Interpreting Image Super-Resolution in Artificial Neural Networks from Global and Local Views

Xiaochen Liu*†, Alexander Jacob†, Wei Song‡, Antonio Liotta*
*Faculty of Engineering. Free University of Bozen-Bolzano, Bolzano, Italy
Email: xiliu@unibz.it, antonio.liotta@unibz.it
†Institute for Earth Observation, Eurac Research, Bolzano, Italy
Email: alexander.jacob@eurac.edu
‡College of Information. Shanghai Ocean University, Shanghai, China
Email: wsong@shou.edu.cn

***Abstract*—Work on image super-resolution (SR), to construct higher-resolution images starting from low-quality ones, has focused primarily on reconstruction algorithms and specific application domains. In this work, we aim at methods to aid interpreting SR inner-working, with a view to improve explainability. We propose a novel gradient-based attribution approach, to provide interpretations from global and local perspectives, dubbed glocal attribution map (GL-AM). After verification with five different SR models, we show that GL-AM: (1) is a powerful tool to understand the principles of SR networks from both global and local views; (2) provides the consensus and variation sensitivity of different models to the input; (3) is more effective to emphasize the features captured by the attention mechanism (for the SR model) through feature re-calibration; (4) is more computationally efficient and more effective as the region of interest increases.**

***Index Terms*—explainable AI, super-resolution, data interpretation, deep learning**

I. Introduction

Deep learning has shown its power in the field of low-level tasks such as image super-resolution (SR). With the help of deep learning (DL), SR has surfaced as an indispensable technological innovation, substantially elevating the granularity and fidelity of images beyond their inherent resolution [2]–[4], [19]. Despite the advances achieved by deep learning in the domain of SR, it concurrently contends with intrinsic complexities that confound its comprehensive elucidation. For instance, we all know that deep learning is deemed a 'black-box' model, and it is often criticized for its lack of transparency. It is, in fact, difficult to comprehend the SR model's transformation of input data into super-resolved outputs. This opacity can limit trust and improvement of SR tasks. Therefore, an in-depth study of the inner mechanisms of these models can help us understand their limitations and discover possible improvements for SR models.

Explainable Artificial Intelligence (XAI) [21] refers to the methodologies and processes designed to make the mechanisms and outcomes of DL algorithms and AI systems more comprehensible and trustworthy for human operators. XAI is paramount in sectors where critical decisions are made, such as healthcare, finance, and environmental monitoring. In these fields, the clarity surrounding the decision-making process of AI systems is as crucial as the decisions themselves. However, the development of XAI algorithms is kind of unbalanced currently, with a greater focus on classification problems [1]. Most methods that visualize and highlight the regions of input contributing most significantly to classification results, such as Grad-CAM [10], focus on gradient-based class activation mappings. Meanwhile, layer-wise relevance propagation [9] decomposes predictions to understand contributions at the neuron level.

These methods are not applicable to SR neural networks. In particular, LAM [11] serves as an XAI method designed to interpret the SR deep learning model by employing the integrated path gradient. Nevertheless, LAM possesses limitations, as it lacks the capability to delve into the internal workings of the model, providing explanations solely for the influence of specific features on the outcomes. Additionally, the computation involved in utilizing the integrated path gradient proves to be relatively time-consuming.

In this paper, we proposed an XAI method from both global and local views, dubbed glocal attribution map (GL-AM), to interpret the deep learning SR model. Globally, GL-AM elucidates the contributions of individual regions within the image to the overall image SR reconstruction. Locally, GL-AM identifies pixels with substantial influence on the SR outcomes. Furthermore, by leveraging both the visualization and quantitative results provided by GL-AM, we delve into the reasons behind the varying performance levels of SR models with different depths. We also explore how the attention mechanism contributes to the efficacy of SR models. Our findings reveal that the attention mechanism significantly enhances the performance of SR networks, allowing them to achieve comparable or superior results with smaller receptive fields. This is particularly evident when combined with functional recalibration, which further optimizes model efficiency. In Fig. (1), we provide some general representative results. GL-AM interprets the low-resolution input pixels that affect the SR results by analyzing the feature map of each convolution layer and its corresponding gradient, rather than using integral path gradient, as shown in Fig. (1) (a) and (b).

