A Multi-Level Supervision Model: A novel approach for Thermal Image Super Resolution

Priya Kansal
Couger Inc
Shibuya, Tokyo, Japan
priya@couger.co.jp

Sabari Nathan
Couger Inc
Shibuya, Tokyo, Japan
sabari@couger.co.jp

Abstract

This paper proposes a novel architecture for thermal image super-resolution. A very large dataset is provided by PBVS 2020 in their super-resolution challenge. This dataset contains the images with three different resolution scales(low, medium, high) [1]. This dataset is used to train the proposed architecture to generate the super-resolution images in x2, x3, x4 scales. The proposed architecture is based on the residual blocks as the base units of the network. Along with this, the coordinate convolution layer and the convolutional block attention Module (CBAM) are also used in the architecture. Further, the multi-level supervision is implemented to supervise the output image resolution similarity with the real image at each block during training. To test the robustness of the proposed model, we evaluated our model on the Thermal-6 dataset [13]. The results show that our model is efficient to achieve the state of art results on the PBVS'2020 dataset. Further the results on the Thermal-6 dataset show that the model has a decent generalization capacity.

1. Introduction

The image super-resolution involves the task of generating the high-resolution images from the low-resolution images [12]. Due to the development of automation in every field, the super-resolution of images is of vital importance. In recent years, imaging techniques have also been developed a lot. Nowadays, the practitioners are able to capture almost all the visible spectrum regions of an image including thermal images. Thermal images are infrared radiation emitted by all objects with different temperatures and temperatures above absolute zero.[13] [4]. Unlike the RGB images, the images captured in the thermal spectral band are not affected by the lighting and other environmental conditions, hence these images have wide applications such as medicine, military, object detection, recognition, and tracking[4]. However, capturing these thermal images with a high resolution is quite expensive because of the expensive equipments[13]. Hence, the requirement of high-resolution thermal images at an affordable cost is the need of time. Researchers are working on the thermal image super-resolution as an alternative to this problem. However, image super-resolution is always a challenging problem. Recently, the remarkable performance of neural network inspired the researchers in this field also. The approach proposed in this paper is also a deep convolutional neural network-based approach, named Multi-Level Supervision model which exploits the coordinate convolutional layer, residual connections, and attention modules. The proposed network is novel in the following way:

- Because of the multilevel supervision, this single model can handle the super-resolution task at three different scales (x2, x3, x4).
- This model is having a low complexity as it exploits the residual connection to focus on the lost information.
- This model is more robust as it is able to retain all the spatial information because of the use of coordinate convolutional layer and CBAM.

Detailed architecture is discussed in section 3.

2. Related work

Due to the wide applications, image super-resolution is widely studied from the last few decades. However, with the recent development in the deep convolutional neural networks and their impressive performance, researchers in this field have also get attracted to the use of convolutional neural networks for super image resolution task. For example, [2] constructs a three-layer deep convolutional neural network for image super-resolution in which the features of the LR input image are extracted and up-sampled in the last layer. The results of this model outperformed most

