NTIRE 2025 Image Super-Resolution ($\times$4) Challenge Factsheet
SMT: Scale-Aware Mamba-Transformer for Image Super-Resolution

Lu Zhao[1], Yuyi Zhang[1,2], Pengyu Yan[1,2], Jiawei Hu[1], Pengwei Liu[1], Fengjun Guo[1]
[1]Intsig Information Co., Ltd., Shanghai, China
[2]South China University of Technology, Guangzhou, China
zlcossiel@gmail.com, yuyi.zhang11@foxmail.com, smithpy6@gmail.com,
jiawei_hu@intsig.net, pengwei_liu@intsig.net, fengjun_guo@intsig.net

1. Team details

- Team name
  BBox
- Team leader name
  Lu Zhao
- Team leader address, phone number, and email
  Address: No. 1268 Wanrong Road, Jing'an District, Shanghai
  Phone: +86 198 2128 9756
  Email: zlcossiel@gmail.com
- Rest of the team members
  Lu Zhao[1], Yuyi Zhang[1,2], Pengyu Yan[1,2], Jiawei Hu[1], Pengwei Liu[1], Fengjun Guo[1]
- Team website URL (if any)
  None
- Affiliation
  [1]Intsig Information Co., Ltd.
  [2]South China University of Technology
- Affiliation of the team and/or team members with NTIRE 2025 sponsors (check the workshop website)
  None
- User names and entries on the NTIRE 2025 Codalab competitions (development/validation and testing phases)
  developmentg phase: User name: ZZXF; entries: 2
  testing phase: User name: ZZXF; entries: 3
- Best scoring entries of the team during the testing phase
  31.9657
- Link to the codes/executables of the solution(s)
  https://github.com/ZZXF11/BBox_Solution_NTIRE2025_ImageSR_x4

2. Method details

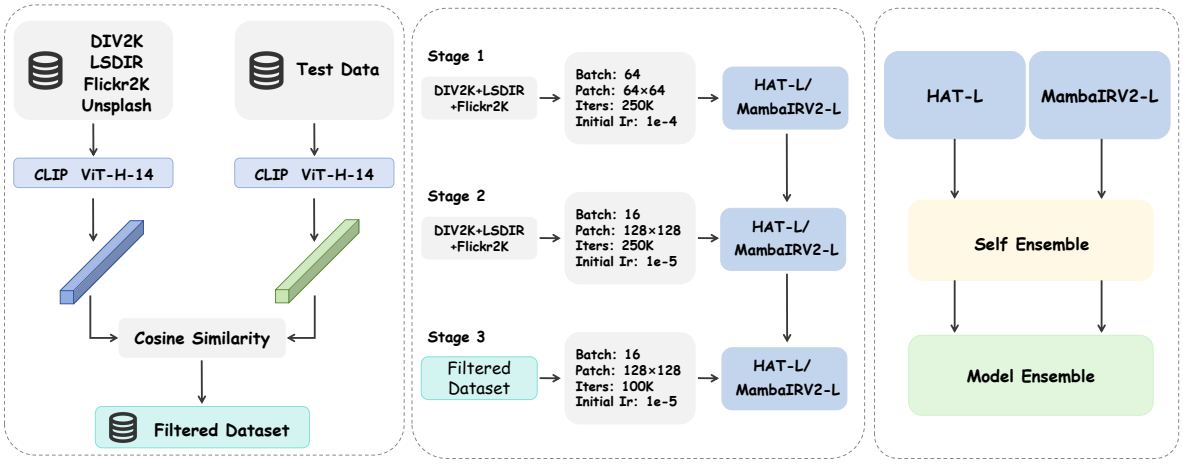
2.1. Description

The solution proposed by the BBox is illustrated in Fig. 1. Transformer-based models have consistently demonstrated remarkable performance in the field of super-resolution, as exemplified by methods such as HAT [2], DAT [3], and SwinIR [8]. Recently, numerous studies have shown that Mamba architectures can also achieve impressive results in this domain, as evidenced by models such as MambaIR [5], MambaIRv2 [6], and S^{3}Mamba [10]. Notably, prior research indicates that Transformer-based models excel at modeling sequential relationships, while Mamba-based approaches are particularly adept at capturing long-range contextual information [4]. Both capabilities are crucial for pixel-dense super-resolution tasks, which require models to simultaneously capture spatial relationships between individual pixels and model their long-range contextual dependencies. Inspired by these complementary strengths, we adopt HAT-L and MambaIRv2-L as our foundation models, training them independently and employing a model ensemble strategy to effectively leverage their advantages, thereby achieving optimal performance.