979-8-3503-5475-1/24/$31.00 ©2024 IEEE

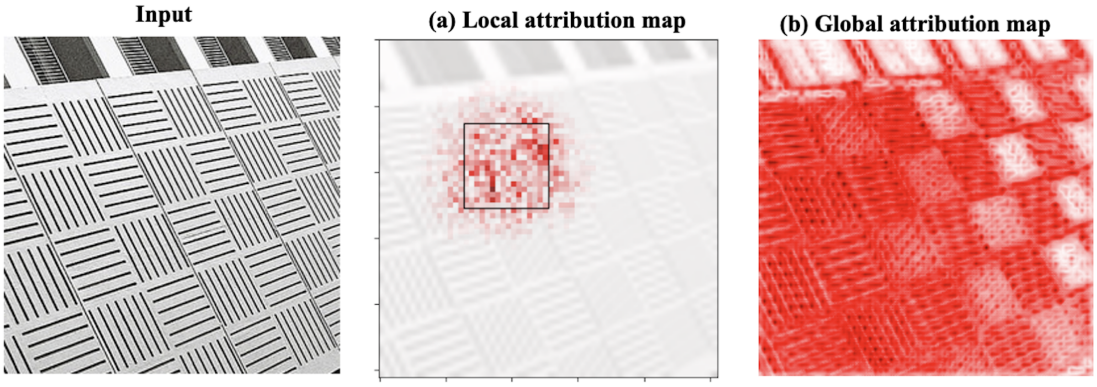

Fig. 1. This demo presents the results of the proposed attribution method GL-AM applied to the image super-resolution (SR) networks CARN [16]. Figure (a) and (b) illustrate the importance of pixels that influence the reconstruction of both the region of interest (ROI) and the overall image. Higher intensity (darker red pixels) indicates a greater influence on the SR outcomes.

II. Related Work

Model explanation and model interpretation are two important topics related to the field of artificial intelligence. Both of them aim to increase the transparency and understanding of neural networks. Because neural networks are often seen as 'black boxes' and their internal decision-making processes are not transparent to the end user. Currently, there are works that try to explore what happens inside the neural network and visualize how the input data affect the outputs [6]–[8], [12], [13]. However, much of this research has been in the service of classification tasks.

Selvaraju et al. [10] proposed Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the decision-making process of convolutional neural networks (CNNs) for classification tasks. Technically, this method highlights the important regions in the image for predicting the class by calculating the input importance based on gradient and classification score. Simonyan et al. [9] proposed saliency maps for visualizing and interpreting how convolutional neural networks (CNNs) classify images and identify the features they perceive as significant. Saliency maps are generated by computing the gradient of the output category with respect to the input image. This gradient implies how much the class score affects the pixel's value, thus indicating the pixel's importance in the network's decision-making process. Pixels with higher gradient values are believed to be more influential, and their collective representation forms the saliency map, visually illustrating the regions crucial for the classification decision. Shrikumar et al. [14] proposed a novel gradient-based interpreting method for CNNs (DeepLIFT), which attributes importance to input features. DeepLIFT distinguishes itself by comparing the activation of each neuron to its 'reference activation' and assigns contribution scores based on the difference, effectively handling scenarios where other methods like gradient-based approaches fail, especially in networks with non-linearities and saturation. The method computes the contribution of each input feature by back-propagating the differences in activation, rather than the gradient, thereby providing a more precise and robust understanding of feature importance, particularly in complex models with high non-linearity.

However, many of the above studies cannot be directly applied to SR neural networks. For example, to avoid the problem of gradient saturation, Gu and Dong [11] proposed local attribution maps (LAM) to interpret the low-resolution input influence on super-solved output in super-resolution tasks, which uses integral gradient. LAM takes the black image, without any texture, as the baseline and interpolates 50 images to fully capture the input texture influence. This granular analysis allows for the creation of maps that visually represent the contribution of different regions of the input image to the super-resolution process. This method helps in understanding which specific features of the input are most significant for the enhanced output, providing valuable insights into the network's functioning but with low efficiency. At the same time, when increasing the size of the local area, the region of interest loses its strong ability to interpret the model from the local view.