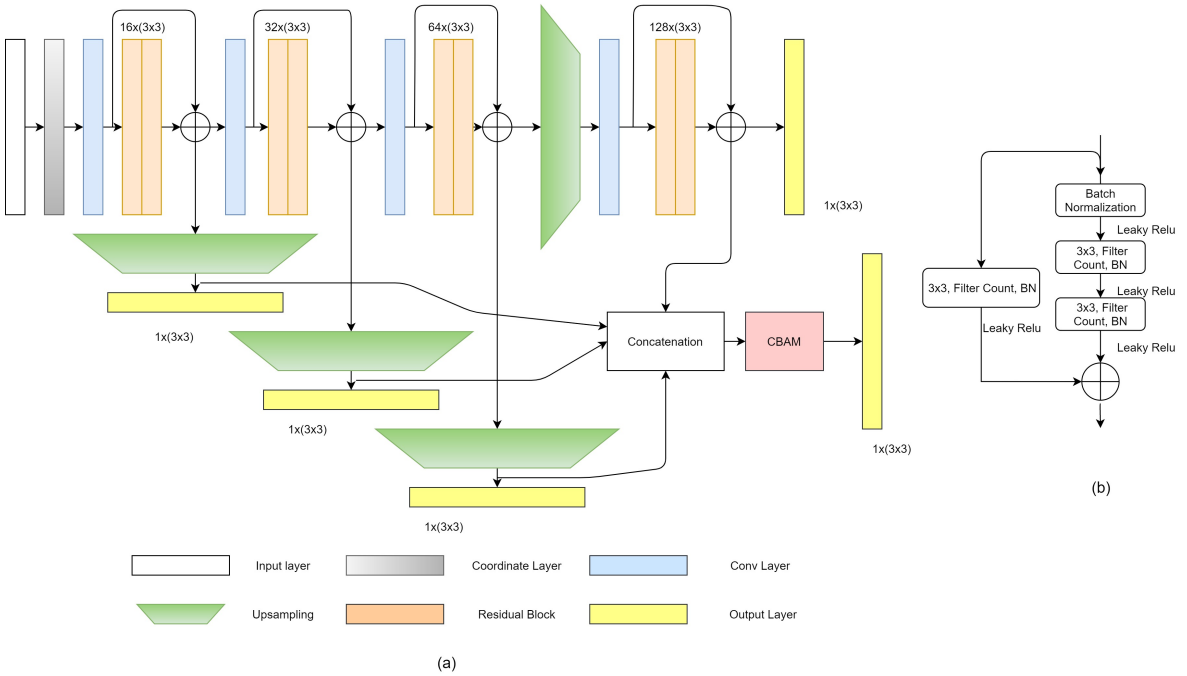

Figure 1: (a) Proposed Model: A Multi-Level Supervision Model; (b) A detailed modified residual unit

of the previous non-deep learning-based methods. Similarly, [7][20] develop a model based on the residual learning which is much deeper than the previous methods. There exist a lot of experimentation to improve the performance of the task of super-resolution such as [10] experimented to improve the speed during training by removing the batch normalization. [3][15][21] propose approaches to reduce the complexity and the cost of the image super-resolution task. Further, the HR images generated using generative adversarial networks have also given some impressive results[9]. However, all these approaches have discussed the super-resolution of the images in the RGB spectrum. There are only a few studies that develop the approaches to generate the high-resolution(HR) images from the low-resolution (LR) thermal images. Recently [14] propose a deep CNN network with residual blocks exploiting dense connection. Similarly, [12] develop a model based on the Cycle GAN architecture for rescaling the thermal images from LR to HR. The present work is also a deep convolutional network that is based on some attentions module and exploits the multi-level supervision to train the network.

3. Proposed Architecture

To deal with the challenge of thermal image super-resolution, we have proposed an attention-based multi-level supervision deep convolutional network. The detailed architecture is presented in Figure 1(a). In the following subsections of this section, we will discuss the details of each part we used in the proposed architecture.

3.1. Residual Units

Residual units have been a proven state-of-art approach for improving the accuracy in many domains[5][16]. In our residual block, the lost information of the previous layer is again infused to the network using the non-identity mapping. This restored information can be used to improve similarity with the high-resolution image. Further, the skip connection helped in dealing with the problem of vanishing or exploding gradient as it bypasses the higher layer gradients directly to the first convolution layer of the residual block. The initial batch-normalization assisted in the fast convergence of the loss. We have used Leaky Relu (0.1) as an activation function in the residual unit as inspired by [22][6]. Figure 1(b) shows the details of the Modified Residual unit.

