Hierarchical Attention Maps for Super Resolution

Zhiqin Zhang¹, Bo Zhang²,*
School of Computer Science, Wuhan Donghu University, Wuhan 430212, China
*Corresponding author’s e-mail: zhangbo@wdu.edu.cn

Abstract: Recent state-of-the-art super-resolution methods have achieved impressive performance on ideal datasets. However, these methods always fail in balance between spatial details and high-level perceptual information. Most of them adopt downsampling step to construct Low-Resolution (LR) and High-Resolution (HR) training which may lose local spatial details. To address this issue, we focus on designing a hierarchical attention maps mechanism for recovering both local spatial details and global perceptual information. By using our novel Hierarchical attention module, we can acquire better High-Resolution (HR) predicted images. Finally, we propose a hierarchical multiple scale feature concatenation module aiming at better perception. Extensive experiments on real-world images demonstrate that our method achieved better visual quality both for perception and quantitative estimation.

1. Introduction
Over the last few years, image restoration tasks such as denoising, deblurring or super-resolution (SR) [1, 2] have experienced significant improvements because of deep learning. super-resolution (SR) task is to increase the resolution of low-quality images, and keep its high fidelity and clarity. The problem is ill-posed since it is difficult to restore any given low-resolution pixel from a multiplicity of solutions. In recent years, deep learning-based methods [3] have achieved remarkable results with respect to image restoration on improving image fidelity, which mainly focuses on developing different network structures to achieve better performance of specific datasets. Most of them use fixed kernel (such as bicubic) layer for downsampling to construct training data pairs. Similarly, in test phase, the input image downsampled by the same kernel is feed to the designed network. Subsequently, The performance will be estimated by comparing the predicted results with Ground Truth (GT) to calculate SSIM and PSNR metrics.

The successes encourage the community to further attempt deep learning on the more challenging video restoration problems. Earlier studies [4,5] treat video restoration as a simple extension of image restoration. The temporal redundancy among neighboring frames is not fully exploited. Recent studies [6] address the aforementioned problem with more elaborated pipelines that typically consist of four components, namely feature extraction, alignment, fusion, and reconstruction. Although video super resolution method can achieve better perceptual quality than single image super resolution, more computation resource will be cost.

In this paper, we focus on a single image super resolution, we propose a novel hierarchical attention maps (HAM) and feature concatenation strategy for Super-Resolution, which contains attentive intra-scale features fusing and inter-scale features concatenation. On one hand, we can reserve key features and abandon useless features from low hierarchical by HAM. On the other hand, by using inter-scale features concatenation, different level information flow can make our model learn enough which is good for model convergence and generalization. The experimental results show that
our method produces clearer and cleaner results compared with several current deep learning methods. The PSNR of the proposed method is 0.12 higher than the best methods enumerated in this paper.

2. Related Work

Recent years have witnessed a paradigm shift from high-end DSLR cameras to smartphone cameras. However, capturing high-quality images with smartphone cameras is challenging. Image degradations are often present in images either due to the limitations of cameras and/or adverse environment situation.

Many studies have introduced prior information to help address the ill-posed SR problem. Early methods explore a smoothing prior such as bicubic interpolation and Lanczos resampling. Image priors such as edge features, statistics and internal patch recurrence are employed to improve performance. Dong et al. [7] train domain specific dictionaries to better recover local structures in a sparse representation framework.

For video super-resolution, temporal alignment plays an important role and has been extensively studied. Several methods [8-10] use optical flow to estimate the motions between images and perform warping. However, accurate flow is difficult to obtain given occlusion and large motions and the complexity of optical flow model hinder the real application of SR.

In contrast to these studies, we explore how to keep both local spatial details and global perceptual information. To achieve this goal, we proposed a method which include two key steps:

1) Intra scale hierarchical attention maps (HAM): for reserving local spatial details

2) Inter scale features concatenation: for better information flow which can improve perceptual visual effect.

