Interpreting Super-Resolution Networks with Local Attribution Maps Supplementary Material

Jinjin Gu[1] Chao Dong[2,3]
[1]School of Electrical and Information Engineering, The University of Sydney.
[2]Key Laboratory of Human-Machine Intelligence-Synergy Systems,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences.
[3]SIAT Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society
jinjin.gu@sydney.edu.au, chao.dong@siat.ac.cn

Abstract

In this supplementary material, we first provide a review for the attribution methods in the literature and discuss their relationship to the interpretation of SR networks in Sec 1. We also provide a review for the SR networks discussed in the main text in Sec 2. The detailed training settings are provided. At last, we provide more qualitative results in Sec 3.

1. Review of Attribution Methods

In this section, we provide a review of attribution methods in the literature that are used for interpreting classification networks. We also discuss their relationship with the interpretation of super-resolution (SR) networks. As presented in the main text, given an input image $I \in \mathbb{R}^d$ and a model $S : \mathbb{R}^d \mapsto \mathbb{R}$ that outputs the probability of I belongs to a certain class, an attribution method provides attribution maps $\mathsf{Attr}_S : \mathbb{R}^d \mapsto \mathbb{R}^d$ for S that are of the same size as the inputs. Each dimension of these attribution maps corresponds to the "relevance" or "importance" of that dimension to the final output, which is often a class-specific score in classification networks.

Gradient w.r.t. I. This method employs the gradient of the predicted probability w.r.t. to the input I [29, 7].

$$\mathsf{Grad}_S(I) = \frac{\partial S(I)}{\partial I} \tag{1}$$

However, the vanilla gradient method suffers from the "saturation" problem that the magnitude of this gradient tends to be small. A little movement toward the direction of the gradient will not change the predicted probability significantly [32]. In Sec 3.4 of the main text, we show that for the interpretation of SR networks, the "saturation" problem also exist. Thus the vanilla gradient method is not appropriate for interpreting SR networks.

The element-wise product of Gradient and the input. This method was proposed to address the saturation problem and reduce visual diffusion [28], denoted as

$$\mathsf{Grad} \odot \mathsf{I}_S(I) = I \odot \frac{\partial S(I)}{\partial I}. \tag{2}$$

Ancona *et al*. [4] show that, for a network with only ReLU activation function and no additive biases, this input gradient product is equivalent to DeepLift [28], and ϵ-LRP [6]. For the interpretation of SR networks, the pixel intensity should not be part of the attribution as the textures and edges may not change when the pixel intensity changes. Directly calculate the product of the input intensity and the gradient will introduce interference factors.

Guided Backpropagation (GBP). This method specifies a change in how to calculate gradients for ReLU activations. Let $\{f^l, f^{l-1}, \dots, f^0\}$ be the feature maps obtained during the forward process by a deep neural network S, and $\{r^l, r^{l-1}, \dots, r^0\}$ be the representation obtained during the backward process. Springenberg *et al*. [31] propose GBP that aims to zero out negative gradients during the computation of r. The map is computed as:

$$r^l = 1_{r^{l+1}>0} 1_{f^l>0} r^{l+1}, \tag{3}$$

where $1_{r^{l+1}>0}$ represents keeping only the positive gradients and $1_{f^l>0}$ indicates keeping only the positive activations. The usage scenarios of this method are relatively limited. For residual networks that are widely used in SR, this method is not valid.

Integrated Gradients (IG). Most relevant to the method proposed in this paper, IG also employs path integration [12], but uses a black image as baseline image and linear interpolation as the path function. IG is defined as:

$$\mathsf{IG}_S(I) = (I - I') \times \int_0^1 \frac{\partial S(I' + \alpha(I - I'))}{\partial I} d\alpha, \tag{4}$$

where I' is the baseline black image and α is the parameter of the interpolation. In Sec 3.4 of the main text, we discuss the differences between the proposed local attribution maps for SR networks and IG.

