SAT: Selective Aggregation Transformer for Image Super-Resolution –Appendix–

Dinh Phu Tran
phutx2000@kaist.ac.kr

Thao Do
thaodo@kaist.ac.kr

Saad Wazir
saad.wazir@kaist.ac.kr

Seongah Kim
kimsa0322@kaist.ac.kr

Seon Kwon Kim
lukaskim@kaist.ac.kr

Daeyoung Kim
kimd@kaist.ac.kr

School of Computing, KAIST, Republic of Korea

A. Implementation Details

A.1. Training Details

For training the SAT model, we use the DF2K dataset, which combines DIV2K [9] and Flickr2K [14], as our training set. To ensure fair comparisons, we adopt the same training configurations as those employed in recent super-resolution (SR) studies [4, 6, 13]. Our model is optimized using the Adam optimizer with parameters set to ($\beta_1 = 0.9, \beta_2 = 0.99$) , a weight decay coefficient $\lambda = 0.0001$, and an initial learning rate of 2×10^{-4}. The $\times 2$ model is trained for 500K iterations. During training, the input patch size is fixed at 64×64, and a MultistepLR scheduler is applied to halve the learning rate at predefined iterations [250000, 400000, 450000, 475000]. The batch size is set to 32 for all training processes. To enhance robustness, the training data is augmented with random horizontal and vertical flips as well as random rotations of $90°$. For the $\times 3$ and $\times 4$ models, we apply finetuning based on the pre-trained $\times 2$ model to save time, training these models for only 250K iterations. The initial learning rate is set to 2×10^{-4}, and a MultistepLR scheduler is used to halve the learning rate at predefined iterations [100000, 150000, 200000, 225000, 240000]. We evaluate our method on five standard benchmark datasets: Set5 [2], Set14 [15], B100 [1], Urban100 [7], and Manga109 [11]. Additionally, the computational cost of all models presented in this paper is measured at an output resolution of 1280×640. For training the SAT-light model, only the DIV2K dataset is used, excluding Flickr2K dataset for a fair comparison with other recent works. All other training strategies remain consistent with those used for the SAT model. All training and testing experiments are conducted on four NVIDIA RTX PRO 6000 96GB GPUs.

A.2. Hyperparameters

Network Hyperparameters. For the standard model, SAT, we set the number of Residual Groups (NGs) to 8, each containing three Local Transformer Blocks (LTB) and three Selective Aggregation Transformer Blocks (SATB), except for the first three Residual Blocks, where we use two blocks for each type of Transformer. The channel dimension, number of attention heads, and the MLP ratio of SAT are set to 228, 6, and 2.0, respectively. Note that we use smaller channel dimension compared to PFT (228 vs. 240). The window size used in the Local Transformer Block is set to 8×32 (equivalent to a squared window size of 16×16 for a fair comparison with other window-based attention approaches). The channel scaling factor in our SAA module for Query and Key representations is set to 0.5, meaning that 50% of the number of channels is reduced for query and key projections. For SAT-light, we reduce the channel dimension from 228 to 40 to decrease the model's computational complexity and parameters. Other hyperparameters will remain the same as those in SAT.

Other Hyperparameters. Besides the network hyperparameters, we also report details of other hyperparameters for increasing the reproducibility of our proposed method. In this work, we set the hyperparameter for our Selective Aggregation Attention (SAA) as follows. The number of clusters K or the number of compressed Key and Value tokens is set to 3% of the token sequence length N of input features. Subsampling factor β for density estimation and separation is set to 4 to balance computational cost and performance, number of neighbors used for local density estimation m is set to $0.1K$; temperature τ in similarity-weighted aggregation is set to 0.1, all ϵ are set to 10^{-6} for numerical stability, or preventing division by zero.

Table A. Comparison on Model complexity and Running Time.

Scale	Method	Params	FLOPs	PSNR (Manga109)	Runtime
×2	HAT [4]	20.6M	5.81T	40.26	**377ms**
×2	ATD [16]	20.1M	6.07T	40.37	466ms
×2	IPG [13]	18.1M	5.35T	40.24	765ms
×2	PFT [10]	19.6M	5.03T	<u>40.49</u>	528ms
×2	SAT (Ours)	19.4M	3.64T	**40.70**	<u>394ms</u>
×3	HAT [4]	20.8M	2.58T	35.53	**209ms**
×3	ATD [16]	20.3M	2.69T	35.63	285ms
×3	IPG [13]	18.3M	2.39T	35.53	447ms
×3	PFT [10]	19.8M	2.23T	<u>35.67</u>	313ms
×3	SAT (Ours)	19.5M	1.63T	**35.87**	<u>231ms</u>
×4	HAT [4]	20.8M	1.45T	32.48	**192ms**
×4	ATD [16]	20.3M	1.52T	32.62	228ms
×4	IPG [13]	18.3M	1.30T	32.53	288ms
×4	PFT [10]	19.8M	1.26T	<u>32.63</u>	230ms
×4	SAT (Ours)	19.5M	0.94T	**32.85**	<u>207ms</u>

