ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095368

LEARNING TO EXPLAIN: A GRADIENT-BASED ATTRIBUTION METHOD FOR INTERPRETING SUPER-RESOLUTION NETWORKS

Anni Yu, Yu-Bin Yang

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

ABSTRACT

DNN-based super-resolution(SR) models inherit the black-box nature of DNN and present low transparency. However, few works focus on interpreting low-level SR models and the limited existing gradient-based interpretability methods often require large computational costs and produce spurious/noisy feature attributions with a mixture of positive and negative signals. In this paper, we propose a gradient-based attribution method L2X(Learning to eXplain) to provide post-hoc visualization and interpretation for SR models by quantifying the attribution of individual features with regard to the SR output and generating a heatmap in pixel/input space. L2X relies on the forward-pass activations and efficiently propagates the important signal from the output neuron through the layers to the input in one pass. Besides, both positive and negative attributions are taken into account during the back-propagation process. We conduct cross-architectural analysis on State-of-the-Art SR networks and investigate the inner workings of them. We experimentally demonstrate the potential of L2X as a research tool to diagnose and visualize the most relevant features of the current SR model and offer insights into SR models to make further improvements.

Index Terms— Super-resolution, interpretability, attribution, deep neural networks, explainable artificial intelligence

1. INTRODUCTION

Deep Neural Networks(DNN) have constantly pushed the State-of-the-Art in image super-resolution(SR). The academic literature has provided diverse model architectures and has offered myriad techniques to improve SR performance. Despite this outstanding achievement, the black-box nature of DNN remains a barrier to adoption when interpretability is considered as a critically important metric or at least as important as model's performance[1]. Interpretation of SR networks entails potential benefits including but not limited to: 1) understanding the inner workings of networks, 2) revealing the most relevant features to the current network, and 3) comparing the differences in information usage between various network architectures.

In our work, we follow the line of decomposing the prediction made by the network into attribution values(also called "relevance" or "contribution") of the input features. As shown in Fig.1, our method operates in a local scope in that it generates an attribution map for a given model and input instance, and features relevant to the SR result are highlighted in the attribution map. The attribution of a feature is calculated with respect to differences from a reference state, which is opted to be a default/neutral state to represent the absence of information. Precisely, we chose the blurred version of the input LR sample as the reference state. Moreover, in view of the well-known "saturation" problem[2], we distinguish the effects of both positive and negative contributions at nonlinear activation layers, which can reveal relevancy missed by other gradient-based methods.

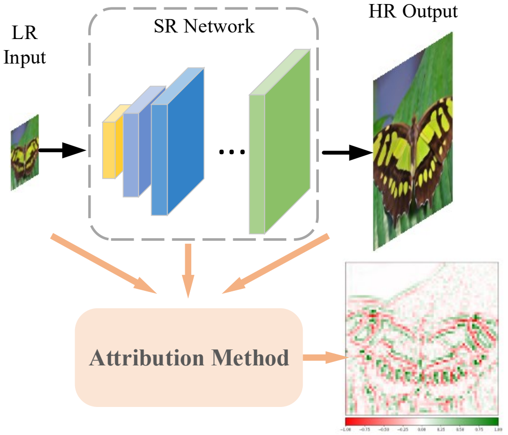

Fig. 1. General settings for L2X. An attribution map is generated for a specific LR input, SR network and HR output.

2. RELATED WORK

SR networks. There has been extensive research on designing powerful and efficient SR networks, including deeper architectures, global and local residual learning[3], recursive learning[4], attention mechanism[5, 6, 7, 8], etc. Recently, Dosovitskiy et al. introduced Vision Transformer(ViT)[9] which utilizes a self-attention mechanism to capture long-range information between sequence elements. Due to its strong representation capability, ViT launched an inevitable revolution in computer vision tasks. Subsequently, Liang et al. proposed SwinIR[10] for image restoration which is based on Swin Transformer. In contrast with offering a taxonomy of SR networks and comparing their performances, few papers work on interpreting and understanding the inner workings of different SR networks, which is the primary theme of our work.