III. Method

To start with, we clarify the local and global explanations in our work. Typically, local explanations provide insights into the decision-making process of a model for a specific instance or prediction. Instead, the global explanations offer an understanding of the overall behavior and decision logic of a model, across all possible inputs [15]. However, we are more concerned with which low-resolution input pixels affect the reconstruction from the local view as well as the reconstruction from the global view in the SR task. More specifically, in our research, the local explanation aims at determining which pixels are responsible for the reconstruction of the SR image, e.g., which pixels contribute to the reconstruction of the region of interest (ROI). On the other hand, the global explanation determines how each part of the region affects the reconstructed SR output. Inspired by the previous studies [10], [11], herein we propose a novel gradient-based interpreting

method, dubbed glocal attribution map (GL-AM), to interpret the model for SR deep learning methods.

A. Local Attribution Map

We define an SR network F, mapping from the $\mathbb{R}^{h\times w} \mapsto \mathbb{R}^{sh\times sw}$, with a scaling factor s. In our analysis, we interpret the functionality of F by recognizing the presence of texture, a specific feature in local areas of the reconstructed SR image rather than focusing on pixel intensity. By contrast to using pixel intensity, focusing on the texture helps to understand the reconstructed high-resolution outputs which are perceptually more accurate and visually pleasing. This is also the key mission of interpreting the SR model.

In our work, we start by applying the same method as shown in [11] to calculate the texture of an ROI having a window size of $l \times l$, denoted as T_{ROI}. We calculate the gradient of the feature map of each network layer in relation to the original inputs, as shown by the blue arrows in Fig. (2), and as formulated in equation (1). In this way, GL-AM could focus more on the sensitivity of each layer's output to the model input, which can more intuitively understand the importance of different input pixels to model output.

$$grad^{local}_{i\to in} = g\left(T_{ROI}\right)_{i\to in} \tag{1}$$

In equation (1), $g()_{i\to in}$ is the function for calculating the gradient of the layer i with respect to the input; and $grad^{local}_{i\to in} \in \mathbb{R}^{c\times h\times w}$, c, h, w denote the number of the channels, the height and the width of the ROI, respectively. To better demonstrate the detailed subtle changes of the SR neural network model, we introduce consensus and variation to the attention of the model. The main purpose of consensus is to aggregate the results of multiple gradient calculations to obtain a comprehensive gradient effect that emphasizes features that appear consistently in all gradient maps. At the same time, variation highlights the characteristics of changes between different gradient maps, helping to identify areas where the model is sensitive to input changes.

Then, we calculated the weights of $grad^{local}_{i\to in}$, $W^{local}_{i\to in} \in \mathbb{R}^{c}$. So the highlighted $grad^{local}_{i\to in}$ is denoted as $grad^{local}_{key_i} \in \mathbb{R}^{c,h,w}$, and $grad^{local}_{key_i} = grad^{local}_{i\to in} \times W^{local}_{i\to in}$. Finally, the local attribution map is calculated by equation (2).

$$G_{local} = \sum_{i=0}^{n} grad^{local}_{key_i} + \sum_{i=1}^{n-1}\left(grad^{local}_{key_i} - grad^{local}_{key_{i-1}}\right) \tag{2}$$

The grad-cam [10] utilizes only the product of the last layer's output and the weights of its gradient to emphasize the key areas. However, this method cannot account for how inputs are processed within the model because it only uses the gradient from the last layer. To address this, GL-AM extends the analysis to incorporate the weighted gradients from each layer of the network. These weighted gradients are combined across the various network layers to better capture the cumulative impact of the different model parameters on the output dimensions.

As delineated in equation (2), the first term represents the consensus among all layers, while the second term captures the cumulative variance of each layer. It is crucial to recognize that the size of gradient maps can vary significantly between different layers. Therefore, we employ the bicubic interpolation method to resize all normalized gradients, $grad^{local}_{key_i}$, to match the output dimensions. Subsequently, we normalize the $grad^{local}_{key_i}$ values to the range of $[0, 1]$. The pixel values of G_{local} represent the impact of all pixels on the construction of the ROI; whereas the darker pixels (higher intensity) indicate a stronger influence w.r.t. the SR results.