B. Comparison of Model Complexity and Inference Time

We compare the model complexity and inference time of our SAT model with several state-of-the-art SR methods, including HAT [4], IPG [13], ATD [16], and PFT [10]. In this experiment, the inference time for all models is measured on a single NVIDIA RTX PRO 6000 GPU at an output resolution of 512×512. As shown in Tab. A, the inference time of our SAT is comparable to existing methods. Our model takes more time than HAT but delivers much better reconstruction performance. Moreover, compared to more recent SR works such as IPG, ATD, and PFT, our proposed method shows lower computational complexity while achieving better SR quality. These results demonstrate that our SAT achieves a more favorable balance between computational complexity and the model's performance, further pushing the boundaries of the SR domain.

C. Additional Experiment and Qualitative Result

C.1. Perceptual Quality Comparison

The paper [3] demonstrates that a high PSNR (Peak Signal-to-Noise Ratio) does not always correspond to better visual quality. To better evaluate perceptual quality, we present LPIPS results, referencing [17], which compares our method with several state-of-the-art (SOTA) approaches. LPIPS is one of the most widely used metrics for super-resolution (SR) in real-world applications.

Unlike PSNR, LPIPS is more closely aligned with human perception. Tab. B shows that SAT achieves the best performance (i.e., the lowest value) across all datasets. These results further highlight the superiority of our method, even when assessed using a perceptual similarity metric like LPIPS.

C.2. Additional Ablation Study

We conduct additional ablation studies to understand our proposed method better. Some ablation studies do not affect the number of parameters or FLOPs, or the differences are minimal; therefore, we only report the performance for these experiments. All ablation studies, including those in the main paper and supplementary material, are performed on four NVIDIA H200 GPUs. Each experiment runs on two H200 GPUs to enhance speed.

Effects of feature norm restoration. Tab. C illustrates the impact of feature norm restoration on model performance across three benchmark datasets. The results show that integrating feature norm restoration consistently improves performance across all tested datasets. The observed gains of 0.03 to 0.07dB indicate that feature norm restoration helps maintain a consistent feature norm distribution after aggregation. Nonetheless, even without feature norm preservation, our primary architectural components are effective at retaining feature representations throughout the network, achieving competitive performance.

Effects of channel scaling factor. To evaluate the effect of the channel scaling factor r_c in our SAA module, we perform an ablation study on benchmark datasets, as detailed in Tab. D. r_c adjusts channel dimensions to reduce channel redundancy and computational complexity. At $r_c = 1.0$ (full channels), the model has the highest complexity (801K Params, 38.5G FLOPs) but yields moderate performance (e.g., 32.46dB on Set5). Reducing to $r_c = 0.8$ slightly lowers Params/FLOPs with minor PSNR drops (e.g., -0.05dB on Set5). Our choice of $r_c = 0.5$ achieves peak performance (32.48dB on Set5, +0.02dB over $r_c = 1.0$; 26.61dB on Urban100, +0.04dB; 31.09dB on Manga109, +0.02dB) while reducing complexity (763K Params, 36.4G FLOPs). Further reducing r_c to 0.3 significantly degrades performance in terms of PSNR. We select $r_c = 0.5$ as the default for an optimal balance between accuracy and efficiency in super-resolution.

Effects of subsampling strategy. Tab. E analyzes the impact of different subsampling strategies on model performance, highlighting significant variations that emphasize the importance of sample selection methods. Random subsampling performs worst across all datasets, indicating inadequate coverage of spatially substantial features. Grid-based subsampling shows slight improvements (Set5: 32.43dB, Urban100: 26.52dB, Manga109: 30.90dB), as it provides more structured spatial coverage, but it remains suboptimal. In contrast, our proposed stratified subsampling strategy outperforms all other methods across the benchmarks (Set5: 32.48dB, Urban100: 26.61dB, Manga109: 31.09dB), achieving particularly noteworthy gains of up to 0.18dB on Manga109 compared to random

Table B. LPIPS comparison with the recent methods. Top-2 results are marked with **bold** and underline. The result is lower is better.

Method	Scale	Params	FLOPs	Set5	Set14	B100	Urban100	Manga109
SwinIR [8]	×4	11.9M	0.76T	0.2079	0.2900	0.3295	0.2470	0.1531
CAT-A [5]		16.6M	1.27T	0.2042	0.2862	0.3241	0.2392	0.1502
HAT [4]		20.8M	1.45T	0.2047	0.2845	0.3243	0.2349	0.1478
IPG [13]		18.3M	1.30T	0.2048	0.2854	0.3262	0.2360	0.1490
ATD [16]		20.3M	1.52T	0.2045	0.2854	0.3231	0.2314	0.1461
PFT [10]		19.8M	1.26T	0.2026	0.2826	0.3227	0.2288	0.1451
SAT (Ours)		19.5M	0.94T	**0.2022**	**0.2800**	**0.3192**	**0.2287**	**0.1438**

Table C. Effects of Feature Norm Restoration.