Network Interpretation. Our work is closely related to network interpretation methods. A plethora of approaches follow the line of back-propagating an important signal from an out neuron through each layer of the network down to the input features, for instances, saliency map[11], deconvolutional networks[12], Guided Backpropagation[13], and DeepLIFT[2], to name just a few. These approaches mainly differ in the handling of activation functions(e.g.,

This work is funded by the Natural Science Foundation of China under Grants 62176119. (Corresponding author: Yu-Bin Yang)

Authorized licensed use limited to: Syracuse University Libraries. Downloaded on August 23,2023 at 14:25:28 UTC from IEEE Xplore. Restrictions apply.

ReLU, MaxPool). A majority of interpretability methods focus on high-level vision tasks including image classification, segmentation, etc. Recently, Gu et al. [14] proposed the Local Attribution Map(LAM) to conduct attribution analysis of low-level SR networks based on Integrated Gradients. Unlike LAM which highlights important regions given a small meaningful image patch, our proposed L2X propagates the gradients with regard to all the input features. Besides, LAM only focuses on the attribution of CNN-based SR networks and cannot distinguish positive and negative attributions of each feature. Moreover, the computation cost for Integrated Gradients is relatively high in that estimating the average gradient requires the numerical calculation of an integral, while L2X only efficiently propagates the gradients in one pass.

3. METHOD

3.1. Overview of L2X

Our method explains the SR network's predictions by assigning each neuron with an attribution score that represents the difference between activation at the input sample with the activation at some reference input. It works in a backward manner and relies on the forward-pass activations to propagate from later layers to preceding layers in the network.

Formally, we consider an SR network F which takes an N-dimensional input $x = [x_1, x_2, \cdots, x_N] \in \mathbb{R}^N$ and produces a T-dimensional output $F(x) = [F_1(x), F_2(x), \cdots, F_T(x)] \in \mathbb{R}^T$, where $T = s^2N$ denotes the total number of output neurons with an upscaling factor s. Given a single target output neuron t, our goal is to determine the attribution $R_t(x_i) = [R_t(x_1), R_t(x_2), \cdots, R_t(x_N)]$ of each input feature x_i to the output $F_t(x)$ with regard to some reference $F_t(\bar{x})$. Thus, the problem converts to how to assign the attribution of neuron i to neuron t, s.t.:

$$\sum_{i=1}^{N} R_t(x_i) = F_t(x) - F_t(\bar{x}) \tag{1}$$

Note that the left side sums up to the difference-from reference $F_t(x) - F_t(\bar{x})$. Therefore, our method satisfies the *Sensitivity-N*[15] property which holds that the sum of the attributions for any subset of features of cardinality N is equal to the variation of the output caused by removing such features. Moreover, in accordance with [2], we also use *multipliers* $m_t(x)$ to multiply with x instead of computing the contribution scores directly. Thus, the attribution score is computed as:

$$R_t(x) = m_t(x)(x - \bar{x}) \tag{2}$$

3.2. The Choice of Reference

As mentioned before, a reference input is a pre-requisite to formulate our method and references for all neurons can be calculated by choosing a reference input and propagating activations through the network. Attribution results vary with the choices of reference inputs. In most cases, a black image with zero-value pixels is the canonical choice, especially for high-level vision tasks[2, 16]. However, such a choice tends to focus on the intensity of the image rather than fine-grained details which are helpful for SR networks. The reference should be chosen to best represent the absence of information[15]. Therefore, in order to obtain insightful interpretations, special reference input is tailored to the SR task in our work. As the low-frequency features of the LR image contribute less to the ultimate SR output than high-frequency features, we choose the blurred version of the input LR image as the reference input, which is computed as the convolution of the LR image with a 2D Gaussian kernel with variance α.

3.3. Separating Positive and Negative Contributions

Let neuron t be a nonlinear transformation of input x, then we have $m_t(x) = \frac{F_t(x) - F_t(\bar{x})}{x - \bar{x}}$. For simplicity, the activation difference from reference $F_t(x) - F_t(\bar{x})$ and the input difference from reference $x - \bar{x}$ are denoted as Δt and Δx, respectively. Then, we introduce Δt^+ and Δt^- to represent the positive and negative components of Δt. The detailed formula for estimating Δt^+ and Δt^- are given as:

$$\Delta t^+ = \frac{\Delta t}{\Delta x} \Delta x^+ \tag{3}$$

$$\Delta t^- = \frac{\Delta t}{\Delta x} \Delta x^- \tag{4}$$

In the situation where $\Delta x \to 0$ and $\Delta t \to 0$(i.e., x infinitely approaches $\bar{x}$), the computation of *multiplier* approaches the derivative $\frac{dt}{dx}$ at $x = \bar{x}$. Therefore, we use the gradient instead of $\frac{\Delta t}{\Delta x}$ for estimating $m_t(x)$ to avoid the numerical instability issue caused by a small denominator.

