ADD: Attribution-Driven Data Augmentation Framework for Boosting Image Super-Resolution

Supplementary Material

1. Related Works

1.1. Image Super-Resolution

Image super-resolution (SR) stands as a pivotal technology in computer vision [16, 27] and image processing. Since the groundbreaking work of SRCNN [4], a multitude of convolutional neural network (CNN) based methods have emerged, including the integration of residual blocks [9, 26], dense blocks [17, 21, 28], and other architectural elements [2, 8, 12]. Some strategies incorporate attention mechanisms, encompassing channel attention [1], non-local attention [15], and adaptive path aggregation. Recently, a series of transformer-based networks [5, 6, 13] have been proposed and have made significant progress in SR. In contrast to the above methods, we propose ADD to boost the performance of existing SR approaches without increasing the inference time or changing architectures.

1.2. Data Augmentation in High-level Vision

Traditional *vanilla* DA strategies in high-level vision tasks include geometric transformations, color transformations, intensity transformation (*e.g.*, Cutout [3], Random Erasing [29]), Mixed-based strategies (*e.g.*, Mixup [25], CutMix [24]), etc. However, these *vanilla* methods are hardly focused on important regions and generate labels that match semantics. A series of *saliency-based* methods have been proposed and have become a hot research topic in high-level vision tasks. SaliencyMix [19] cuts the maximum saliency region and pastes it to the corresponding region in another image. AttentiveMix [20] further divides the image into blocks and selects the k blocks with the highest saliency to paste onto another image. Puzzlemix[10] and Co-Mixup [11] propose combinatorial optimization strategies to find optimal mixup masks by maximizing saliency information. Auto-Mix [14] simplifies the calculation of saliency information and adaptively generates mixed samples. The *saliency-based* DA methods have gained much popularity in the high-level computer vision community, surpassing *vanilla* methods and alleviating the information loss problem, which motivates us to propose *saliency-based* DA methods in low-level vision tasks.

1.3. Data Augmentation in Low-level Vision

Currently, the DA strategies in low-level visual tasks are still limited to *vanilla* methods. An early pioneering work [18] proposed seven methods, including geometric transformation DA approaches (*i.e.*, rotation and flipping), to improve the performance of SR. [7] introduces Mixup and demonstrates its ability to alleviate overfitting problems in SR models. [23] makes a comprehensive analysis of existing DA strategies applied to the single image SR task and proposes Cutblur. CutMIB [22] further transferred Cutblur to the light-field super-resolution and effectively improved network performance. However, these *vanilla* methods have information loss issues similar to high-level vision tasks, making it easy to lose important areas during the augmentation process. In this paper, we introduce saliency into DA in low-level tasks by designing a new attribution analysis approach and further proposing new *saliency-based* DA methods ADD and ADD+.

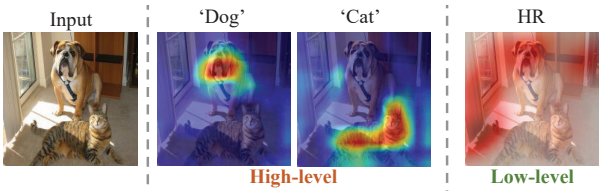

Figure 1. Differences in saliency map of high-level and low-level vision tasks.

2. Differences and challenges in High-level and Low-level Vision Tasks

The disparity between saliency methods for high-level and low-level vision tasks is substantial. As shown in Figure 1, salient regions in high-level tasks are associated with semantic concepts such as "cats" and "dogs," whereas low-level tasks emphasize edge textures and other fundamental features. Directly transferring saliency-based DA methods from high-level to low-level vision tasks poses several critical challenges.

Task objective mismatch. High-level vision tasks prioritize semantic information, while low-level tasks focus on pixel-level details. For instance, in object detection and image classification, saliency methods highlight significant semantic regions to identify objects and scenes using high-level features. Conversely, low-level tasks like image denoising and super-resolution require the processing of local details and textures through precise pixel-level operations. Employing high-level saliency methods in low-level tasks can overlook essential image details and textures, hindering the achievement of desired detail restoration.