B. Global Attribution Map

We have implemented different computational methods to generate the global attribution map. In contrast to the local attribution map, we use gradient maps from subsequent layers relative to earlier ones, which are applicable to nearly all neural network architectures and more efficient as well, as shown in the red arrow lines in equation (2). Similarly to the local explanation, we first compute the global gradient, $grad^{global}_{j\to i}, j = i + 1$. However, for the global explanation, we calculate the gradient of the output feature map for each network layer in relation to its adjacent previous layer, as detailed in equation (3).

$$grad^{global}_{j\to i} = g\left(T_{Global}\right)_{j\to i} \tag{3}$$

In equation (3), T_{Global} computes the texture of the feature map from layer $j \to i$, and the subscripts $j = i + 1$. And $grad^{global}_{j\to i} \in \mathbb{R}^{C,H,W}$, here, the capital C, H, W denotes the channel number and the size during the global reconstruction. Meanwhile, we still calculate $W^{global}_{j\to i} \in \mathbb{R}^{C}$, the weight of $grad^{global}_{j\to i}$, to help GL-AM capture the highlight of gradient from the global view. This is denoted as $grad^{global}_{key_j} \in \mathbb{R}^{C}$, and $grad^{global}_{key_j} = grad^{global}_{j\to i} \times W^{global}_{j\to i}$. However, the difference from the local explanation here is that we not only use the gradient but also take the feature map of the layer j into consideration. In this way, we could better understand the real effect from input to output, instead of only exploring the contributing pixels, as depicted in equation (4).

$$fea^{j\to i}_{global} = fea_j \times grad^{global}_{key_j} \tag{4}$$

In (4), fea_j is the feature map of layer j. Next, following the same idea to distinguish the consensus as well as the variance in each layer, we obtain the global attribution map, G_{global} in (5).

$$G_{global} = \sum_{i=0}^{n} fea^{j\to i}_{global} + \sum_{j=1}^{n-1}\left(fea^{j}_{global} - fea^{j-1}_{global}\right) \tag{5}$$

IV. Experiments

A. Settings

In our work, we collect images that are challenging for SR networks as the test set for the following analysis from