Feature Norm Restoration	Set5	Urban100	Manga 109
w/o Feature Norm Restoration	32.45	26.54	31.06
w/ Feature Norm Restoration	**32.48**	**26.61**	**31.09**

Table D. Effects of channel scaling factor r_c.

r_c	Params	FLOPs	Set5	Urban100	Manga 109
$r_c = 1.0$	801K	38.5G	32.46	26.57	31.07
$r_c = 0.8$	786K	37.7G	32.41	26.55	31.02
$r_c = 0.5$	763K	36.4G	**32.48**	**26.61**	**31.09**
$r_c = 0.3$	748K	35.9G	32.37	25.43	30.94

Table E. Effects of subsampling strategy.

Sampling Strategy	Set5	Urban100	Manga 109
Random Subsampling	32.40	26.48	30.91
Grid-based Subsampling	32.43	26.52	30.90
Stratified Subsampling (Ours)	**32.48**	**26.61**	**31.09**

sampling. These results demonstrate that stratified subsampling effectively captures representative samples from the feature space, ensuring comprehensive coverage of both spatial and semantic information while maintaining computational efficiency.

Effects of subsampling factor. To assess the impact of the subsampling factor β for efficient density estimation and separation, we conduct an ablation on benchmark datasets, as shown in Tab. F. With $\beta = 2$, performance is lowest (32.31dB on Set5), as limited subsampling may miss optimal centers. Increasing to $\beta = 4$ yields significant gains (e.g., +0.17dB on Set5, +0.18dB on Urban100, +0.21dB on Manga109), improving density estimation and center diversity, and further doubling to $\beta = 8$ offers negligible or minor drops (-0.02dB on Set5 and Urban100, no change on Manga109). We select $\beta = 4$ as the default, achieving near-optimal PSNR with better efficiency than higher values, suitable for lightweight deployment.

Table F. Effects of subsampling factor β.

β	Set5	Urban100	Manga 109
$\beta = 2$	32.31	26.43	30.88
$\beta = 4$	**32.48**	**26.61**	**31.09**
$\beta = 8$	32.46	26.59	**31.09**

Table G. Effects of aggregation method.

Aggregation Method	Set5	Urban100	Manga 109
Uniform Average	32.41	26.53	30.99
No aggregation	32.32	26.37	30.79
Similarity-weighted (Ours)	**32.48**	**26.61**	**31.09**

Effects of aggregation method. Tab. G presents the results of an ablation study on aggregation methods. Our similarity-weighted aggregation consistently outperforms uniform averaging across all benchmarks, achieving improvements of 0.07 to 0.10dB. The most significant gains are observed in the Urban100 (+0.08dB) and Manga109 (+0.10dB) datasets, which feature structured content with clear semantic boundaries. In contrast, using cluster centers directly without aggregation results in the poorest performance, showing decreases of -0.09 to -0.20dB compared to the uniform average. This clearly demonstrates that aggregating information from **all cluster members** is crucial. The similarity weighting effectively reduces the influence of boundary tokens with ambiguous cluster membership, leading to more representative cluster prototypes.

Effects of number of neighbors for local density estimation. To explore the impact of number of neighbors m on our local density estimation, we conducted experiments on benchmark datasets, as illustrated in Tab. H. The parameter m defines the number of nearest neighbors used to compute local densities, which influences the quality of the cluster centers. We established a baseline performance with a fixed $m = 30$, achieving a peak signal-to-noise ratio (PSNR) of 32.40dB on the Set5 dataset. As we increased m to 50, we observed improvements of +0.03dB and +0.11dB, respectively, on the Set5 and Urban100 datasets, allowing for bet-

Table H. Effects of number of neighbors for local density estimation m.

m	Set5	Urban100	Manga109
$m = 30$	32.40	26.43	30.86
$m = 50$	<u>32.43</u>	<u>26.54</u>	<u>30.94</u>
$m = 100$	32.39	26.44	30.90
$m = 0.1K$	**32.48**	**26.61**	**31.09**

ter capture of neighborhood structures. At $m = 100$, performance declined, showing a decrease of -0.1dB on Urban100 compared to $m = 50$. Our adaptive choice of $m = 0.1K$ (K is set to 3% of the input sequence length) produced the best results, achieving 32.48dB on Set5 (+0.05dB compared to $m = 50$), 26.61dB on Urban100 (+0.07dB), and 31.09dB on Manga109 (+0.15dB). This approach dynamically adjusts to token distributions, leading to enhanced clustering. We have selected $m = 0.1K$ as the default due to its adaptive efficiency and superior performance.

