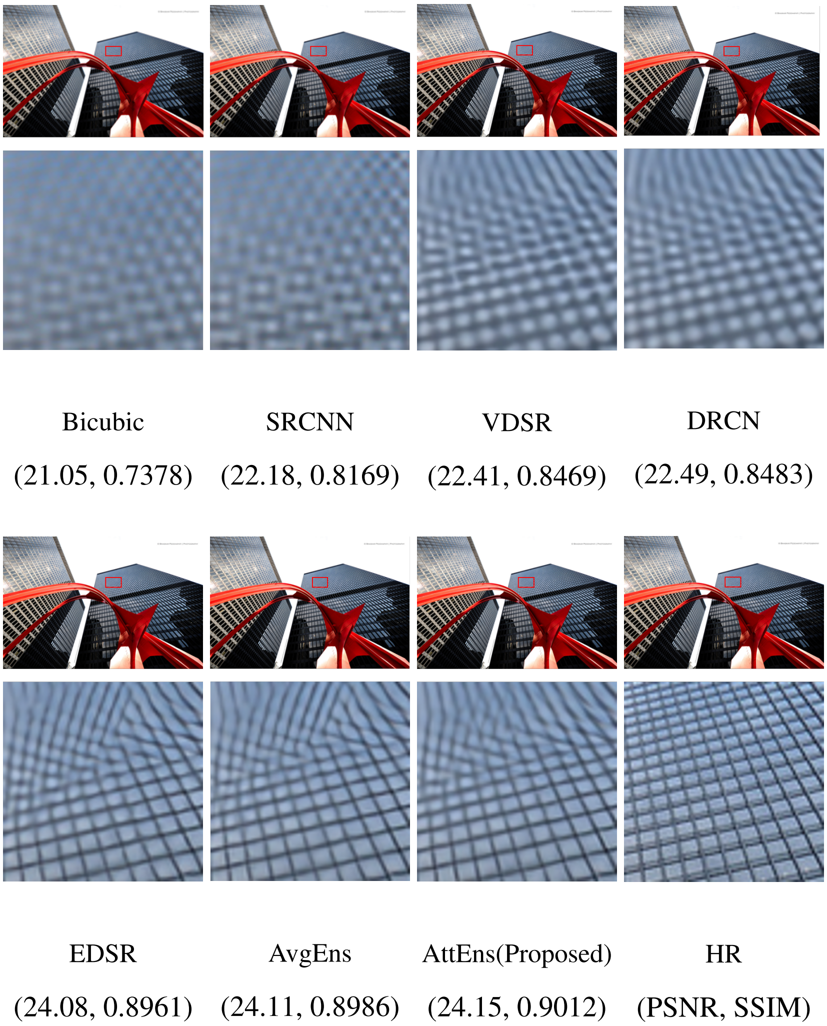
EDSR (27.12, 0.9419)

AvgEns (27.32, 0.9444)

AttEns(Proposed) (27.53, 0.9484)

HR (PSNR, SSIM)

Figure 4.2: Super-resolution results of ”img040” of Urban100 with scale factor x4.



Bicubic (21.05, 0.7378)

SRCNN (22.18, 0.8169)

VDSR (22.41, 0.8469)

DRCN (22.49, 0.8483)