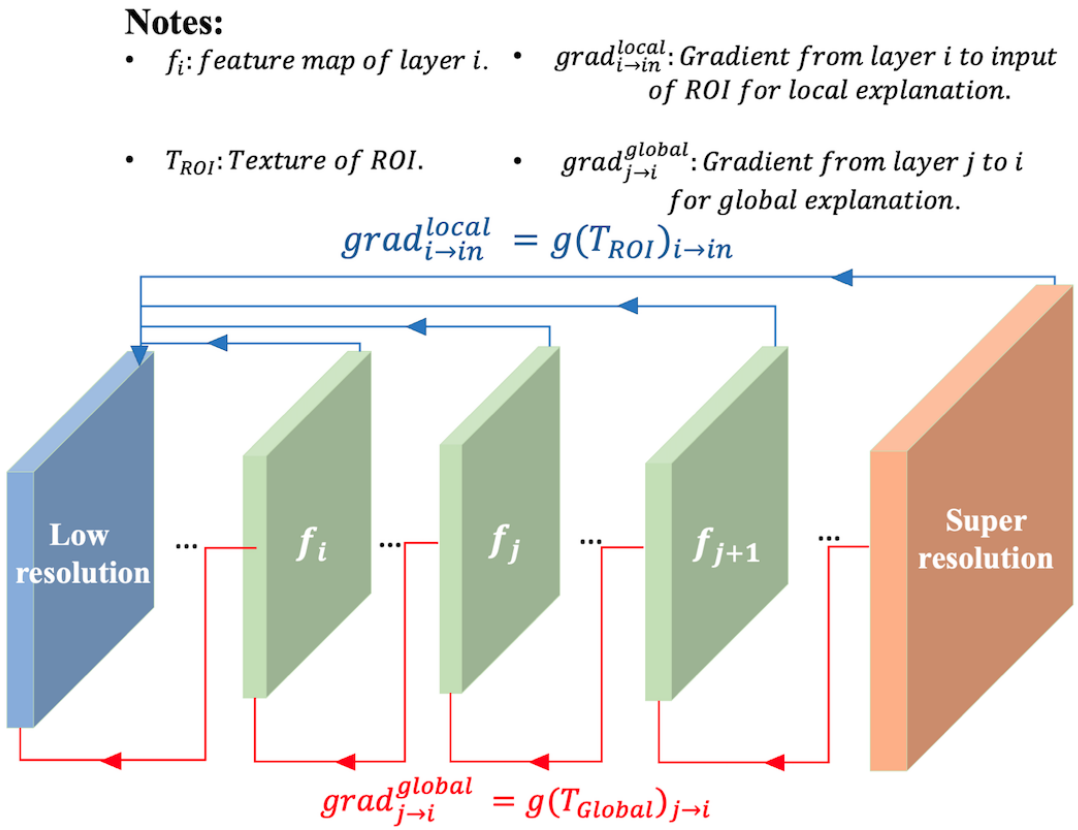

Fig. 2. Overview of how to calculate the gradients map of GL-AM. At the top of this figure are the notes. The blue arrows indicate how we calculate the gradient maps for each layer relative to the inputs of the ROI, and the red arrows represent the method we calculate the gradient maps of each subsequent layer relative to the previous layer for global reconstruction.

DIV2K [24] and Urban100 [25]. The images we chose exhibit a low average PSNR score and significant discrepancies in performance across various SR networks. The size of high-resolution images is 256 × 256, and generate low-resolution (input image) images with the size of 64 × 64 by bicubic interpolation. Thus, the up-scaling factor in our work is 4. When selecting metrics for quantitative evaluation for the SR neural network, we follow the suggestion of Gu et al. [22] and employ both PSNR and MSE as perceptual similarity metrics and quantitative metrics, respectively. We apply GL-AM to different SR neural networks, e.g., CARN, Residual Dense Block Network (RRDBNet) [17], Residual Non-local Attention Networks (RNAN) [19], Residual Channel Attention Networks (RCAN) [20], and Second-order Attention Network (SAN) [18]. With the exception of CARN and RRDBNet, the other models all use the attention mechanism.

B. Local Attribution Map

1) General Interpretation: Fig. (3) shows the local interpretation results of different models, CARN, RRDBNet, SAN, RNAN, and RCAN. Take CARN as an example in Fig. (3), the red pixels in the local attribution map represent the attributed pixels that influence the reconstruction of the ROI, delineated by the black box - darker pixels indicate a stronger influence w.r.t. the SR results. The Sum is the pure local attribution map without overlapping the input image, and it is the combination of normalized 'Variation' and normalized 'Consensus'. At last, we show the diffusion index (DI), as well as the PSNR of the ROI. A larger DI indicates that more pixels are involved in the reconstruction.

In Fig. (3), the Sum of RRDBNet reveals that it possesses the largest receptive field among all models, due to its deep residual structure and a deeper structure - a key determinant of the receptive field range. Furthermore, RNAN, RCAN, and SAN have a smaller range of contributing pixels, and the pixels are nearly all within the ROI, compared to the CARN and RRDBNet during the local reconstruction. The red pixels have a darker color and a certain texture, indicating that the model using the attention mechanism has a stronger ability to capture, and make use of texture details, compared with models not having the attention mechanism, which means the attention mechanism helps SR neural network model find more effective information from a smaller receptive field.

Additionally, we employ the DI [11] to assess SR network performance. DI is a metric that measures the range and intensity of the pixels that have been taken into consideration by SR networks, during image reconstruction. This offers insights beyond MSE or PSNR. A high DI implies extensive pixel consideration, as seen with RRDBNet's broad receptive