C.3. Additional Qualitative Result

Visualization of pixel-wise absolute error between SR and HR images. We further analyze the pixel-wise absolute error between the super-resolution (SR) outputs and the ground-truth images. As illustrated in Fig. A, the reconstruction errors of existing methods, including PFT, are disproportionately concentrated in high-frequency regions such as edges and fine textures. Although PFT demonstrates competitive overall performance and performs better than the early approach RCAN, it still struggles to recover these challenging areas. In contrast, our method, SAT, considerably reduces these localized errors by effectively allocating computation to the critical high-frequency regions. This results in a more faithful structural restoration while maintaining lower computational complexity compared to PFT. We observe that our method struggles in challenging areas, indicating there is still room for improvement based on our approach and observations. We believe that our study can serve as a strong baseline and motivate further research in the future.

Visual comparison of SAT. To qualitatively assess the reconstruction performance of our SAT model compared to other state-of-the-art methods, we present visual examples in Fig. B and C. These comparisons highlight the advantages of our approach in restoring sharp edges and fine textures from severely degraded low-resolution inputs. By selectively focusing on critical regions, our model produces cleaner edges and achieves more accurate, visually pleasing reconstructions with low computational complexity.

Visualization of cluster center selection. We offer an enhanced visualization of the cluster center selection method applied to low-resolution input across various SAA (Selective Aggregation Attention) layers, specifically focusing on the final SAA layers of Residual Blocks 1, 3, 5, 7, and 8. Fig. D illustrates that earlier layers maintain a broad spatial coverage, while the deeper layers increasingly focus on semantically significant regions, such as edges and pattern features. This progressive adaptation highlights the content-aware nature of our DTA (Density-driven Token Aggregation) algorithm, enabling efficient compression without exhaustive spatial coverage. The visualization demonstrates that the selected centers provide sufficient diversity, allowing the attention mechanism to function effectively.

D. Theoretical Proof of Theorem 3.2

D.1. Statement

We first formalize the setting. All feature vectors $\mathbf{x} \in \mathbb{R}^d$ lie in a bounded domain $\|\mathbf{x}\| \le B$, and the similarity function $s(\mathbf{x}, \mathbf{y}) \in [0, 1]$ is L_s–Lipschitz in both arguments:

$$|s(\mathbf{x}, \mathbf{y}) - s(\mathbf{x}', \mathbf{y}')| \le L_s\big(\|\mathbf{x} - \mathbf{x}'\| + \|\mathbf{y} - \mathbf{y}'\|\big).$$

The empirical density estimator ρ_i is defined as the average similarity to the m nearest neighbors of $\mathbf{x}_i$ in feature space.

Theorem 3.2 (Approximation Quality). *Let* $\mathbf{O}^* = \mathit{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ *denote exact self-attention, and let* $\mathbf{O} = \mathit{SAA}(\mathbf{Q}, \mathbf{K}', \mathbf{V}')$ *denote Selective Aggregation Attention using density-driven token clustering and uniform subsampling. Assume:*

(i) *(Lipschitz density)* The density field $\rho(\mathbf{x})$ induced by s is L–Lipschitz:

$$|\rho(\mathbf{x}) - \rho(\mathbf{y})| \le L\|\mathbf{x} - \mathbf{y}\|.$$

(ii) *(Cluster separation)* There exists a ground-truth clustering $\{C_k\}$ such that

$$\min_{\mathbf{x} \in C_k, \mathbf{y} \in C_\ell} \|\mathbf{x} - \mathbf{y}\| \ge \epsilon \quad (k \ne \ell),$$

which ensures stability of m–NN neighborhoods.

(iii) *(Score margin)* Let $\gamma_i = \rho_i \delta_i$, where

$$\delta_i = \min_{j : \rho_j > \rho_i} (1 - s(\mathbf{x}_i, \mathbf{x}_j)).$$

The K highest scores satisfy a margin $\gamma_{(K)} - \gamma_{(K+1)} \ge \Delta_{\min} > 0$.

(iv) *(Within-cluster variation)* Each cluster C_k satisfies

$$\|\mathbf{x} - \mathbf{c}_k\| \le \sigma,$$

where $\mathbf{c}_k$ is the true cluster center.

(v) *(Subsampling size)* $S = \beta K$ with $\beta \ge 2$, and subsampling is uniform without replacement.