SmoothGrad and VarGrad. SmoothGrad [30] and VarGrad [1] are proposed to relieve the situation where the attribution graph is full of noise. The SmoothGrad is defined as:

$$\mathsf{SmoothGrad}_S(I) = \frac{1}{N}\sum_{i=1}^{N}\mathsf{Grad}_S(I+n_i), \qquad (5)$$

where n_i are the noise vectors and $n_i \sim \mathcal{N}(0, \sigma^2)$ are sampled from a Gaussian distribution. Similar to SmoothGrad, a variance analog of SmoothGrad can be defined as:

$$\mathsf{VarGrad}_S(I) = \mathcal{V}(\mathsf{Grad}_S(I+n_i)), \qquad (6)$$

where $\mathcal{V}$ represents to the variance. Seo *et al.* [26] theoretically analyze VarGrad showing that it is independent of the gradient, and captures higher order partial derivatives. For SR networks, adding noise to the input is destructive to the output image [24, 13]. Thus both SmoothGrad and VarGrad can not be used to interpret SR networks. On the other hand, SmoothGrad and VarGrad also face the challenge of gradient saturation.

CAM, GradCAM and Guided GradCAM. Different from the aforementioned gradient-based attribution methods, Class Activation Mapping (CAM) [40] generates class activation maps using the global average pooling in convolution neural networks. A CAM map for a particular category indicates the discriminative image regions used by the network to identify that category. Combining gradient-based methods and CAM, Selvaraju *et al.* [25] further propose GradCAM that corresponds to the gradient of the class score w.r.t. the feature map of the last convolution unit. For pixel level granularity, GradCAM can be combined with Guided Backpropagation through an element-wise product. Since CAM is specially designed for high-level vision networks with global pooling layers, it cannot be easily adapted to low-level vision models such as SR networks.

Perturbation-Based Methods. Different from the above works that require the mathematical details of the model, there are works that treat deep models as black-boxes. These methods usually localize the discriminative image regions by performing perturbation to the input. For instance, Fong and Vedaldi [11] propose to explain neural networks that are based on learning the minimal deletion to an image that changes the model prediction. Similar to SmoothGrad and VarGrad, the sensitivity of SR networks to disturbances and perturbation makes it difficult to use these approach to explain.

2. Collection of Models

In this section, we first describe the training settings in our experiments and then briefly review the used SR networks. We use DIV2K training set [2] for training and the size of LR image is 64×64. For optimization, we use Adam [17] with the default settings that $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initialized as 1×10^{-4} and decayed linearly at every 2×10^{5} updates. The size of minibatch is set to 16. We next briefly review the used SR networks.

Early methods with fully convolutional architectures. These methods include SRCNN [9], FSRCNN [10] and ESPCN [27]. What they have in common is that they only use stacked convolution layers without residual or other deep modules. SRCNN is the first deep SR network that consists of only three convolution layers without upsampling layer – it takes the bicubic interpolation result as input. FSRCNN consists of eight convolution layers and uses deconvolution layer as the upsampling layer. In ESPCN, pixel shuffle is used innovatively as an upsampling operation, and this operation is used on a large scale by subsequent SR networks. In addition to the above networks, DDBPN [14] and LapSRN [18] are also in the form of fully convolution networks with different convolution strategies. LapSRN is a network with progressive upsampling operations that super-resolves low-resolution images in a coarse-to-fine laplacian pyramid framework. DDBPN exploits iterative up- and downsampling layers, aiming at providing an error feedback mechanism for projection errors at each stage.

Networks with residual and dense connections. These methods date back to SRResNet [19] that first introduce residual connections [15] to deep SR networks. Some methods are proposed to improve the residual structure such as EDSR [21], CARN [3] and MSRN [20]. Spatial feature transformation blocks are also introduced to SR networks [34, 13] to achieve interactive SR. Inspired by dense connection network [16], RDN [38] and SRDenseNet [33] with dense architecture was proposed. Combining residual blocks and dense connections, residual-in-residual dense net (RRBDNet) [35] was proposed. Recently, DRLN [5] employs cascading residual on the residual structure to allow the flow of low-frequency information to focus on learning high and mid-level features.