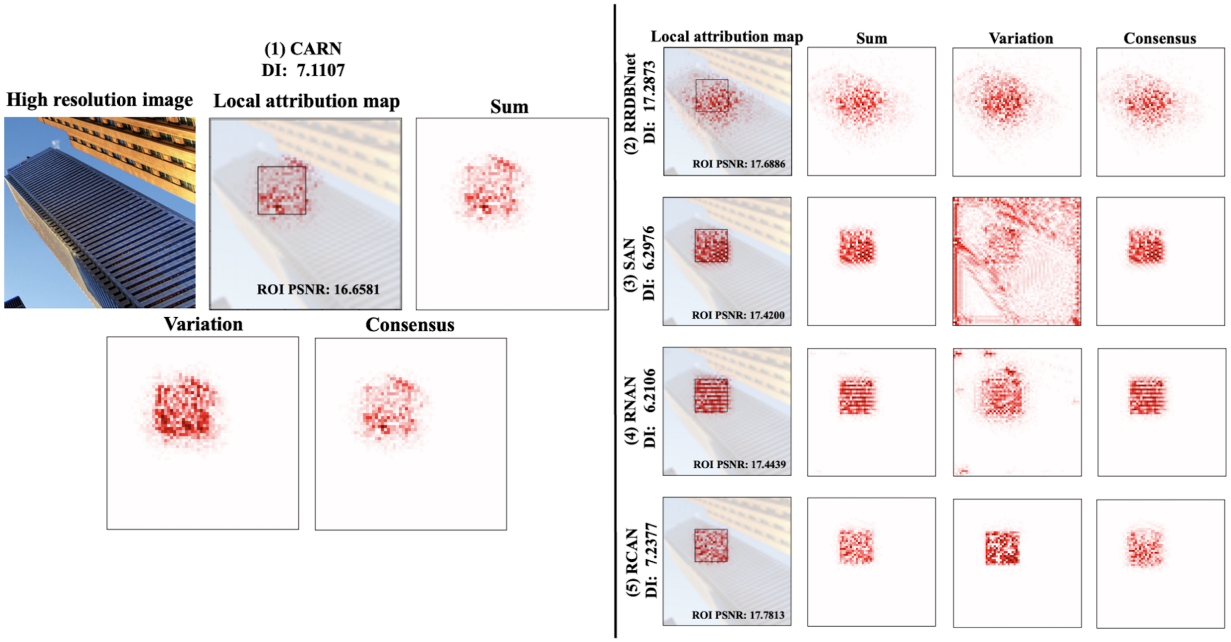

Fig. 3. Local attribution maps of 5 different models. The local attribution map is displayed by overlaying the Sum on the high-resolution image. Consensus emphasizes the highlighting of important pixels that the model consistently agrees on across all the layers by accumulating gradient maps, which reflects the model's learning and reinforcement of common features. Variation emphasizes the model's sensitivity to changes in inputs. Finally, a larger DI [11] indicates more pixels are involved during the reconstruction.

field (DI: 17.2873) since it does not have a deeper network structure. Attention-based models exhibit lower DI values even if compared to CARN which is a shallower network model. This verifies that the SR model using the attention mechanism can achieve equivalent or even better performance by using a smaller receptive field.

2) Insights of Attention Mechanism by Local Attribution Map: The Variation in Fig. (3) highlights the model's sensitivity to changes in inputs, showcasing the unique capability of each layer to explore different areas; in contrast, the Consensus emphasizes pixels that consistently appear across all gradient maps. It's important to note that the Sum is the normalized result of both Variation and Consensus. Therefore, in the Sum, some values with very small variations may not be displayed. Unlike RRDBNet and CARN, which do not employ attention mechanisms, RNAN implements a non-local attention mechanism to process information across the entire image, resulting in a wider range of contributed pixels, as evident in the corners of the Variation in Fig. (3). SAN utilizes a second-order attention mechanism, leveraging second-order feature statistics to capture more complex interactions between features. This approach explains why the Variation of SAN is particularly focused on regions rich in texture details, such as the main body of skyscrapers, which provide extensive information. Simultaneously, RCAN, employing feature recalibration, achieves a more stable attribution map from both the Variation and Consensus aspects, and it records the highest PSNR. In summary, in SR tasks, it is essential not only to utilize information from a broader range but also to assign different weights to various types of information.

C. Global Attribution Map

Fig. (4) illustrates the global attribution maps of different models. According to the results, CARN and RRDBNet struggle to capture hard areas, notably the main body of skyscrapers with intensive textures, indicating that these models do not utilize the input data appropriately. CARN allocates more attention to skyscraper silhouettes, suggesting it struggles to capture fine texture details due to its shallow network. Additionally, CARN ignores the hard areas with intensive textures, specifically the top of the skyscraper. RRDBNet, with its deeper structure, overly focuses on detailed textures and pays equal attention to areas regardless of texture intensity, thereby neglecting key aspects. This results in RRDBNet having the highest MSE and lowest PSNR among the models, as shown in Table (I). Notably, the global attribution map of models using attention mechanisms shows that these mechanisms help in capturing the silhouette of the building as well as much of the texture details. From Fig. (4) RCAN, we observe that RCAN's global attribution map shows attention to both intensively textured areas and edges of major objects, compared to Fig. (4) RNAN and Fig. (4) SAN. This is due to RCAN's use of an attention mechanism, which selectively weights various feature maps to suppress irrelevant information and recalibrates features simultaneously, thus allowing it to utilize input information more appropriately and effectively. RNAN employs the non-local attention mechanism to capture long-range dependencies in images, which allows for a distinct and clear representation of edges and texture features. This is also why RNAN's global attribution map shows that it