3.2. Convolutional Block Attention Module (CBAM)

The CBAM block inspired by [18][6] is used to create spatial attention on the channel attention of the fused output of all the side layers. First of all, to create channel attention, the spatial dimension is squeezed using max pooling and average pooling simultaneously. Exploiting both the features improves the representation power of the network. Next, the concatenated pooling features are passed to the convolutional layer and activated between 0 and 1. The output is further passed to the spatial attention module. The spatial

module includes the 1x1 convolution operation of the max-pooling of all neurons. This attention is further added to the original input.

3.3. Multi-level Supervision

Inspired by [6][11], to improve the image resolution at different scales using one model, we used multi-level supervision. This helped the model to learn the features according to the original high-resolution image. As the receptive field increases across successive layers, predictions computed at different layers embed spatial information at different levels. The network is able to update the weights more efficiently, and propagate the gradient in intermediary level to learn the features at each intermediary scale. The output of the first three residual blocks is directly up-sampled and supervised. However, the up-sampled output of the third residual block is further passed to another residual block before supervision. The resulted information is then concatenated along with the three up-sampled layers again and the attention discussed in the previous section is applied. This multilevel supervision guided the network to generate the HR images progressively from the different resolutions respectively. This enabled a single architecture to work on all the three kind of resolution tasks. The detailed ablation study shows that this multilevel supervision is improving the results significantly.

4. Experimental Setup

4.1. Dataset

PBVS'2020 Dataset For generating super-resolution images, PBVS'2020 provides thermal image captured using three different thermal cameras with three different resolutions (low *160X120*, mid *320X240* and high *640X480*). A total of 951 images for training and 50 images for testing with each resolution are shared in the development phase whereas 20 images with each resolution are shared for validation phase[12]. A sample image from each resolution is shown in Figure2.

Thermal6 Dataset For testing the robustness of the proposed architecture and the model trained, we tested the results on the Thermal6 dataset also. Thermal6 dataset is acquired using a Tau2 camera with a resolution of 640 x 512. A total of 101 images are there in the dataset which includes the indoor and outdoor environment in day and night both [13].

4.2. Training

The network is trained for five outputs which include the four side layers and one fused output layer with low-resolution input images. All the outputs are supervised using the loss proposed in Eq.4. Total three models are trained to get the high-resolution images which are double (x2), triple (x3) and four times(x4) in scale when compared to input images. Adam optimizer is used to update the weights while training. The learning rate is initialized with 0.001 and reduced after 15 epochs to 10 percent if validation loss does not improve. The batch size is set to 4. The total epochs are set to 500. However, training is stopped early when the network started overfitting. The dataset is trained using Nvidia 1080 GTX GPU. The model is evaluated using Peak-Signal-to Noise Ratio (PSNR) and Structural Similarity Index(SSIM) loss.

4.3. Loss Function

To supervise the model outputs, a combination of three different loss functions are used: mean squared error (MSE), SSIM Loss and Sobel edge loss (SOBEL Loss). MSE is used for maintaining the consistency between input and output; it is defined as:

$$MSE = \frac{1}{N}\sum_{p=1}^{p}(f(x) - x)^2 \tag{1}$$

where *f(x)* is the pixel value of generated HR image and *x* is the pixel value of HR real image.

Pixel wise Structural Similarity Loss is defined as:

$$\text{SSIM Loss} = \frac{1}{N}\sum_{p=1}^{p}(1 - SSIM(p)) \tag{2}$$

where *SSIM(p)* structural similarity index[17] for pixel p.

Sobel loss is the mean squared error of the Sobel edge information of the real image and the generated image. A Sobel filter to detect the edges is applied to the generated and real image and then this information is used to calculate the mean squared error which is equal to the Sobel loss. More information can be found in [19][8]. A mathematical representation is given in equation 3:

$$\text{SOBEL Loss} = \frac{1}{N}\sum_{i=1}^{N}(S(f(x)) - S(x))^2 \tag{3}$$

where *S(f(x))* is the sobel edge information of the generated HR image and *S(x)* is the sobel edge information of the real image.