Networks with attention modules. In addition to innovations in various short connections, attention modules are also used to improve the performance of SR networks. Zhang *et al.* [36] propose channel attention that compute attention weights w.r.t. the whole channel. Zhao *et al.* [39] propose pixel attention that compute attention weights using 1×1 convolution for each pixel. Non-local operation is also introduced in the form of attention module in [37, 22]. SAN [8] utilizes both non-local attention modules and second-order channel attention. Recently, CSNLN [23] employs cross-scale non-Local attention module with integration into a recurrent neural network to learn cross-scale feature correlation.

3. More Results

In this section, we exhibit more results. We first show more examples of the "area of interest". In Figure 8 of the main text, we have shown the five images with the smallest area of interest and also five images with the largest area of interest. In Figure 1, we show more images with their area of interest and the rank indices are also marked. In Figure 2, Figure 3, Figure 4, and Figure 5, we show more LAM results.

References

[1] Julius Adebayo, Justin Gilmer, Ian Goodfellow, and Been Kim. Local explanation methods for deep neural networks lack sensitivity to parameter values. *arXiv preprint arXiv:1810.03307*, 2018. 2

[2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 2

[3] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 252–268, 2018. 2

[4] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018. 1

[5] Saeed Anwar and Nick Barnes. Densely residual laplacian super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[6] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 1

[7] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010. 1

[8] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 2

[9] 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. 2

[10] 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

[11] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017. 2

[12] Eric J Friedman. Paths and consistency in additive cost sharing. *International Journal of Game Theory*, 32(4):501–518, 2004. 1

[13] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019. 2

[14] 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*, pages 1664–1673, 2018. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[18] 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*, pages 624–632, 2017. 2

[19] 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

[20] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 517–532, 2018. 2

[21] 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

[22] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *Advances in Neural Information Processing Systems*, pages 1673–1682, 2018. 2

