Image Super-Resolution via Hierarchical Attention-Based Multi-References Sampling

Marco Pesavento, Marco Volino, Adrian Hilton
{m.pesavento,m.volino,a.hilton}@surrey.ac.uk

Centre of Vision, Speech and Signal Processing (CVSSP),
University of Surrey

1 Introduction

To overcome the limitations of single image super-resolution (SR) approaches that produce blurry super-resolved images, recent research has introduced the sub-problem of reference image super-resolution (RefSR). Given a low resolution (LR) input image and a similar high resolution (HR) reference image, RefSR approaches estimate a SR image. Reference super-resolution with a single reference image has been demonstrated to improve performances over general SR methods achieving large up-scaling with reduced visual artefacts. We generalise reference super-resolution to use multiple reference images giving a pool of image features and propose a novel attention-based sampling approach to learn the perceptual similarity between reference features and the LR input. As shown in Figure 2, given N_M reference images, our approach produces a 4× SR image which is perceptually plausible and has a similar level of detail to the ground-truth HR image. An extended version of this short paper will appear in ICCV [2].

2 Method

The problem of multiple-reference super-resolution can be stated as follows: given a LR input I_{LR} and a set of HR reference images $\{I^m_{ref}\}^{N_M}_{m=1}$, estimate a spatially coherent SR output I_{SR} with the structure of I_{LR} and the appearance detail resolution of the multiple-reference images. Figure 1 presents an overview of the proposed approach to achieve multiple-reference super-resolution, which comprises the following stages.

Feature Extraction: to reduce GPU memory consumption with multiple reference images, the LR input I_{LR} and HR reference images $\{I^m_{ref}\}^{N_M}_{m=1}$ are divided into N_I and N_R sub-parts, respectively. Image features are extracted from these parts using a pre-trained VGG-19 network. The input vector is further divided into subvectors to focus the learning attention on input features while computing similarity maps with reference features.

Hierarchical Attention-based Similarity: the objective of this stage is to map the features of the LR input to the most similar features of the HR reference images. The output is a feature vector that contains the values of these most similar reference features. A hierarchical approach of similarity mapping is performed over $l = N_L$ levels. For every level l of the hierarchy, a similarity map between LR input subvectors and reference features is computed:

$$s^l_k = \phi^c(I_{LR}) * \frac{P_k(O^{l-1,r,m}_{ref})}{||P_k(O^{l-1,r,m}_{ref})||} \tag{1}$$

$k = c$ if $l = 1$, $k = r$ or $k = m$ otherwise. P is the patch derived from the application of the patch-match approach: patches of the reference features $O^{l-1,r,m}_{ref}$ are convoluted with subvectors $\phi^c(I_{LR})$ of the LR input to compute the similarity. When the similarity map s^l_k is evaluated, a vector O^l_{ref} containing the most similar features of O^{l-1}_{ref} is created by applying either one of two distinct approaches:

1. **Input attention mapping** ($l = 1$): in the first level a feature vector is created by maximising over every subvector of the input:

$$O^{1,r,m}_{ref}(x,y) = P_{k^*}(\phi^r(I^m_{ref}))(x,y) \tag{2}$$

$$k^* = \operatorname*{argmax}_{k=c} s^1_k(x,y)$$

$O^{1,r,m}_{ref}(x,y)$ is the (x,y) value of the k^* patch $P(\phi^r(I^m_{ref}))$ whose s^1 is the highest among all the similarity values $s^1_k(x,y)$ for each subvector of the LR input feature vector.

2. **Reference attention mapping** ($l > 1$): for subsequent levels of the hierarchy, a feature vector is created by maximising a new similarity s^l_k map over the feature vector created in the previous level.

$$O^{l,k}_{ref}(x,y) = O^{l-1,k^*}_{ref}(x,y) \tag{3}$$

$$k^* = \operatorname*{argmax}_{k} s^l_k(x,y)$$

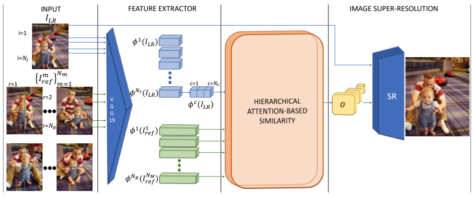


Figure 1: Overview of the approach. Given a LR input image and N_M reference images, the approach produces an HR reconstruction of the LR input image exploiting the references.

$k = r$ or $k = m$ depending on which level is processed. The value of $O^{l,k}_{ref}$ in the (x,y) position is the value of $O^{l-1,k}_{ref}$ with the highest s^l among all the $s^l_k(x,y)$ of $O^{l-1,k}_{ref}$.

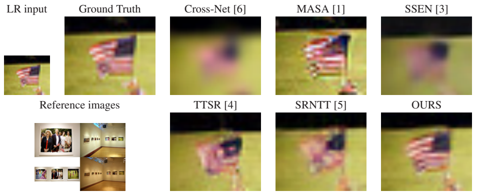
The final output, obtained when the similarity mapping is performed for all the levels of the hierarchy, is a feature vector which contains the features of the references that are most similar to the features of the LR input.

Image Super-resolution: given the feature similarity mapping O, a generative adversarial network super-resolves the LR input to obtain the SR output which maintains the spatial coherence of the input with the HR appearance detail of the reference images. We modified the architecture of the generator of [5] by eliminating the batch normalization layers since they reduce the accuracy for dense pixel value predictions.

3 Results

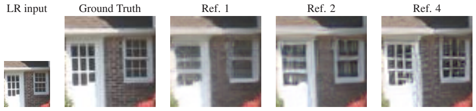
We evaluate our method by comparing with state-of-the-art single RefSR approaches. Figure 2 shows the superiority of our approach.

We also confirmed (Figure 3) that increasing the number of reference images will lead to an improvement of the performance of the approach.



PSNR/ SSIM	Cross-net [6]	MASA [1]	SSEN [3]	TTSR [4]	SRNTT [5]	OURS
	26.00/.7576	24.84/.7311	22.71/.7169	25.59/.7645	26.42/.7738	**27.49/.8145**

Figure 2: Qualitative (top) and quantitative (bottom) comparisons with RefSR approaches.



PSNR/ SSIM	Ref. 1	Ref. 2	Ref. 4
	26.77/.7882	27.30/.8087	**27.49/.8145**

Figure 3: Qualitative (top) and quantitative (bottom) results of using different numbers of reference images.

[1] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. *arXiv preprint arXiv:2106.02299*, 2021.

[2] Marco Pesavento, Marco Volino, and Adrian Hilton. Attention-based multi-reference learning for image super-resolution. *arXiv preprint arXiv:2108.13697*, 2021.

[3] Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8434, 2020.

[4] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020.

[5] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7982–7991, 2019.

[6] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Cross-net: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 88–104, 2018.