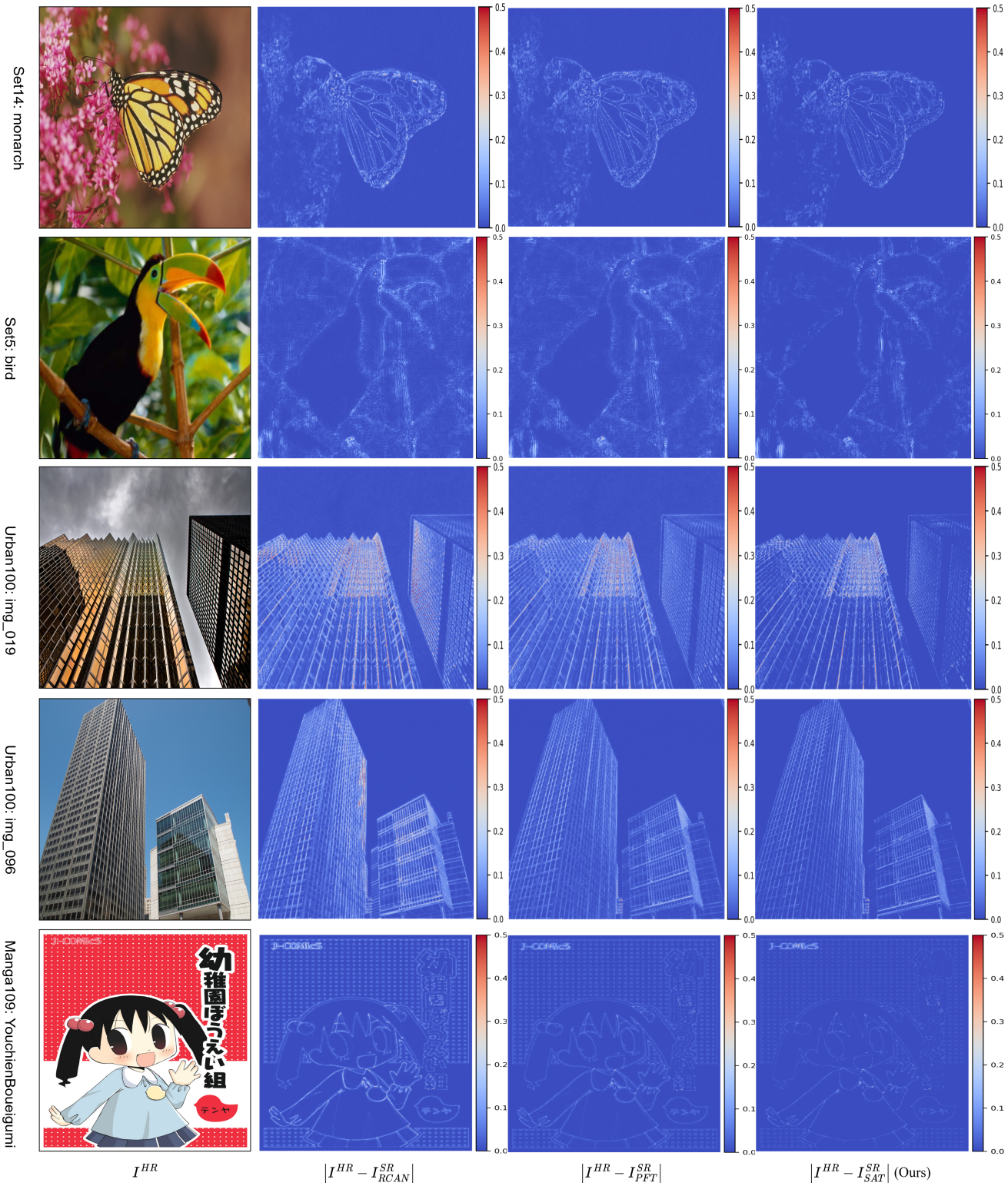

Figure A. Visualization of pixel-wise absolute error between SR and HR images across diverse samples from benchmark datasets, including Set5, Set14, Urban100, and Manga109. This visualization shows that our SAT can handle these challenging regions better than other methods, including PFT.