[23] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5690–5699, 2020. 2

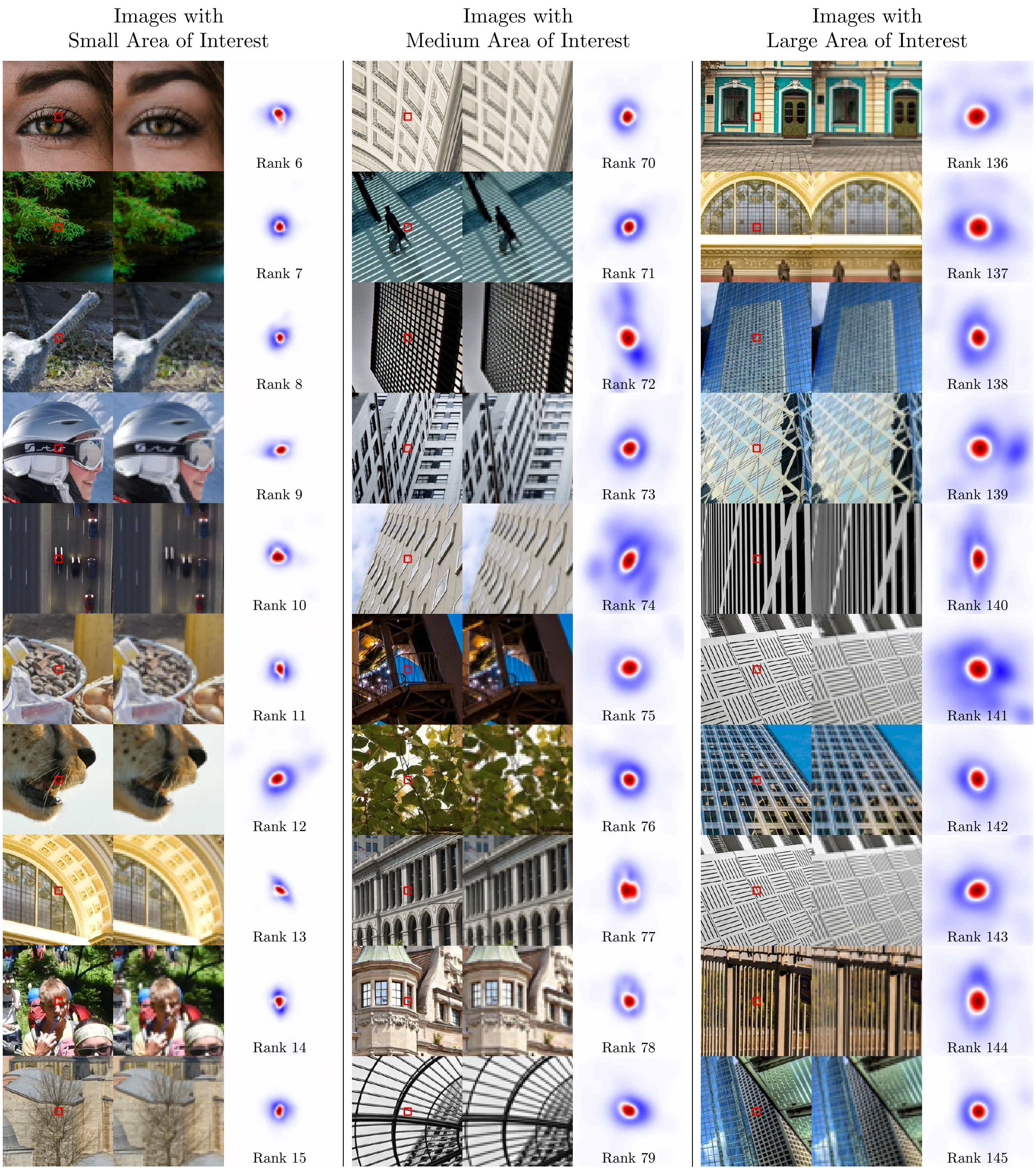

Figure 1: The heat maps exhibit the area of interest for different SR networks. The pixels with red color are noticed by almost all SR networks while the areas marked with blue represents the differences between the SR networks with large LAM interest areas and those with small interest areas. The rank indices indicate the ranking order of the largest diffusion index of images' lam results.

[24] Guocheng Qian, Jinjin Gu, Jimmy S Ren, Chao Dong, Furong Zhao, and Juan Lin. Trinity of pixel enhancement:

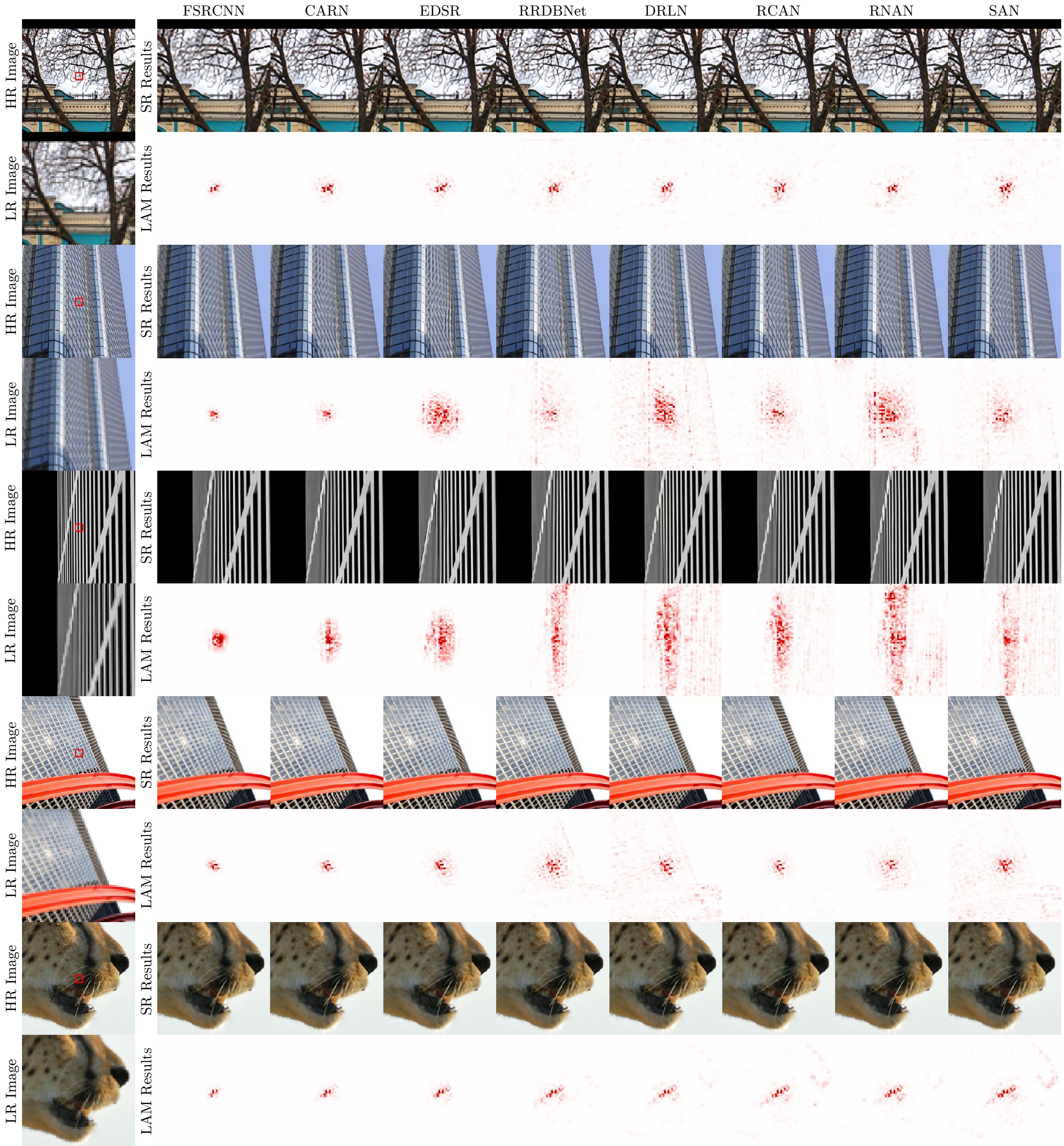

Figure 2: Comparison of the SR results and LAM attribution results of different SR networks. The LAM results visualize the importance of different pixel w.r.t. the SR results.

a joint solution for demosaicking, denoising and super-resolution. *arXiv preprint arXiv:1905.02538*, 2019. 2