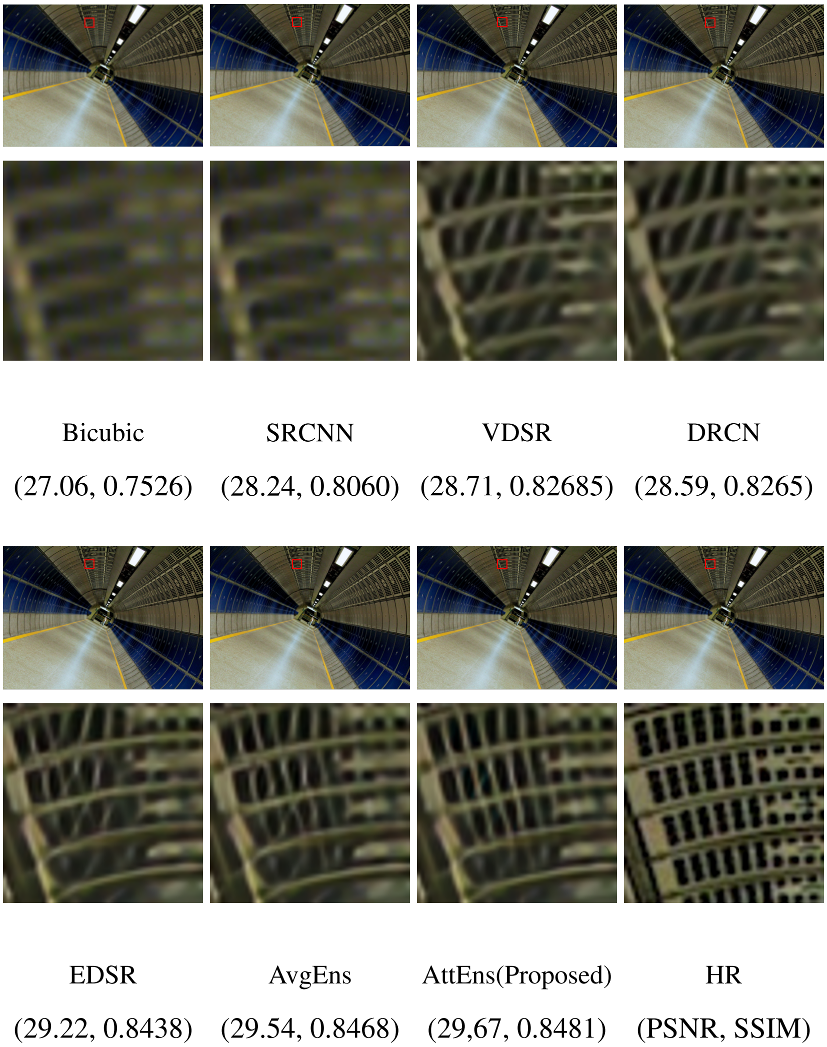
EDSR (24.08, 0.8961)

AvgEns (24.11, 0.8986)

AttEns(Proposed) (24.15, 0.9012)

HR (PSNR, SSIM)

Figure 4.3: Super-resolution results of ”img062” of Urban100 with scale factor x3.



Bicubic (27.06, 0.7526)

SRCNN (28.24, 0.8060)

VDSR (28.71, 0.82685)

DRCN (28.59, 0.8265)

EDSR (29.22, 0.8438)

AvgEns (29.54, 0.8468)

AttEns(Proposed) (29,67, 0.8481)

HR (PSNR, SSIM)

Figure 4.4: Super-resolution results of ”img078” of Urban100 with scale factor x3.

4.3.3 Effect of Attention Mask

In this section, we check whether the attention mask plays the role actually we want. The following figures 4.5 and 4.6 show each output, weight mask, and final output of the ensemble. The final output is represented by the equation $output = mask * outputA + (1 - mask) * outputB$. Therefore, the more white the attention mask is, the greater the weight of outputA. In the first figure, output A is closer to the ground truth so the color of the mask should be close to white. In the second figure, outputB is close to the ground truth so the color of the mask should be close to black. We can see that the color of the mask we expect is the same as the actual color.