In this manner, the resulting attribution map can detect both positive and negative evidence that might be present in the input. Yet, other attribution methods tend to take the absolute value(e.g., Sensitivity Analysis[11]) or simply eliminate negative signals during the back-propagation.

3.4. Back-propagation Rules

In this subsection, we show the rules for back-propagating activations to assign contribution scores for each neuron to its immediate inputs.

Let $R_i^{(l)}$ be the attribution score of neuron i of layer l. As Eq.(1), the back-propagation process starts from the output layer L, assigning the attribution of target neuron t equal to the activation difference between t and reference $\bar{t}$, and the attribution of other neurons to zero.

$$R_i^{(L)} = \begin{cases} F_i(x) - F_i(\bar{x}), & \text{if neuron } i \text{ is the target neuron} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Consider $z_{ij} = b + x_i w_{ij}$ the weighted activation of neuron i onto neuron j and b as the additive bias term. Then, the prediction $F_t(x)$ is redistributed through hidden layers down to the input layer. This recursive rule is defined in Eq.(6).

$$R_i^{(l)} = \sum_j \frac{z_{ij} - \bar{z}_{ij}}{\sum_{i'} z_{i'j} - \sum_{i'} \bar{z}_{i'j}} r_j^{(l+1)} \tag{6}$$

Where $\bar{z}_{ij} = b^{(l)} + \bar{x}_i^{(l)} w_{ij}^{(l,l+1)}$ is the weighted activation of neuron i onto neuron j given the reference input $\bar{x}$.

4. EXPERIMENTS

4.1. Experimental Settings

In this subsection, we describe the experimental settings in detail, including datasets, evaluation metrics, and SR networks. We use three benchmark datasets in SR: DIV2K[17] as the training and validation set, Urban100[18] and Manga109[19] as test sets, and obtain their ×4 SR results based on different models. We evaluate the

Authorized licensed use limited to: Syracuse University Libraries. Downloaded on August 23,2023 at 14:25:28 UTC from IEEE Xplore. Restrictions apply.