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

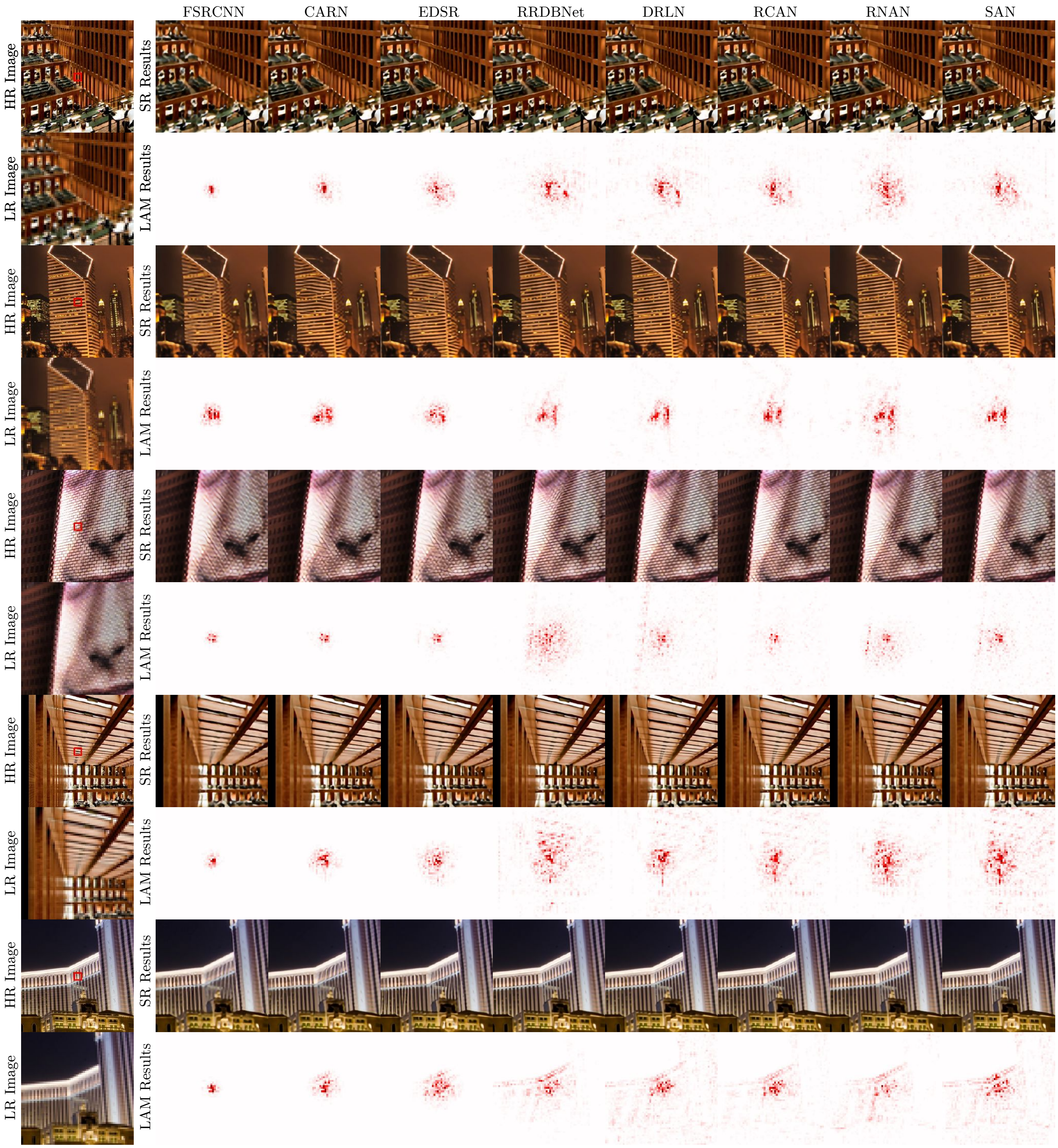

Figure 3: Comparison of the SR results and LAM attribution results of different SR networks. The LAM results visualize the importance of different pixel w.r.t. the SR results.

[26] Junghoon Seo, Jeongyeol Choe, Jamyoung Koo, Seunghyeon Jeon, Beomsu Kim, and Taegyun Jeon. Noise-adding methods of saliency map as series of higher order partial derivative. *arXiv preprint arXiv:1806.03000*, 2018. 2