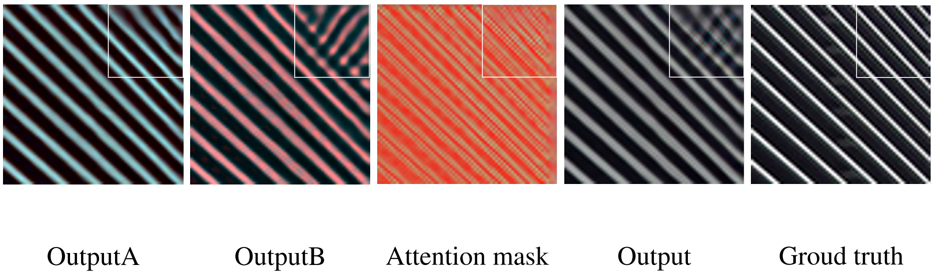

OutputA OutputB Attention mask Output Groud truth

Figure 4.5: Effect of attention mask in ”img040” (x4)

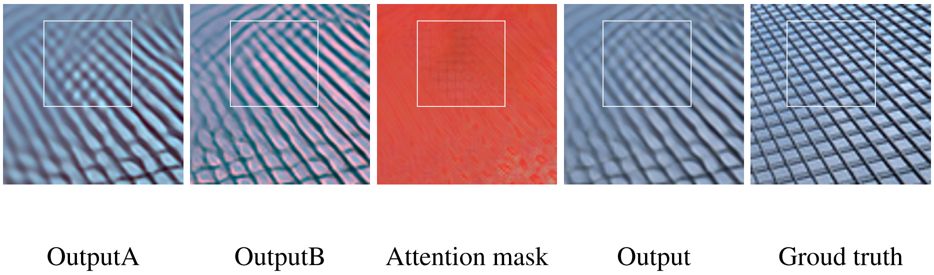

OutputA OutputB Attention mask Output Groud truth

Figure 4.6: Effect of attention mask in ”img062” (x4)

4.3.4 Hard Pixel Complement

In this section, we discuss how hard pixels are handled by the proposed method, and how the networks in the ensemble complement each other's hard pixels. To confirm and evaluate effect of the proposed method, the proposed method was compared with average ensemble which is the general ensemble.

The conventional methods for evaluating super-resolution are PSNR and SSIM. However, since these evaluation methods are applied to the entire image, a method for hard pixels is needed. For quantitative evaluation in hard pixels, we have to define the hard pixels numerically. The degree of difficulty is not easy to define. But we know that hard pixel means poorly performing pixels in the network output and performance is determined by the pixel difference. Therefore, based on the output of the baseline network and ground truth, we can define the hard pixel from the pixel difference. Also, we define $n - hardpixel$ is the pixel group of which pixel's difference is greater than a certain value n. The reason for defining $n - hardpixel$ is that the actual evaluation requires a pixel set to calculate the PSNR. The definition is as follow if the maximum of the pixel value is 255.

$$n - hardpixel \Leftrightarrow |I_{i,j,k} - O_{i,j,k}| > n$$

Where $I_{i,j,k}$ is the pixel value of the i color, j, k position of the ground truth and $O_{i,j,k}$ is the counterpart position value of prediction. The larger the value of n, the harder the $n - hardpixel$ becomes.

The proposed method aims to improve the performance in hard pixels by doing ensemble the baseline networks. To measure the performance improvement in hard pixels, we need to measure the PSNR only for the hard pixels of the baseline network. This is illustrated in Figure 4.7. The y-axis value is the difference between the baseline's hard pixel PSNR and the hard pixel PSNR of the methods. The x-axis value is the threshold n which is defined from $n - hardpixel$. The proposed method obtains a higher PSNR than average ensemble method's PSNR.

Next, in order to measure the complementary characteristics, it is necessary to confirm that each network of the ensemble performs well on each other's hard pixels. As described above, the PSNR of the other network is measured for $n - hardpixel$ of each network and is shown in Figure 4.8. Since two networks used in the ensemble, each method in the graph is represented by two lines. The hard pixel PSNRs of the proposed method is higher than average ensemble's. The increasing slope is also steep. So we conclude that the proposed method is robust to hard pixels than the existing ensemble method.