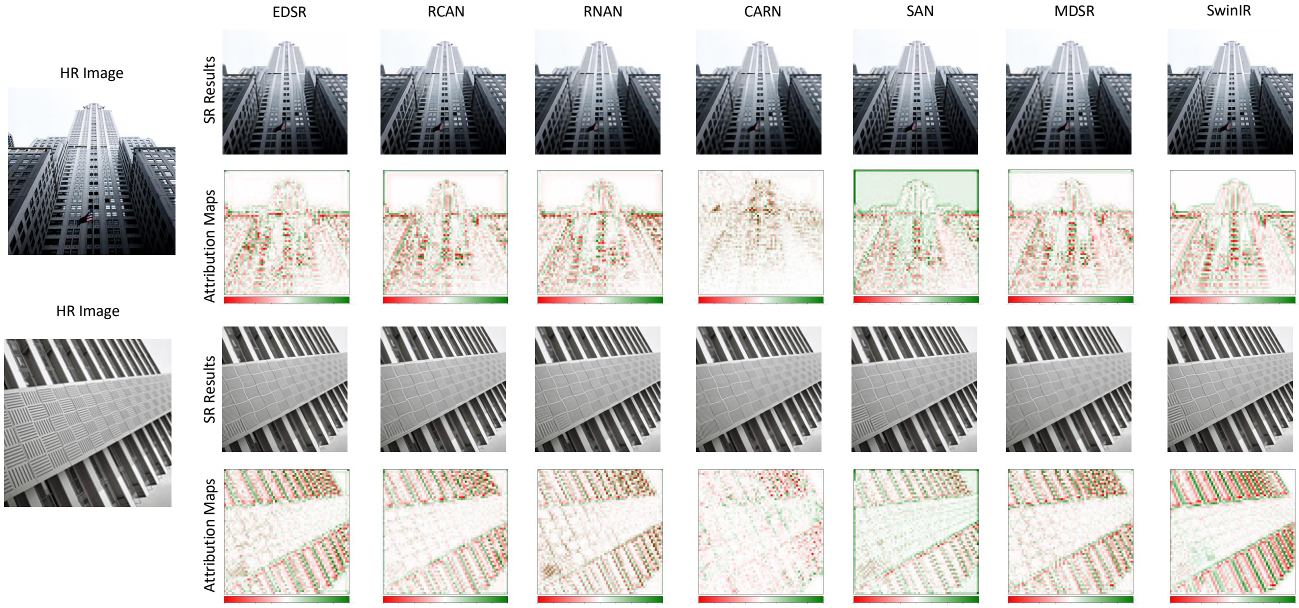

Fig. 2. Comparison of SR results and attribution maps generated by different SR networks.

SR performance with the signal-to-noise ratio (PSNR) and structural similarity image index (SSIM) metrics. In our experiment, up to 10 state-of-the-art SR networks are trained for analysis and interpretation, which consists of 8 CNN-based networks and 2 Transformer-based networks listed in Table 1.

Table 1 shows the quantitative results of different SR networks on Urban100 and Manga109 datasets at upscaling factor ×4 . SwinIR and its variant Swin2SR surpass CNN-based SR models by a large margin on both testsets. This invokes the motivation to interpret and distinguish the inner workings of different models and investigate the representations learnt by their hidden neurons.

Table 1. Quantitative results(average PSNR/SSIM) with different SR networks on two benchmark datasets: Urban100 and Manga109

Method	Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM
EDSR[3]	26.53	0.7995	30.97	0.9146
MDSR[20]	26.07	0.7851	30.57	0.9085
CARN[4]	26.35	0.7741	30.63	0.9100
RCAN[5]	26.79	0.8072	31.24	0.9175
DRLN[7]	26.56	0.7998	31.10	0.9152
HAN[8]	25.10	0.7497	28.72	0.8772
SAN[21]	25.63	0.7692	29.81	0.8998
MSRN[22]	26.12	0.7866	30.56	0.9088
SwinIR[10]	27.41	0.8235	32.09	0.9258
Swin2SR[23]	27.44	0.8243	32.03	0.9254

The attribution maps generated by L2X on some representative SR networks and their corresponding SR results are presented in Fig.2. The proposed L2X can be applied to diagnose and visualize the information learnt by SR models.

In the first example, the attention-based SAN network and transformer-based SwinIR both reconstruct a sharper edge of the skyscraper apart from the sky background, while the lightweight CARN restores a blurry edge or even extracts misleading features of the background. This observation is confirmed in their corresponding attribution maps where pixels highlighted in green indicate great relevancy w.r.t. the SR results, whereas pixels highlighted in red indicate low relevancy. Besides, SwinIR has a stronger representation capability compared to other SR networks in that it can capture clear features regardless of image intensity(like the windows in the shadow). In the second example, only SAN and SwinIR reconstruct the correct and clear textures of the building, while the other models reconstruct blurred textures. This shows the superiority of SAN to utilize both non-local attention and second-order channel attention. The evidence can be found in the below attribution maps where SAN and SwinIR assign positive attributions to the corresponding features.

4.2. Ablation Study

In order to quantitatively evaluate the reliability and fidelity of the attribution maps generated by the proposed L2X method, we use an efficient remove-and-replace strategy to remove the supposedly informative features according to the attribution results from the input and replace them with the per channel mean. This process is also termed as "mask". Then, the corrupted LR inputs are fed into the same SR network.

In Fig.3(a), the first row presents the PSNR variations in terms of masking features in the range from 5% to 40% on the Urban100 dataset, and the second row presents the SSIM variations accordingly. The dashed orange line indicates the results when we randomly remove-and-replace certain features while the solid blue line for the results when features are masked according to their attribution scores in descending order.