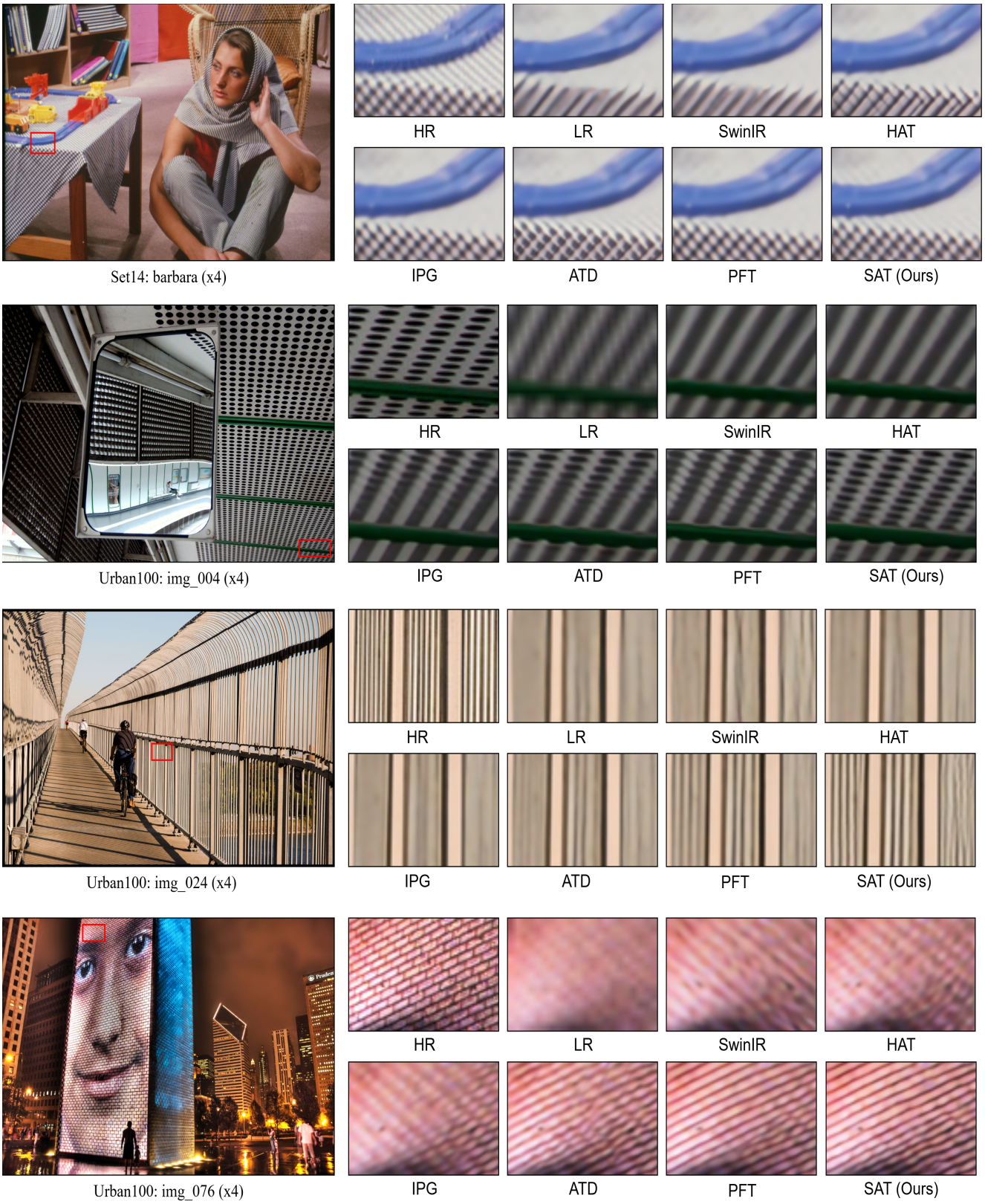

Figure B. Qualitative comparison of visual results between our SAT and other state-of-the-art super-resolution methods on the Set14 and Urban100 datasets.

Figure C. Qualitative comparison of visual results between our SAT and other state-of-the-art SR methods on Manga109 dataset.