Feature requirement inconsistency. High-level vision

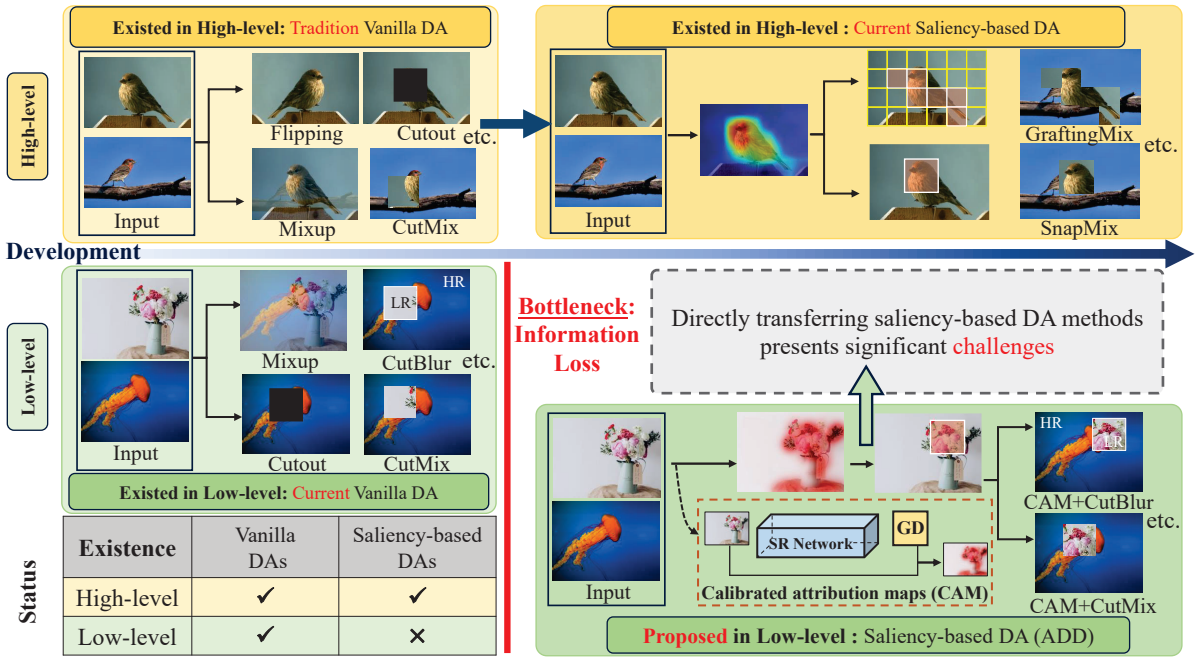

Figure 2. The critical bottleneck of DA methods in low-level vision tasks is **information loss** problems. Directly transferring the saliency-based DA methods from high-level tasks is challenging, and we propose ADD to tackle it.

tasks depend on deep semantic features, whereas low-level tasks rely on shallow detail features. Saliency methods in high-level tasks typically extract deep features to identify key regions, but low-level tasks need to retain and restore image details using shallow features. Directly transferring these methods results in a mismatch in feature extraction layers, which fails to capture the detailed information required for low-level tasks, thereby compromising image quality.

Data characteristics difference. Data in high-level vision tasks are rich in semantic information, whereas data in low-level tasks focus on detail and noise processing. High-level tasks often include clear semantic annotations, enabling saliency methods to effectively highlight target regions. In contrast, low-level tasks primarily deal with local features and noise. Using saliency methods from high-level tasks can lead to over-smoothing of details or inadequate noise handling, as these methods are not optimized for the characteristics of low-level vision data.

Method design inappropriateness. High-level vision tasks require methods optimized for semantic information, while low-level tasks need methods optimized for details. Saliency methods for high-level tasks often involve complex context modeling and global feature extraction, which are effective for handling global image information. However, low-level tasks require high-resolution feature maps and fine convolution operations. From a model design perspective, saliency methods tailored for high-level tasks are inherently unsuitable for low-level tasks, as their reliance on global context modeling and coarse feature extraction fails to preserve the fine-grained details critical to low-level vision applications.