Since our method describes the marginal effect of a feature on the SR results in terms of the same input with such features being removed, we observed an expected larger performance drop when compared to a random removal strategy. For instance, PSNR descends from 26.53dB to 13.97dB for EDSR at mask-ratio 40% when

Authorized licensed use limited to: Syracuse University Libraries. Downloaded on August 23,2023 at 14:25:28 UTC from IEEE Xplore. Restrictions apply.

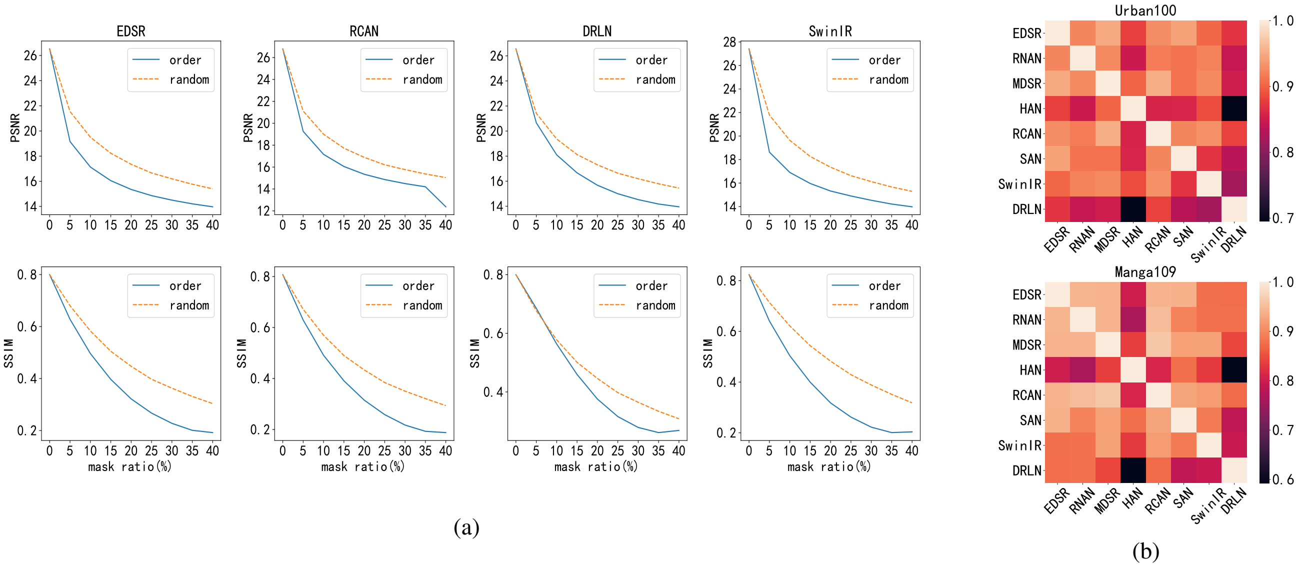

Fig. 3. (a) PSNR and SSIM results for different SR networks with regard to varying mask ratios on Urban100 dataset; (b) CKA analysis of L2X attribution results on Urban100 and Manga109 datasets.

features are masked in attribution order, compared to a relatively small drop from 26.53dB to 15.40dB in random order. Moreover, with the growth of mask-ratio, both PSNR and SSIM witness a rapid decline at an early stage($<$ 10%), and this downward trend becomes stable. This indicates a phenomenon of "the law of the vital few" which confirms the fact that the majority of feature contributions come from the minority of features.

Furthermore, we observe differences in information usage among different SR models. The average gap between two lines in the first row is around 1.6dB with the maximum approaching 3.15dB for SwinIR and the minimum approaching 0.76dB for DRLN at mask-ratio 5%. This indicates that SwinIR is more sensitive to feature removal compared with other SR networks, whereas DRLN is more efficient in the usage of information in that its PSNR at 5% mask-ratio is 20.66, even surpassing all the others including the best SwinIR. What's more, the SSIM of ordered masking slightly exceeds that of random masking at mask-ratio 5% for DRLN, which is different from all the other models.