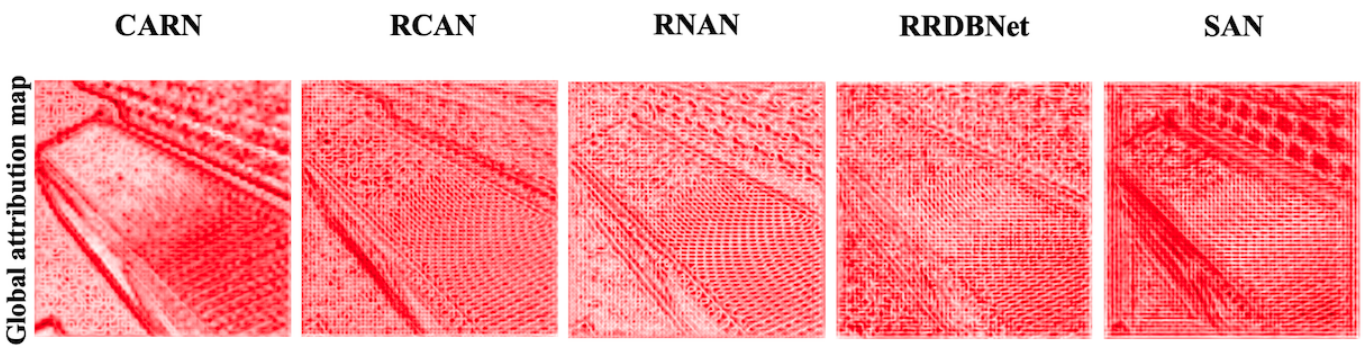

Fig. 4. Global attribution map of different models.

equally allocates attention across the edges and texture areas of skyscrapers, compared to RCAN or SAN. On the other hand, SAN employs second-order attention and reveals significantly more details by capturing more complex and subtle features. Yet, overemphasis on attention mechanisms can also lead to negative effects, especially during global reconstruction, e.g., aliasing and noise, which results in a higher MSE (0.0892) for SAN—the second-largest value after RRDBNet.

TABLE I
GLOBAL STATISTICS, PSNR AND MSE

Matrix \ Model	CARN	RCAN	RNAN	RRDBNet	SAN
PSNR	10.5000	10.5995	10.6625	10.4332	10.4938
MSE	0.0891	0.0871	0.0858	0.0905	0.0892

D. Complementary Findings

1) Attention Mechanism: Based on the analysis of GL-AM in (IV-B) and (IV-C), the PSNR of global reconstruction from different models is obviously lower than the ROI's, which means the SR model performs better in local reconstruction, especially with the help of attention mechanism. However, attention mechanisms don't always perform well. For instance, SAN, using a second-order attention mechanism, suffers from the artifact at the edge of the global attribution map when allocating the model's attention to different areas, as shown in Fig. (4) SAN. However, the SR model with the help of feature re-calibration - e.g., RCAN - can better emphasize the features of the hard areas and assign different weights to the features of simple areas to distinguish them. That is the reason why the global attribution map of RCAN appears more clearer details from the building than the global attribution map of RNAN and SAN in Fig. (4).

2) Comparison between GL-AM and LAM: LAM, which relies on the integral gradient method, necessitates the generation of 50 interpolated images to adequately capture the necessary information. This requirement leads to the execution of the SR model 50 times, significantly increasing the computational load, especially for models with deeper structures. In contrast, GL-AM efficiently obtains the attribution map by running the SR model just once, utilizing gradient calculations at each layer.

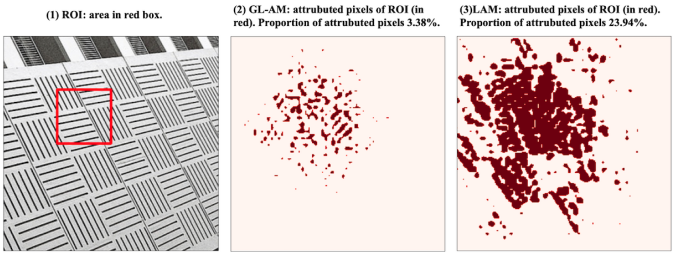

Fig. 5. Comparison between GL-AM and LAM from local view. LAM is oversensitive to input. LAM may become ineffective when increasing ROI.