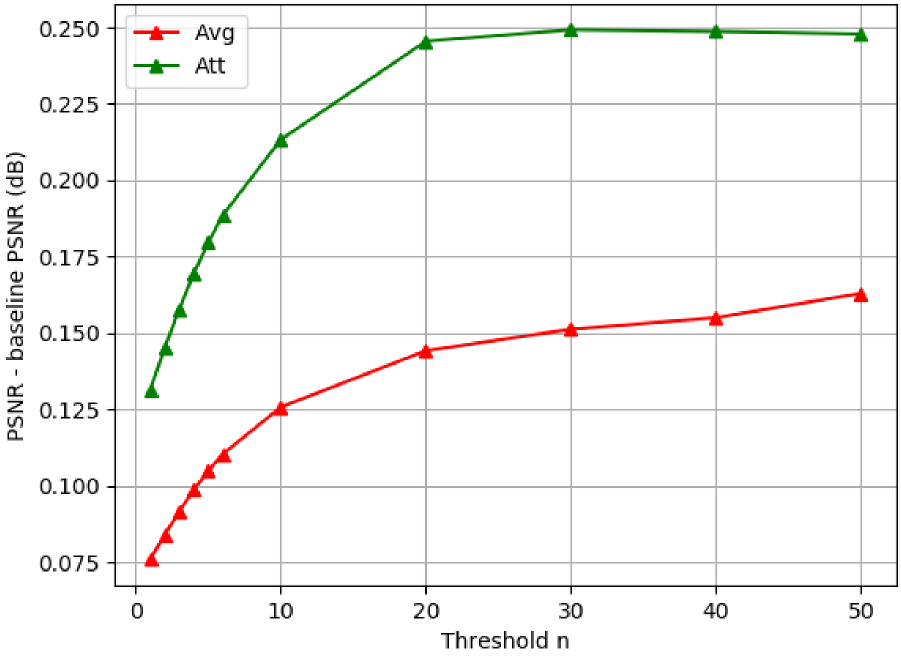


Figure 4.7: Hard pixel PSNR test on DIV2K validation (x3)

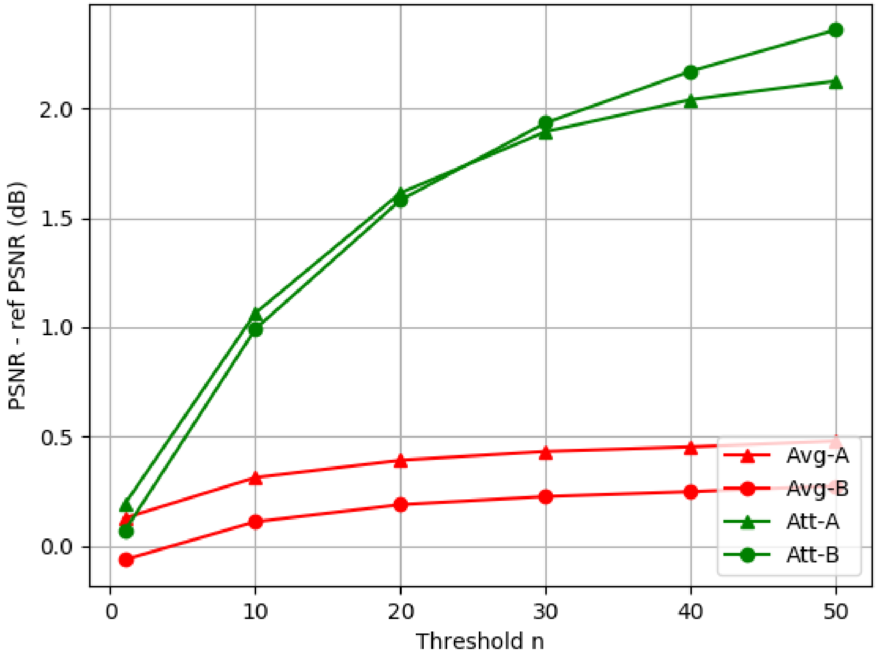


Figure 4.8: Complement PSNR test on DIV2K validation (x3)

Chapter 5

Conclusion

5.1 Summary

In this thesis, attention ensemble network was proposed to handle hard pixels of a network. Ensemble makes weak learner complement each other so it can alleviate hard pixels. Also, general ensembles can not handle hard pixels and easy pixels separately. To solve this issue, we adopted attention network for pixelwise ensemble.