3. Information Loss in High-level and Low-level Vision Tasks

High-level vision tasks. Data augmentation (DA), especially cropping, often results in the loss of critical semantic information or essential image regions, which can undermine the representativeness of training samples. This issue is particularly pronounced in high-level vision tasks such as image classification, object detection, and image segmentation. For instance, in image classification, cropping an image may exclude crucial object parts, leading to incomplete semantic information. If the head of a cat is cropped out, the remaining image content may fail to convey the semantic category *cat*, increasing the likelihood of misclassification. To address this, researchers have developed *saliency-based* DA strategies leveraging class activation maps and attention mechanisms to emphasize salient regions, thereby minimizing information loss during training.

Low-level vision tasks. In low-level vision tasks, the issue of information loss pertains to the degradation of fine-grained details during DA, as outlined in the main paper

Algorithm 1 CAM

Input: $\mathcal{I}^{LR} \in R^{H \times W \times C}$: Low-resolution imgaes
Input: $\mathcal{F}$: Pretrained SR model
Output: $\mathcal{I}_s^{LR} \in R^{H \times W \times C}$: Saliency maps of input images

$\gamma_{pb}(a) \leftarrow \omega(\sigma - \alpha\sigma) \otimes \mathcal{I}^{LR}$: Generate blurred LR images $s.t. a = 0.05, 0.1, 0.15, ..., 0.95$

$\mathcal{I}_{pb}^{SR} \leftarrow \mathcal{F}(\gamma_{pb}(a_i))$: Input blurred images to obtain reconstructed images

$\boldsymbol{G}_{sum} \leftarrow \boldsymbol{GD}(\mathcal{I}_{pb}^{SR})$: Convert to gradient scalar ▷ Eq. (2)

$\boldsymbol{G}_{pd}(i) \leftarrow \frac{\partial \boldsymbol{G}_{sum}}{\partial \gamma_{pb}((a_i)} d\alpha$: Obtain gradient through back-propagation ▷ Eq. (3)

$\gamma_{pb}(i) \leftarrow \gamma_{pb}(a_i) - \gamma_{pb}(a_{i+1})$: Calculate the difference between adjacent blurred images ▷ Eq. (4)

Initial: $\mathcal{L}_{OA} = \|\gamma_{pb}(0) - \mathcal{I}^{LR}\|_1$: The overall loss function

for $i = 1$ to k **do**

  $\phi_i^{CAM} \leftarrow (\gamma_{pb}(i) - \gamma_{pb}(i+1)) \times \boldsymbol{G}_{pd}(i)$: Obtain the initial gradient of the step i ▷ Eq. (4)

  ▷ Obtain threshold T_f for selecting the lowest absolute gradient value with a fraction of p_f ▷ Eq. (5)

  $M_f = \mathbf{1}_{\|\mathcal{I}_i^{LR}\| \leq T_f}$: Mask of the selected gradients ▷ Eq. (5)

  ▷ Achieve the limited range $[\alpha_{\text{min}}, \alpha_{\text{max}}]$ of the α_i in the step i

  ▷ Calculate the target loss $\mathcal{L}_{TG}$, current loss $\mathcal{L}_{CU}$, mask fraction loss $\mathcal{L}_{MF}$, and calculate the calibration factor δ:
  $\mathcal{L}_{TG} = \mathcal{L}_{OA} \times (1 - \frac{i}{k})$
  $\mathcal{L}_{CU} = \|\gamma(\frac{i}{k}) - \mathcal{I}^{LR}\|_1$
  $\mathcal{L}_{MF} = \|M_f \odot (\gamma(a_i) - \gamma(a_{\text{max}}))\|_1$ ▷ Eq. (6)
  $\delta = \frac{\mathcal{L}_{CU} - \mathcal{L}_{TG}}{\mathcal{L}_{MF}}$ ▷ Eq. (6)

  $\gamma_c(a_i) = \gamma_{pb}(a_i + \delta \times (a_{\text{max}} - a_i))$: Calibrate the blurred LR

  $\psi_i^{CAM} = \phi_i^{CAM} + (\gamma_c(a_i) - \gamma(a_i)) \times \phi_i^{CAM}$: Calibrate the gradient of the step i

end for

return $\mathcal{I}_s^{LR} = \sum_{i=0}^{k} \psi_i^{CAM}$ ▷ Eq. (7)