In Fig. (5), we illustrate this comparison with attribution maps from both GL-AM and LAM, viewed locally. The ROI, highlighted in a red box in Fig. (5) (1), measures 64×64 pixels in high resolution. To focus on attributed pixels, we employ a threshold method where we exclude background areas by using the average pixel value of the entire image as a threshold. Since the background usually displays visual uniformity, pixel values in these areas are densely clustered with low standard deviation. Consequently, only pixels with values exceeding the overall mean are included in the attribution map.

Fig. (5) (2) and (3) display the attribution maps generated by GL-AM and LAM, respectively. In GL-AM, attributed pixels constitute just 3.38% of the total pixel count, focusing on essential details without overwhelming the visual representation. Conversely, LAM's attribution accounts for 23.94% of pixels, with a dominance of red pixels that can detract from the visual clarity and accuracy. This excessive sensitivity in LAM's approach, particularly noticeable as the ROI size increases, can lead to an overemphasis on texture-based gradients, potentially misleading in detailed image analysis

V. CONCLUSIONS

In this paper, we propose a useful global-local attribution map (GL-AM), to visualize and understand principles of SR neural networks. GL-AM explores which input pixels contribute to the SR output. At the same time, GL-AM

gives us some possible hints to improve the SR networks or other low-level vision tasks. For instance, it is better to use feature re-calibration when using the attention mechanism. Compared to LAM, GL-AM is more computationally efficient and still effective with the expansion of the ROI. However, GL-AM is still a kind of preliminary results, which need to be further improved. For instance, on the one hand, there are many attributed pixels, but with low values in the local attribution map. It is difficult to distinguish them from other non-contributing pixels, e.g., when comparing the difference between GL-AM and LAM. On the other hand, we need to propose quantitative indicators to help us better explore and understand the SR model, such as a tool similar to the diffusion index in LAM.

VI. Acknowledgments

This project has been supported by European Union PNRR Funding under Italian DM 352/2022.

REFERENCES

[1] S. Letzgus, P. Wagner, J. Lederer, W. Samek, K.-R. Müller, and G. Montavon, "Toward Explainable Artificial Intelligence for Regression Models: A methodological perspective," IEEE Signal Processing Magazine, vol. 39, no. 4, pp. 40-58, 2022. DOI: 10.1109/MSP.2022.3153277

[2] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a GAN to learn how to do image degradation first," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 185-200.

[3] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 217-233.

[4] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 517-532.

[5] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295-307, 2015.

[6] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," Advances in neural information processing systems, vol. 30, 2017.

[7] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," arXiv preprint arXiv:1312.6034, 2013.

[8] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in International conference on machine learning, 2017, pp. 3319-3328.

[9] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," PloS one, vol. 10, no. 7, pp. e0130140, 2015.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618-626.

[11] J. Gu and C. 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.

[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921-2929.

[13] E. Strumbelj and I. Kononenko, "An efficient explanation of individual classifications using game theory," The Journal of Machine Learning Research, vol. 11, pp. 1-18, 2010.

[14] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in International Conference on Machine Learning, 2017, pp. 3145-3153.

[15] A. M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney, "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review," Applied Sciences, vol. 11, no. 11, pp. 5088, 2021, MDPI.

[16] N. Ahn, B. Kang, and K.-A. 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.

[17] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0-0.

[18] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 624-632.

[19] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," arXiv preprint arXiv:1903.10082, 2019.

[20] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. 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.

[21] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, and others, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information fusion, vol. 58, pp. 82-115, 2020, Elsevier.

[22] J. Gu, H. Cai, H. Chen, X. Ye, J. S. Ren, and C. Dong, "Pipal: a large-scale image quality assessment dataset for perceptual image restoration," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI, 2020, pp. 633-651, Springer.

[23] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," Advances in neural information processing systems, vol. 31, 2018.

[24] E. Agustsson and R. 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, 2017, pp. 126-135.

[25] J.-B. Huang, A. Singh, and N. 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.