The proposed method is more advantageous than the simple average ensemble method in several ways. First, in PSNR and SSIM, which are traditional evaluation methods, the proposed method shows the best performance. We defined the hard pixel and also showed that the proposed method get higher PSNR in hard pixels. In addition, it was shown that the ensemble network complements each other's hard pixels better than the simple average ensemble.

5.2 Future Directions

The attention network should predict which baseline network output is close to the ground truth. To do so, the attention network must have receptive field more than the baseline network's. However, even though the receptive field of the baseline network is very large, the receptive field of the attention network of the proposed structure is not large. We tried additional layer experiment and Resnet [31] structure experiment to increase the receptive field of the attention network, but failed to improve the performance. If we find a way to increase the receptive field of the attention network, it will definitely help the system improve performance. Also, there is tradeoff between easy pixels and hard pixels. Due to the tradeoff, the performance on easy pixels dropped slightly. If we solve this, we can get better results.

Bibliography

[1] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, et al. “Ntire 2017 challenge on single image super- resolution: Methods and results.” In *CVPRW* 2017.

[2] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi. “Low-complexity single-image super-resolution based on nonnegative neighbor embedding.” In *BMVC* 2012.

[3] R. Zeyde, M. Elad, and M. Protter. “On single image scale-up using sparse-representations.” In *Proceedings of the International Conference on Curves and Surfaces* 2010.

[4] D. Martin, C. Fowlkes, D. Tal, and J. Malik. “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecologi- cal statistics.” In *ICCV* 2001.

[5] J.-B. Huang, A. Singh, and N. Ahuja. “Single image super- resolution from transformed self-exemplars.” In *CVPR* 2015.

[6] C. E. Duchon. “Lanczos filtering in one and two dimensions.” *Journal of Applied Meteorology (JAM)*, 18(8):1016–1022, 1979.

[7] C. Dong, C. C. Loy, K. He, and X. Tang. “Learning a deep convolutional network for image super-resolution.” In *ECCV* 2014.

[8] J. Kim, J. Kwon Lee, and K. M. Lee. “Accurate image super- resolution using very deep convolutional networks.” In *CVPR* 2016.

[9] J. Kim, J. Kwon Lee, and K. M. Lee. “Deeply-recursive convolutional network for image super-resolution.” In *CVPR* 2016.

[10] . Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. “Photo-realistic single image super-resolution using a generative adversarial network.” arXiv preprint arXiv:1609.04802, 2016.

[11] B. Lim, S. Son, H. Kim, S. Nah, K.M. Lee. “Enhanced deep residual networks for single image super-resolution.” In *CVPRW* 2017.

[12] L. Wang, Z. Huang, Y. Gong, C. Pan, “Ensemble based deep networks for image super-resolution,”*Pattern Recognition* 68 (2017) 191–198.

[13] Y. Liu, Y. Wang, N. Li, X. Cheng, Y. Zhang, Y. Huang, G. Lu. “An attention-based approach for single image super resolution.” arXiv preprint arXiv:1807.06779, 2018.

[14] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks”, In *ECCV* 2018.

[15] WS. Lai et al. “Deep laplacian pyramid networks for fast and accurate super resolution.” In *CVPR* 2017.

[16] SJ. Park, H. Son, S. Cho, KS. Hong, S. Lee. “SRFeat: single image super-resolution with feature discrimination.” In *ECCV* 2018.

[17] J. Yang, J. Wright, T. S. Huang, Y. Ma. “Image super-resolution via sparse domain representation,” *Image Process* IEEE Trans. 19 (11) (2010) 2861-2873.

[18] D. Glasner, S. Bagon, M. Irani. “Super-resolution from a single image.” In *ICCV* 2009.