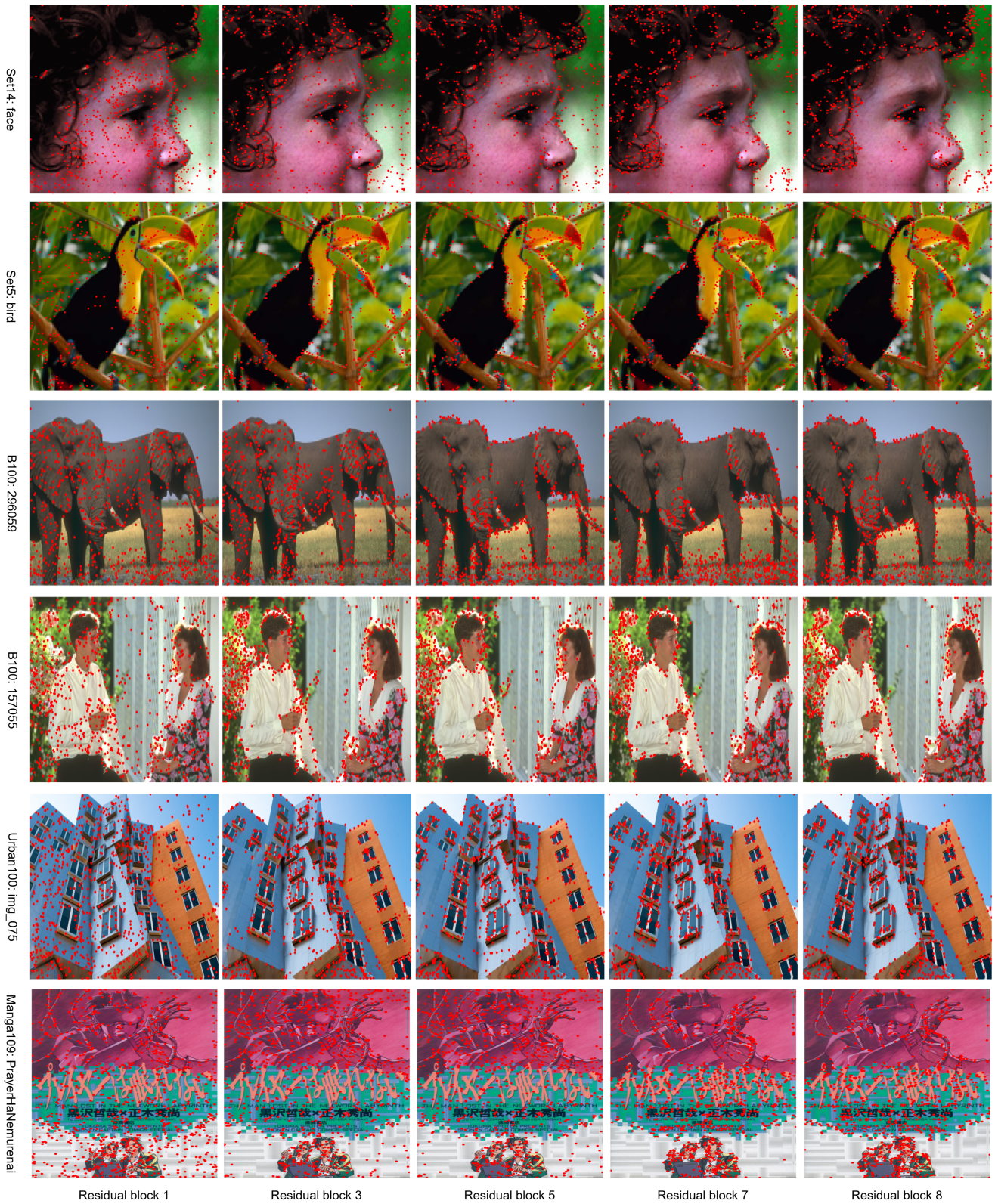

Figure D. Visualization on low-resolution input for cluster center selection (red points) across different network layers. Early layers maintain broad spatial coverage, while deeper layers increasingly concentrate on semantically salient regions such as edges and pattern features. This progressive adaptation enables efficient compression without exhaustive spatial coverage.

Then for some constants $C_1, C_2 > 0$, with probability at least $1-\delta$,

$$\|\mathbf{O} - \mathbf{O}^*\|_F \;\le\; C_1 L \sqrt{\frac{NK\log(\delta^{-1})}{S}} \;+\; C_2 \|\mathbf{V}\|_F \frac{K}{N}. \tag{1}$$

The first term corresponds to clustering error (density estimation + center approximation); the second term corresponds to attention compression.

D.2. Proof Overview

We decompose the error into a clustering part and an attention-compression part:

$$\|\mathbf{O} - \mathbf{O}^*\|_F \le \underbrace{\|\mathbf{O} - \mathbf{O}_{\text{exact}}\|_F}_{E_{\text{cluster}}} + \underbrace{\|\mathbf{O}_{\text{exact}} - \mathbf{O}^*\|_F}_{E_{\text{attn}}},$$

where $\mathbf{O}_{\text{exact}}$ uses the *true* top-K cluster centers but follows the same SAA aggregation rule. We bound the two terms separately.

D.3. Step 1: Density Estimation and Concentration

For each subsampled token $i \in \mathcal{S}$,

$$\rho_i = \frac{1}{m} \sum_{j \in \mathcal{N}_m^{\text{full}}(i)} s(\mathbf{x}_i, \mathbf{x}_j), \qquad \tilde{\rho}_i = \frac{1}{m} \sum_{j \in \mathcal{N}_m^{\mathcal{S}}(i)} s(\mathbf{x}_i, \mathbf{x}_j).$$

Lemma 1 (Concentration). Since $s \in [0,1]$ and subsampling is uniform, applying Hoeffding gives

$$|\tilde{\rho}_i - \mathbb{E}[\tilde{\rho}_i]| \le \sqrt{\frac{2\log(2S\delta^{-1})}{m}} \quad \text{with prob. } \ge \frac{1-\delta}{S}.$$

Lemma 2 (Bias). Lipschitz continuity of ρ and of s implies that neighbor sets in the full population and subsample differ by $O(\sqrt{\frac{1}{S}})$ in feature distance, yielding

$$|\mathbb{E}[\tilde{\rho}_i] - \rho_i| \le \frac{CL}{\sqrt{S}}.$$

Consequence. Combining Lemmas 1–2:

$$|\tilde{\rho}_i - \rho_i| \le O\left(\frac{L}{\sqrt{S}}\right).$$

D.4. Step 2: Separation and Score Stability

Define scores $\gamma_i = \rho_i \delta_i$. Lipschitz continuity of s gives:

Lemma 3. If $|\tilde{\rho}_i - \rho_i| \le \varepsilon_\rho$, then

$$|\tilde{\delta}_i - \delta_i| \le C\varepsilon_\rho.$$

Hence

$$|\tilde{\gamma}_i - \gamma_i| \le C\varepsilon_\rho = O\left(\frac{L}{\sqrt{S}}\right).$$

Lemma 4 (Top-K stability). If $\max_i |\tilde{\gamma}_i - \gamma_i| \le \Delta_{\min}/2$, then $\tilde{\gamma}$ and γ select exactly the same top-K tokens. Since $S = \beta K$ with $\beta \ge 2$, the condition holds with probability $\ge 1-\delta$.