4.3. Exploration with L2X

For quantitative analysis of representations learnt by hidden neurons across different SR models, we use centered kernel alignment(CKA) which computes the similarity of attribution matrices[24]. CKA takes as input $\mathbf{R_1} \in \mathbb{R}^{m \times n_1}$ and $\mathbf{R_2} \in \mathbb{R}^{m \times n_2}$ which are attribution matrices of two layers of different models calculated by L2X, with n_1 and n_2 neurons, respectively, evaluated on the same amount of m samples. Let $\mathbf{K} = \mathbf{R_1}\mathbf{R_1}^T$ and $\mathbf{L} = \mathbf{R_2}\mathbf{R_2}^T$ represent the Gram matrices for the two layers which measure the similarity of a pair of samples according to their attribution scores, CKA formulates:

$$CKA(\mathbf{K}, \mathbf{L}) = \frac{HISC(\mathbf{K}, \mathbf{L})}{\sqrt{HSIC(\mathbf{K}, \mathbf{K})HSIC(\mathbf{L}, \mathbf{L})}} \tag{7}$$

where $HSIC(\mathbf{K}, \mathbf{L})$ represents the Hilbert-Schmidt Independence Criterion. The resulting CKA measure is between 0 and 1, and the closer it is to 1, the more dependent the pair are, conversely, the closer it is to 0, the more independent they are.

Fig.3(b) shows the heatmaps of similarities between attribution results of different SR models on Urban100 and Manga109 datasets, respectively. In this experiment, we only consider positive attributions. The heatmap indicates fairly significant similarities across models in that high values(above 0.7) appear off the main diagonal. EDSR, RNAN, MDSR and SAN share correlated feature-level information learnt by neurons of the networks in that the paired CKA measure is above 0.8.

HAN gets the lowest PSNR among all the other SR networks and obtains a distinct and unusual attribution compared with others. In this case, involving more informative features may be helpful to performance gain. Although the performance of DRLN is close to other CNN-based methods, like RCAN and EDSR, it assigns slightly different attributions to the input features. This indicates that the way to utilize the information within receptive area also plays an important role in designing more powerful SR networks.

5. CONCLUSION

In this work, we present L2X, a gradient-based attribution method for visualizing and interpreting SR networks. L2X quantifies the attributions of input features w.r.t. the SR output and generates a heatmap highlighting relevant features for a given model and an input instance. By separately considering both positive and negative attributions, L2X can reveal relevancy missed by other gradient-based attribution methods. We conduct experiments to quantitatively evaluate the fidelity of the proposed L2X. Besides, we also conduct cross-architecture experiments among 10 representative SR networks and arrive at several useful findings of the learnt information within neurons of different networks. These findings are also very pertinent to demonstrate the potential of L2X as a research tool for diagnosing and visualizing the SR models.

Authorized licensed use limited to: Syracuse University Libraries. Downloaded on August 23,2023 at 14:25:28 UTC from IEEE Xplore. Restrictions apply.

6. REFERENCES

[1] Zachary C Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[2] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*. PMLR, 2017, pp. 3145–3153.

[3] 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*, 2017, pp. 136–144.

[4] 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)*, 2018, pp. 252–268.

[5] 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)*, 2018, pp. 286–301.

[6] 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.

[7] Saeed Anwar and Nick Barnes, "Densely residual laplacian super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1192–1204, 2022.

[8] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen, "Single image super-resolution via a holistic attention network," in *European conference on computer vision*. Springer, 2020, pp. 191–207.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[10] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.

[11] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[12] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[13] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[14] Jinjin Gu and Chao Dong, "Interpreting super-resolution networks with local attribution maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9199–9208.

[15] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross, "Gradient-based attribution methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 169–191. Springer, 2019.

[16] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[17] Eirikur Agustsson and Radu Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[18] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5197–5206.

[19] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa, "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools and Applications*, vol. 76, no. 20, pp. 21811–21838, 2017.

[20] 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*, 2017, pp. 136–144.

[21] 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/CVF conference on computer vision and pattern recognition*, 2019, pp. 11065–11074.

[22] 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)*, September 2018.

[23] Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte, "Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration," *arXiv preprint arXiv:2209.11345*, 2022.

[24] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton, "Similarity of neural network representations revisited," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3519–3529.

Authorized licensed use limited to: Syracuse University Libraries. Downloaded on August 23,2023 at 14:25:28 UTC from IEEE Xplore. Restrictions apply.