[27] 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

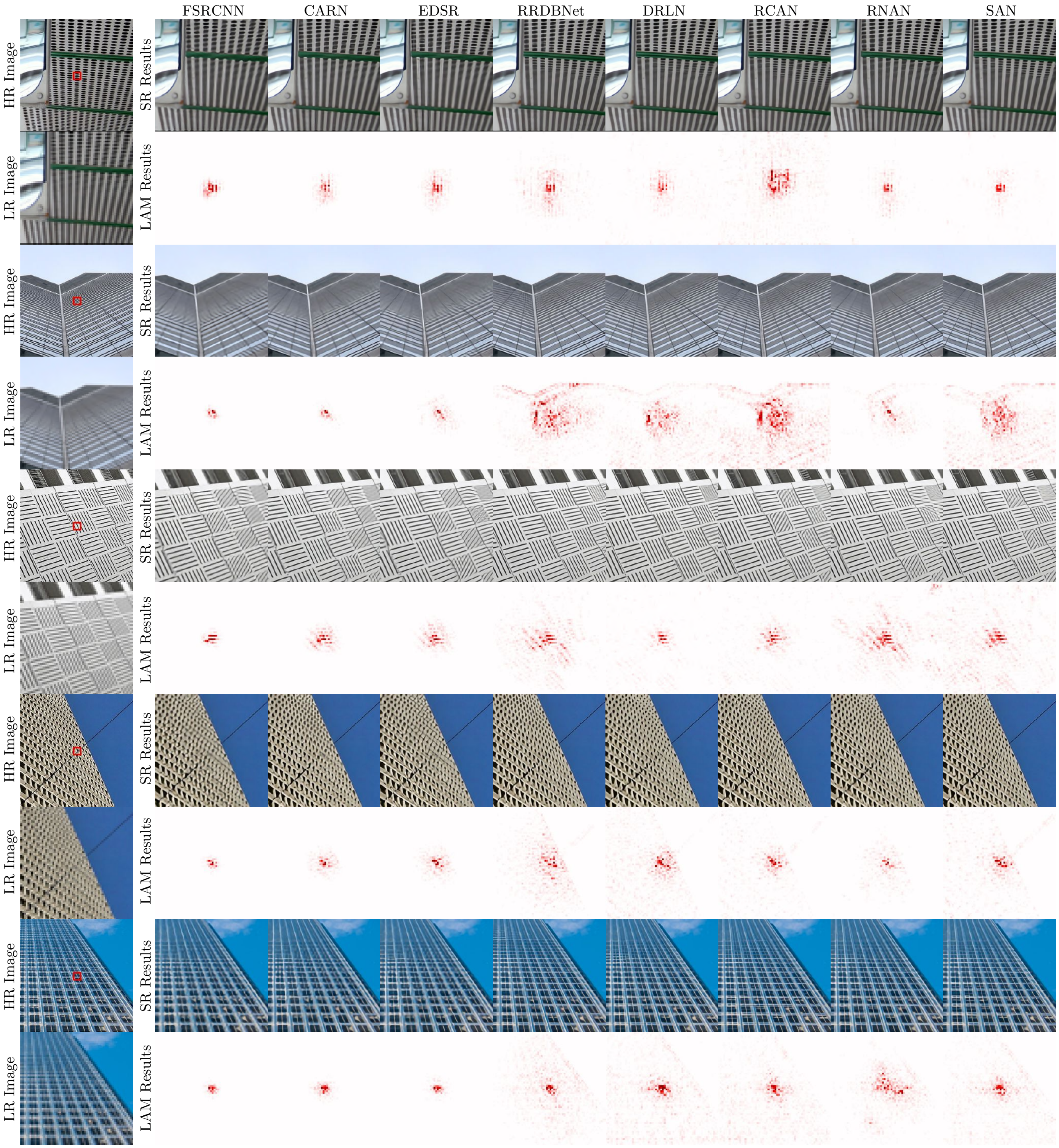

Figure 4: Comparison of the SR results and LAM attribution results of different SR networks. The LAM results visualize the importance of different pixel w.r.t. the SR results.