[19] L. Zhang, P. Wang, C. Shen, L. Liu, W. Wei, Y. Zhang, A. Hengel. “Adaptive importance learning for improving lightweight image super-resolution network.” arXiv preprint arXiv:1806.01576, 2018.

[20] J. T. Barron. “A more general robust loss function.” arXiv:1701.03077, 2017.

[21] X. Li, Z. Liu, P. Luo, C. C. Loy, X. Tang. “Not all pixels are equal: difficulty-aware semantic segmentation via deep layer cascade.” In *CVPR* 2017.

[22] S. Cho, J. Wang and S. Lee. “Handling outliers in none-blind image deconvolution.” In *ICCV* 2011.

[23] A. Shrivastava, A. Gupta, R. Girshick. “Training region-based object detectors with online hard example mining.” In *CVPR* 2016.

[24] O. Canevet, F. Fleuret. “Large scale hard sample mining with monte carlo tree search.” In *CVPR* 2016.

[25] S. Ren, K. He, R. Girshick, J. Sun. “Faster R-CNN: towrads real-time object detection with region proposal networks.” In *NIPS* 2015.

[26] J. Hou, S. Zhang, and L. Dai. “Gaussian prediction based attention for online end-to-end speech recognition.” *Proc. Interspeech* 2017, pp. 3692–3696, 2017.

[27] A. Vaswani, N. Shazeer, N. Parmar et al. “Attention is all you need.” arXiv:1706.03762, 2017.

[28] K. Li, Z. Wu, K. C. Peng, J. Ernst, Y. Fu. “Tell me where to look: Guided attention inference network.” In *CVPR* 2018.

[29] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang. “Residual attention network for image classification” In *CVPR* 2017.

[30] L. Yao, A. Torabi, K. Cho et al. “Describing videos by exploiting temporal structur.” In *ICCV* 2015.

[31] K. He, X. Zhang, S. Ren, J. Sun. “Deep residual learning for image recognition” In *CVPR* 2016.

초록

단일 영상 초해상도 문제는 매우 빠르게 발전하고 있지만, 단순히 성능을 높이는 데에 방점을 두고 있다. 그러나 이 문제에서는 각 픽셀 위치의 특성에 따라 에러의 정도가 다르다. 그러므로 에러가 높은 특정 픽셀들, 즉, 하드 픽셀에 집중하여 접근하는 것이 효과적일 것이다. 하드 픽셀을 다루기 위해서는 추가적인 시스템이 필요하다. 그러나 이 시스템에서도 마찬가지로 기존과 다른 특성의 하드 픽셀을 얻게 된다. 그러므로 서로의 하드 픽셀을 보완해 주는 구조가 이상적인 시스템이라고 할 수 있다. 앙상블은 쉽게 시스템의 성능을 높일 수 있는 방법이며, 머신러닝에서도 많이 쓰이고 있다. 앙상블은 각각의 시스템을 서로 보완하게 해준다. 따라서 위 문제를 앙상블 방법으로 접근하여 각 네트워크가 서로의 하드 픽셀을 보완하게 한다. 또한 일반적인 앙상블과는 달리, 픽셀 별로 난이도가 다르므로 픽셀별 가중치가 달라야 한다. 또한 어떤 픽셀에 집중해야 할지에 대한 정보가 전달되어야 한다. 어텐션 네트워크는 이 역할에 제격이므로, 앙상블에 어텐션 방법을 추가적으로 적용하였다. 우리는 제안하는 방법을 통해 기존 네트워크보다 높은 성능을 얻었다. 또한 하드 픽셀을 정의하고 이에 대해 정확도를 측정하였다. 그럼으로서 실제로 어텐션 네트워크를 적용했을 때 상호 보완적인지와, 전체 시스템이 하드 픽셀에 특히 강인함을 보였다.

주요어: 단일 영상 초해상도 복원, 하드 픽셀 핸들링, 딥러닝, 앙상블, 어텐션

학번: 2018-20864