Total loss is the sum of the all three losses.

$$\text{Total Loss} = \text{MSE} + \text{SSIM Loss} + \text{SOBEL Loss} \tag{4}$$

5. Experimental Results

5.1. Results

Table 1 shows the results of our proposed model on the test images of the PBVS'2020 dataset and Thermal6 dataset.

Scale	PBVS'2020 (Val)		PBVS'2020 (Test)		Thermal6	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
x2 Scale	31.88	0.9364	25.45	0.8529	40.89	0.9616
x3 Scale	30.75	0.9260	25.96	0.8271	38.95	0.9502
x4 Scale	31.92	0.9320	27.31	0.8498	37.60	0.9427

Table 1: Results of the Proposed Model on PBVS'2020 and Thermal 6

Coord Layer	CBAM	Multilevel Supervision	x2 scale		x3 scale		x4 scale	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
yes	yes	yes	**31.40**	**0.9326**	**30.23**	**0.9199**	**31.44**	**0.9267**
yes	yes	no	30.04	0.8332	29.54	0.8457	30.31	0.8757
yes	no	yes	30.24	0.8387	28.53	0.8984	28.68	0.8854
yes	no	no	29.18	0.8244	27.27	0.8026	29.55	0.8315
no	yes	yes	30.76	0.8578	29.51	0.8490	29.36	0.8207
no	yes	no	29.98	0.8312	28.58	0.8591	27.43	0.8549
no	no	yes	31.18	0.8595	27.40	0.8177	27.89	0.8535
no	no	no	30.87	0.8547	28.47	0.8336	27.43	0.8296

Table 2: PSNR/SSIM of Different Experimentation in terms of COORD Layer, CBAM and Multi level supervision on PBVS'2020 Validation set

5.2. Ablation Study

To prove the efficiency of our proposed architecture, a wide range of ablation studies have been performed. Table 2 shows the quantitative results calculated on the experimentation of using coordinate convolution (coord) layer, multilevel supervision (MS) and attention module while training. The results clearly show that using the proposed architecture with coord layer, MS and attention module is giving quite impressive results when compared to the scenarios where we are not using anyone of them.

Table 3 shows the results on the Thermal 6 dataset. The model which was trained on the PBVS'2020 dataset is used for calculating the evaluation metrics on the Thermal6 dataset. The results are at par with the previous approaches, which shows that the model is having a good generalization capacity.

Scale	Bicubic Model	TISR[14]	Ours
x2 scale	39.59	41.24	40.89
x3 scale	37.68	39.62	38.95
x4 scale	34.98	37.85	37.60

Table 3: Results on Thermal 6 dataset compared to the state of art

Table 4 shows the results of the model trained for the x2 scale for low-resolution images. The results are compared with the bicubic interpolation as a baseline and other published work. Since the results are available only in the x2 scale, we have presented only x2 scale results in the comparison table. Figure 2 and Figure 3 depicts the real image and generated images of PBVS'2020 dataset and Thermal6 dataset.

Method	PSNR	SSIM
Bicubic	16.46	0.6695
TISR[14]	17.01	0.6704
PBVS[12]	21.50	0.7218
Ours (val)	**31.92**	**0.9320**
Ours (test)	25.45	0.8529

Table 4: Results on LR set in x2 scale factor, compared with its HR registered test set of PBVS'2020.

6. Conclusion

This present paper proposes an attention-based multilevel supervised network to create high-resolution images. The network exploits the residual learning and coordinate convolutional layer to retain the spatial information throughout training, which helps to improve the robustness of this model. The multilevel supervision enabled the model to learn the resolution hierarchy throughout the network. Three models are trained using the same network for three different resolution scales. The quantitative results and in-depth ablation study results show that the proposed network is not only efficient enough to achieve the results on the PBVS'2020 dataset but also able to generalize the performance on other datasets. In the future, we will exploit the same architecture for images in the RGB spectrum and for other image restoration tasks in RGB as well as a thermal spectrum.