3. Hierarchical Attention and Feature Concatenation

The whole model structure was consisted of two U-Net structure which was showed in figure 1, we adopted two cascade UNet subnet, the bottom unet was used to model traditional SR procedure, the top unet is for refining the bottom restoration results, HAM module was used to refine key information from bottom unet, then the refined information was propagated to top unet. We used HAM reserve key local spatial details and features concatenation get global image perceptual effect. The details was showed as follows:

![Figure1. The whole structure of The propose method](image1)

Intra scale hierarchical attention maps (HAM):

The illustration was showed in figure 2, where yellow conv block represents convolution, green GP block represents global pooling which transform 2d conv output to 1d vector for attention.

![Figure2.illustration of hierarchical attention maps](image2)
Inter scale features concatenation:
The whole net structure is consisted of two unet, every unet have several corresponding downsampling and upsampling layers, the same scale features will be concatenated to obtain multiple scale features which is good for perception field. The fusing details were showed in figure 1.

4. Experiments
We focus on real-world super-resolution, so we evaluate our method on the DPED dataset, in which the photos suffer from degradation problems such as blur, noise, etc. There is no ground truth in the test set, so we only show the results of visual comparison. The visual compared results were showed in figure 3, where column one represents ground truth original high-resolution images, column two are our method results and the last column are bicubic results. Our method can get more clear results.

![Figure3](image.png)

Figure3. The visual comparison results of different SR methods

We also compared with bicubic on quantitative estimation metrics, PSNR and SSIM. The results was showed in table1, our method get 5% relative improvement on PSNR.

| Method       | PSNR  | SSIM   |
|--------------|-------|--------|
| Bicubic      | 17.21 | 0.2051 |
| ESRGAN       | 19.06 | 0.2423 |
| EDSR         | 25.31 | 0.6383 |
| Ours         | 26.08 | 0.6219 |

5. Conclusion
We have explored the joint use of Intra-scale hierarchical attention maps and inter-scale features concatenation for recovering the clear details and retaining the global perception features in SR. A hierarchical attention maps (HAM) block has been proposed to efficiently incorporate the bottom unet features into the top unet subnet. Thanks to inter-scale features concatenation in unet, our method is capable of generating distinct and rich textures in a super-resolved image. Extensive experiments demonstrate the capability of method in achieving visually pleasing textures while retaining global semantic perception information, outperforming previous methods.

References
[1] Saeed Anwar and Nick Barnes. (2019) Real image denoising with feature attention. In: International Conference on Computer Vision. Seoul, Korea (South). pp.3155-3164.
[2] Saeed Anwar and Nick Barnes. (2020) Densely residual laplacian super-resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.13140/RG.2.2.15072.33285
[3] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. (2019) Unprocessing images for learned raw denoising. In Computer Vision and Pattern Recognition. Long Beach, CA, USA, pp.11028-11037.
[4] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. (2018) Learning to see in the dark. In Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, pp.3291-3300.
[5] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. (2017) Dslr-quality photos on mobile devices with deep convolutional networks. In: International Conference on Computer Vision. Venice, Italy, pp. 3277–3285.
[6] Saeed Anwar, Salman Khan, and Nick Barnes. (2019) A deep journey into super-resolution: A survey. ACM Computing Surveys. https://doi.org/10.1145/3390462
[7] Liyuan Pan, Yuchao Dai, Miaomiao Liu, and Fatih Porikli. (2017) Simultaneous stereo video deblurring and scene flow estimation. In: Computer Vision and Pattern Recognition. Honolulu, HI, USA, pp.6986-6996.
[8] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. (2019) Recurrent back-projection network for video super-resolution. In: Computer Vision and Pattern Recognition. Long Beach, CA, USA, pp.3892-3901.
[9] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In CVPR, 2018
[10] Jose Caballero, Christian Ledig, Aitken Andrew, Acosta Alejandro, Johannes Tottz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In CVPR, 2017