Thus

$$\mathcal{C} = \mathcal{C}^*, \qquad \alpha = \alpha^*,$$

except possibly for $O(\frac{L}{\sqrt{S}})$ boundary tokens.

D.5. Step 3: Center and Assignment Perturbation

Let $\epsilon_c = O(\frac{L}{\sqrt{S}})$ be the center-location error induced by Step 1. By classical stability of centroid-based clustering [12], we have:

Lemma 5.

$$\|\mathbf{K}' - \mathbf{K}'_{\text{exact}}\|_F \le C\sqrt{K}\,\epsilon_c = O\left(\frac{L\sqrt{K}}{\sqrt{S}}\right),$$

and the same bound holds for $\mathbf{V}'$.

D.6. Step 4: Sensitivity of Attention to Perturbed Keys/Values

Define

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}'^\top}{\sqrt{d}}\right), \qquad \mathbf{A}_{\text{exact}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}'^\top_{\text{exact}}}{\sqrt{d}}\right).$$

Because softmax is Lipschitz on bounded domains,

$$\|\mathbf{A} - \mathbf{A}_{\text{exact}}\|_F \le C\,\frac{\|\mathbf{Q}\|_F}{\sqrt{d}} \|\mathbf{K}' - \mathbf{K}'_{\text{exact}}\|_F.$$

Since $\|\mathbf{Q}\|_F = O(\sqrt{N})$ and Lemma 5 applies,

$$\|\mathbf{A} - \mathbf{A}_{\text{exact}}\|_F = O\left(\sqrt{N} \cdot \frac{L\sqrt{K}}{\sqrt{S}}\right).$$

Thus

$$E_{\text{cluster}} \le C_1 L \sqrt{\frac{NK\log(\delta^{-1})}{S}}.$$

D.7. Step 5: Bounding Attention Compression Error

Under exact centers,

$$\mathbf{O}_{\text{exact}} = \mathbf{A}_{\text{exact}} \mathbf{V}'_{\text{exact}}, \qquad \mathbf{O}^* = \mathbf{A}^* \mathbf{V}.$$

The variation bound $\|\mathbf{x} - \mathbf{c}_k\| \le \sigma$ implies that all values within a cluster differ from their center by at most σ, hence

$$\|\mathbf{V} - \tilde{\mathbf{V}}\|_F \le C\sigma\sqrt{NK}.$$

Lemma 6. Under bounded cluster variation,

$$\|\mathbf{A}^* - \mathbf{A}_{\text{exact}}\|_F \leq C\frac{K}{\sqrt{N}}.$$

Consequently,

$$E_{\text{attn}} \leq C_2\|\mathbf{V}\|_F\frac{K}{N}.$$

D.8. Final Bound

Combining the contributions of clustering and attention compression,

$$\|\mathbf{O} - \mathbf{O}^*\|_F \;\leq\; C_1 L\sqrt{\frac{NK\log(\delta^{-1})}{S}} + C_2\|\mathbf{V}\|_F\frac{K}{N},$$

which completes the proof.

References

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010. 1

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 1

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 2

[4] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22367–22377, 2023. 1, 2, 3

[5] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Cross aggregation transformer for image restoration. *Advances in Neural Information Processing Systems*, 35:25478–25490, 2022. 3

[6] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, and Xiaokang Yang. Recursive generalization transformer for image super-resolution. *arXiv preprint arXiv:2303.06373*, 2023. 1

[7] 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*, pages 5197–5206, 2015. 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. 3

[9] 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. 1

[10] Wei Long, Xingyu Zhou, Leheng Zhang, and Shuhang Gu. Progressive focused transformer for single image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2279–2288, 2025. 2, 3

[11] 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*, 76:21811–21838, 2017. 1

[12] David Pollard. Strong consistency of k-means clustering. *The annals of statistics*, pages 135–140, 1981. 9

[13] Yuchuan Tian, Hanting Chen, Chao Xu, and Yunhe Wang. Image processing gnn: Breaking rigidity in super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24117, 2024. 1, 2, 3

[14] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 1

[15] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. 1

[16] Leheng Zhang, Yawei Li, Xingyu Zhou, Xiaorui Zhao, and Shuhang Gu. Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2856–2865, 2024. 2, 3

[17] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2