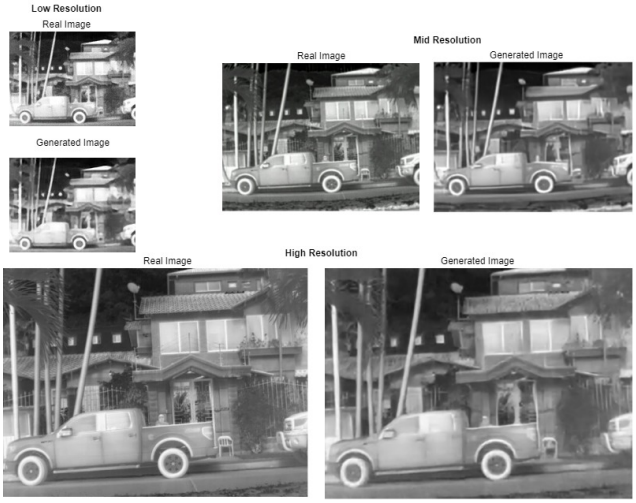

Figure 2: PBVS'2020 dataset: Generated output and the Real Output Images in three different scale

Figure 3: Thermal dataset: Real Image and Generated output image in three different scale

References

[1] IEEE PBVS'2020. ://vcipl-okstate.org/pbvs/20/challenge.html. 1

[2] 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*, 38(2):295–307, 2015. 1

[3] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016. 2

[4] Rikke Gade and Thomas B Moeslund. Thermal cameras and applications: a survey. *Machine vision and applications*, 25(1):245–262, 2014. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on CVPR*, Jun 2016. 2

[6] Priya Kansal and Sabari Nathan. Eyenet: Attention based convolutional encoder-decoder network for eye region segmentation, 2019. 2, 3

[7] 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*, pages 1646–1654, 2016. 2

[8] Josef Kittler. On the accuracy of the sobel edge detector. *Image and Vision Computing*, 1(1):37–42, 1983. 3

[9] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[10] 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*, pages 136–144, 2017. 2

[11] Sabari Nathan and Priya Kansal. Skeletonnet: Shape pixel to skeleton pixel. In *Proceedings of the IEEE Conference on CVPRw*, 2019. 3

[12] Rafael E Rivadeneira, Angel D Sappa, and Boris X Vintimilla. Thermal image super-resolution: a novel architecture and dataset. 1, 2, 3, 4

[13] Rafael E Rivadeneira, Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Thermal image superresolution through deep convolutional neural network. In *International Conference on Image Analysis and Recognition*, pages 417–426. Springer, 2019. 1, 3

[14] Rafael E Rivadeneira, Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Thermal image superresolution through deep convolutional neural network. In *International Conference on Image Analysis and Recognition*, pages 417–426. Springer, 2019. 2, 4

[15] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 2

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[17] 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. 3

[18] 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)*, pages 3–19, 2018. 2

[19] Shanxin Yuan, Radu Timofte, Gregory Slabaugh, Ales Leonardis, Bolun Zheng, Xin Ye, Xiang Tian, Yaowu Chen, Xi Cheng, Zhenyong Fu, et al. Aim 2019 challenge on image demoireing: Methods and results. *arXiv preprint arXiv:1911.03461*, 2019. 3

[20] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017. 2

[21] Lei Zhang, Peng Wang, Chunhua Shen, Lingqiao Liu, Wei Wei, Yanning Zhang, and Anton Van Den Hengel. Adaptive importance learning for improving lightweight image super-resolution network. *International Journal of Computer Vision*, 128(2):479–499, 2020. 2

[22] Xiaohu Zhang, Yuexian Zou, and Wei Shi. Dilated convolution neural network with leakyrelu for environmental sound classification. In *2017 22nd International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2017. 2