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1874–1883, 2016. 2

[28] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016. 1

[29] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image

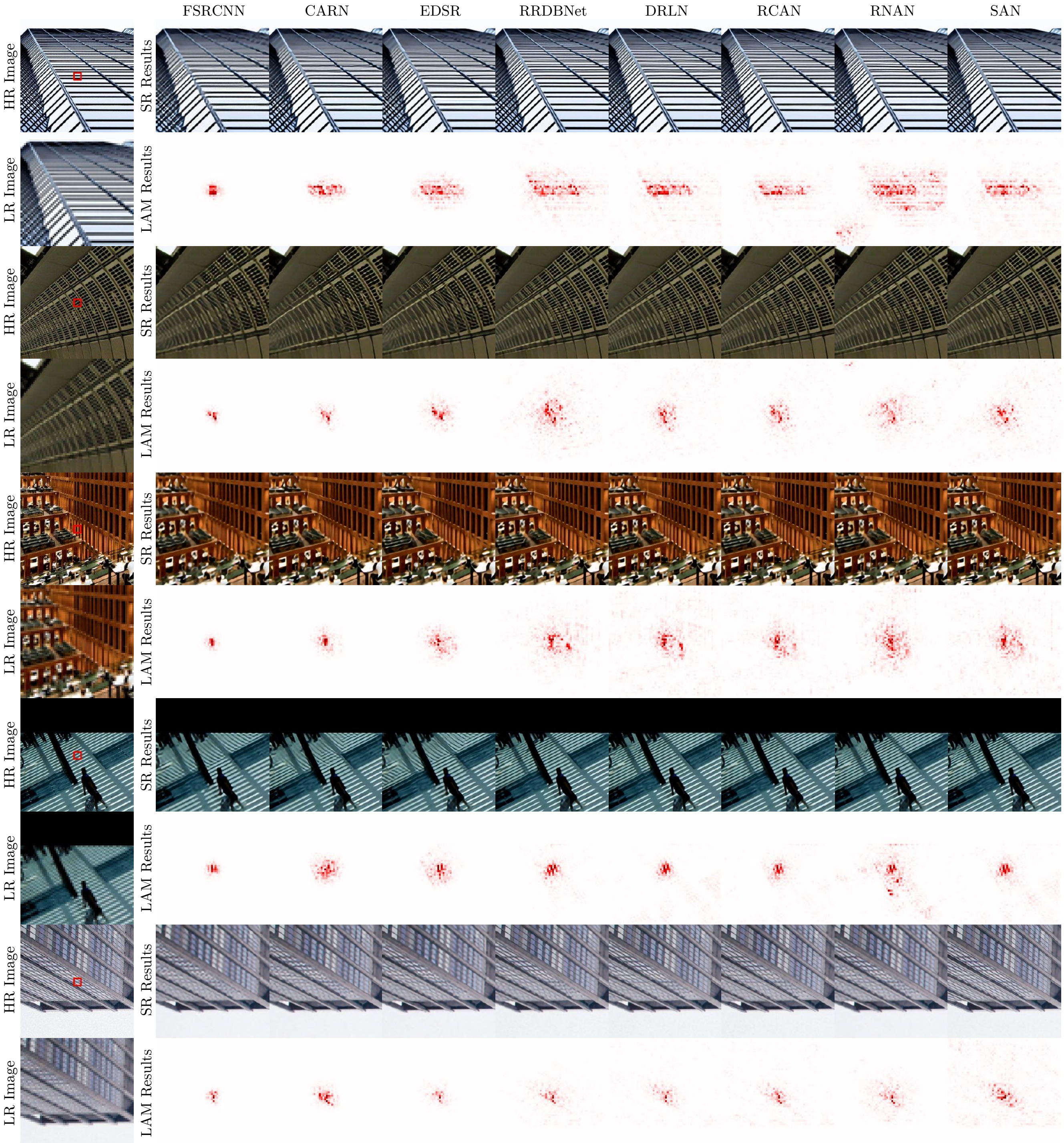

Figure 5: Comparison of the SR results and LAM attribution results of different SR networks. The LAM results visualize the importance of different pixel w.r.t. the SR results.

classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1

[30] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 2

[31] J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015. 1

[32] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020. 1

[33] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4799–4807, 2017. 2

[34] 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*, pages 606–615, 2018. 2

[35] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision*, pages 63–79. Springer, 2018. 2

[36] 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)*, pages 286–301, 2018. 2

[37] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 2

[38] 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*, pages 2472–2481, 2018. 2

[39] Hengyuan Zhao, Xiangtao Kong, Jingwen He, Yu Qiao, and Chao Dong. Efficient image super-resolution using pixel attention. In *European Conference on Computer Vision Workshop (ECCVW)*. Springer, 2020. 2

[40] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2