Sec. (1). This leads to augmented samples that may not effectively represent the original detailed features. For example, in super-resolution tasks, cropping can lead to the loss of intricate edge details and texture information, which are essential for accurate reconstruction. Unfortunately, on the one hand, existing saliency-based approaches in low-level vision tasks often struggle to identify relevant regions due to challenges like background noise. On the other hand, the inherent differences between high-level and low-level vision tasks (detailed in the supplementary Sec. (2)) render high-level saliency methods unsuitable for direct application to low-level tasks. We extend gradient-based attribution techniques and introduce two novel features, proposing a tailored attribution-driven DA method for low-level vision tasks, as shown in Fig. 2. This approach overcomes the limitations of existing methods, addressing the critical bottleneck of information loss and significantly advancing DA strategies in low-level vision applications.

4. Pseudocode for proposed CAM

Algorithm (1) contains a detailed pseudocode of the proposed CAM (main paper Sec. (3.3)). The saliency map $\mathcal{I}_s^{LR}$ obtained from pseudocode will be input into the proposed saliency-based DA methods ADD and the enhanced version ADD+ (main paper Sec. (3.4)).

References

[1] 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, CVPR*, pages 11065–11074, 2019. 1

[2] Xin Deng, Yutong Zhang, Mai Xu, Shuhang Gu, and Yiping Duan. Deep coupled feedback network for joint exposure fusion and image super-resolution. *IEEE Trans. Image Process.*, 30:3098–3112, 2021. 1

[3] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *ArXiv preprint*, abs/1708.04552, 2017. 1

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of the European Conference on Computer Vision, ECCV*, pages 184–199, 2014. 1

[5] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 12114–12124, 2022. 1

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations, ICLR*, 2021. 1

[7] Ruicheng Feng, Jinjin Gu, Yu Qiao, and Chao Dong. Suppressing model overfitting for image super-resolution networks. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition Workshops, CVPRW, pages 1964–1973, 2019. 1

[8] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *Proceedings of the European Conference on Computer Vision, ECCV*, pages 645–660, 2020. 1

[9] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1646–1654, 2016. 1

[10] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 5275–5285, 2020. 1

[11] Jang-Hyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. In *Proceedings of the International Conference on Learning Representations, ICLR*, 2021. 1

[12] 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/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5835–5843, 2017. 1

[13] 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, ICCV*, pages 1833–1844, 2021. 1

[14] Jihao Liu, Boxiao Liu, Hang Zhou, Hongsheng Li, and Yu Liu. Tokenmix: Rethinking image mixing for data augmentation in vision transformers. In *Proceedings of the European Conference on Computer Vision, ECCV*, pages 455–471, 2022. 1

[15] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3517–3526, 2021. 1

[16] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*, pages 4501–4510, 2017. 1

[17] Jialu Sui, Xianping Ma, Xiaokang Zhang, and Man-On Pun. GCRDN: global context-driven residual dense network for remote sensing image superresolution. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 16:4457–4468, 2023. 1

[18] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1865–1873, 2016. 1

[19] A. F. M. Shahab Uddin, Mst. Sirazam Monira, Wheemyung Shin, TaeChoong Chung, and Sung-Ho Bae. Saliencymix: A saliency guided data augmentation strategy for better regularization. In *Proceedings of the International Conference on Learning Representations, ICLR*, 2021. 1

[20] Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 3642–3646, 2020. 1

[21] 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 *Proceedings of the European Conference on Computer Vision Workshops, ECCVW*, pages 63–79, 2018. 1

[22] Zeyu Xiao, Yutong Liu, Ruisheng Gao, and Zhiwei Xiong. Cutmib: Boosting light field super-resolution via multi-view image blending. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1672–1682, 2023. 1

[23] Jaejun Yoo, Namhyuk Ahn, and Kyung-Ah Sohn. Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 8372–8381, 2020. 1

[24] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*, pages 6022–6031. IEEE, 2019. 1

[25] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations, ICLR*, 2018. 1

[26] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6360–6376, 2022. 1

[27] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2472–2481, 2018. 1

[28] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(7):2480–2495, 2021. 1

[29] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI*, pages 13001–13008, 2020. 1