2.2. Implementation Details

The training dataset comprises DIV2K, LSDIR, Flickr2K, and selected Unsplash [1] datasets. To augment the training data, we implement random flip and rotation strategies. For the HAT model, we initialize the network using pre-trained weights from the HAT-L model, which was previously trained on the ImageNet dataset. In contrast, we train the MambaIRv2-L model from scratch without pre-training.

Specifically, our training follows a three-stage progressive strategy with a multi-scale loss that computes losses at x2, x3, and x4 resolutions to enhance performance across different scales. (1) In the first stage, we train on the DIV2K, LSDIR, and Flickr2K datasets. We set the patch size to 64 and batch size to 64, conducting 250K iterations with an initial learning rate of 1×10^{-4}. During this stage, we minimize the L1 loss using the Adam optimizer [7]. (2) In the second stage, while maintaining the same training datasets, we increase the patch size to 128 and reduce the batch size to 16. We also switch from L1 loss to MSE loss. This stage comprises 250K iterations with a reduced initial



(a) Dataset Filtering Pipeline (b) Model Training Pipeline (c) Ensemble Pipeline

Figure 1. Team BBox.

learning rate of 1×10^{-5}. (3) In the final stage, to achieve superior model performance, we implement semantic selection using CLIP [9] features. We filter out 2,000 similar images from DIV2K, LSDIR, Flickr2K, and Unsplash datasets for further fine-tuning. During this stage, the learning rate is set to 1×10^{-5}, and the model is trained for an additional 100K iterations. The multi-step learning rate decay method is applied across all three training phases. For both MambaIRv2-L and HAT-L models, we repeat the aforementioned three training stages five times to ensure optimal convergence.

In the inference stage, we employ a self-ensemble strategy to enhance the performance of both HAT-L and MambaIRv2-L models. For MambaIRv2-L specifically, we utilize multiple sliding windows of varying sizes during inference and integrate these results. Finally, we implement a weighted fusion method [11] to generate outputs from the adaptive combination of HAT-L and MambaIRv2-L models.

References

[1] Unsplash full dataset 1.2.2. https://unsplash.com/data, 2024. Accessed: 2024-03-21. 1

[2] Xiangyu Chen, Xintao Wang, Wenlong Zhang, Xiangtao Kong, Yu Qiao, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, 2023. 1

[3] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12312–12321, 2023. 1

[4] Zheng Chen, Zongwei Wu, Eduard Zamfir, Kai Zhang, Yulun Zhang, Radu Timofte, Xiaokang Yang, Hongyuan Yu, Cheng Wan, Yuxin Hong, et al. Ntire 2024 challenge on image super-resolution (x4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6108–6132, 2024. 1

[5] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *ECCV*. Springer, 2024. 1

[6] Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. Mambairv2: Attentive state space restoration. In *CVPR*, 2025. 1

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[8] 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*, pages 1833–1844, 2021. 1

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[10] Peizhe Xia, Long Peng, Xin Di, Renjing Pei, Yang Wang, Yang Cao, and Zheng-Jun Zha. S^3mamba: Arbitrary-scale super-resolution via scaleable state space model. *arXiv preprint arXiv:2411.11906*, 2024. 1

[11] Yulun Zhang, Kai Zhang, Zheng Chen, Yawei Li, Radu Timofte, Junpei Zhang, Kexin Zhang, Rui Peng, Yanbiao Ma, Licheng Jia, et al. Ntire 2023 challenge on image super-resolution (x4): Methods and results. In *CVPRW*, 2023. 2