Structure-Constrained GAN: Selective Structure Guided Adversarial Network for Lightweight Image Super-Resolution

1th Na Zhang
Zhejiang Sci-Tech University
Hangzhou, China
103185968@qq.com

2rd Haidi Xu
Zhejiang Sci-Tech University
Hangzhou, China
x1078303266@gmail.com

3nd Cheng Xu
Zhejiang Sci-Tech University
Hangzhou, China
958996432@qq.com

4st Xiaoan Bao
Zhejiang Sci-Tech University
Hangzhou, China
baoxiaoan@zstu.edu.cn

5th Hainan Chen
Zhejiang Sci-Tech University
Hangzhou, China
hnchen@zstu.edu.cn

6th Qingqi Zhang*
Hangzhou Institute of Medicine
Chinese Academy of Sciences
Hangzhou, China
zhangtsingtsi@gmail.com

***Abstract*—Image Super-Resolution(SR) faces a long-standing trade-off between perception and distortion. Perceptual quality approaches based on adversarial training enhance visual realism by synthesizing high-frequency details, yet often introduce hallucinated textures and grainy artifacts in regions that are intrinsically smooth. When coupled with global modeling backbones such as state space models, the adversarial signal can be diffused through long-range interactions, making background artifacts more likely and more widespread. To address these issues, Structure-Constrained GAN (SCGAN) is proposed, which is a lightweight SR framework that explicitly injects structure awareness into both generator design and adversarial supervision. The architecture features a Selective Progressive Attention Module (SPAM) to prune redundant interactions in regions with poor texture, complemented by a branch based on Mamba for global context modeling. These streams are integrated via an Adaptive Gating Fusion Module (AGFM), which employs weights that vary spatially to amplify texture synthesis while suppressing artifact leakage in flat areas. Furthermore, a Variance Guided Adversarial Training (VGAT) strategy restricts the focus of the discriminator to regions dominant in texture, aligning perceptual objectives with structural constraints. Quantitative evaluations on five benchmarks show that SCGAN achieves SOTA performance, outperforming MambaIRv2-light by up to 0.26 dB in PSNR and reducing FLOPs by 18.8% with only 835K parameters.**

***Index Terms*—Image Super-Resolution, State Space Models, Sparse Attention, Adaptive Gating, Adversarial Learning**

I. Introduction

Image Restoration (IR) and Image Super-Resolution (SR) aim to recover high-quality images from degraded observations, enabling diverse downstream applications in computer vision, such as medical imaging [1], object detection [2], and remote sensing [3]. Driven by distinct optimization objectives, recent research in this field generally bifurcates into two dominant paradigms: fidelity-oriented and perceptual-quality-oriented approaches. While fidelity-oriented methods prioritize minimizing pixel-wise errors to yield high metric scores, they often suffer from over-smoothing and loss of high-frequency details. Conversely, perceptual quality-oriented methods focus on generating realistic textures to satisfy human visual perception, often at the cost of signal fidelity. This inherent conflict highlights the fundamental perception–distortion trade-off [4]. For perceptual quality-oriented approaches, adversarial training is widely adopted. Seminal works such as SRGAN [5] and ESRGAN [6] show that adversarial objectives encourage models to hallucinate visually plausible high-frequency textures. However, this benefit can be a double-edged sword: without explicit local structural constraints, the discriminator-driven supervision may bias the generator toward texture synthesis even in regions that should remain smooth, thereby introducing artifacts, as shown in Fig. 1(b). Consequently, although perceptual quality improves, objective fidelity metrics

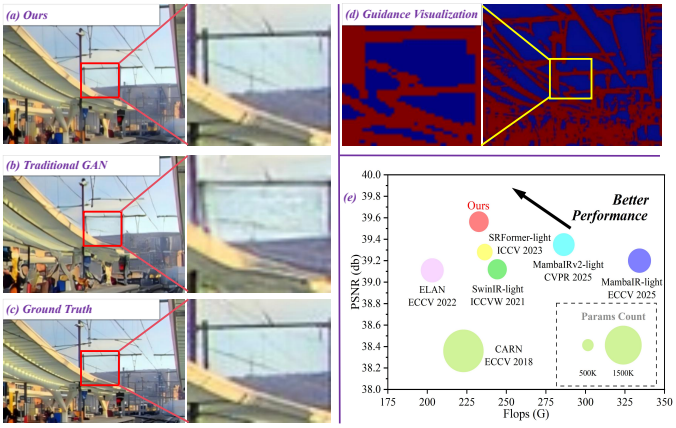

Fig. 1. (a) The restoration result of our proposed SCGAN. (b) The result of other GAN-based methods, exhibiting noticeable artifacts and hallucinations. (c) The Ground Truth (GT). (d) Visualization of our proposed guidance strategy. (e) Performance trade-off comparison (Bubble Chart) with other state-of-the-art (SOTA) methods.

*Corresponding author: Qingqi Zhang (zhangtsingtsi@gmail.com).

(e.g., PSNR/SSIM) often degrade due to increased pixel-wise deviations and structural inconsistencies. Complementing these adversarial strategies , the pursuit of superior perceptual quality has also driven a paradigm shift in network architectures. While early CNN-based methods [7] struggled to generate globally consistent textures due to limited receptive fields, Transformers [8] reshaped the landscape by leveraging attention mechanisms. Despite their strong global modeling ability, the quadratic complexity of Transformers is prohibitive for high-resolution generation. Thus, State Space Models (SSMs), particularly Mamba [9], [10], have emerged as a compelling alternative that preserves global capacity crucial for high-quality texture synthesis while reducing computational burden. Mamba combines the receptive field advantages of Transformers with linear complexity, exhibiting promising prospects for pushing the boundaries of perceptual quality in low-level vision tasks. [11]However, Mamba, an architecture with a global receptive field, is more prone to generating hallucinations or background noise across the entire image during adversarial training.

Given this persistent tension, a natural question arises: how can we bridge the gap between fidelity-oriented and perceptual-quality-oriented paradigms? One common strategy is to augment fidelity-driven restoration with auxiliary guidance signals. Nevertheless, existing guidance is often applied in a largely uniform manner and remains insufficiently selective: it fails to reliably distinguish structure-rich regions that truly require high-frequency recovery from intrinsically smooth areas that should remain texture-free. This imperfect guidance frequently leads to an undesirable trade-off, residual blur around informative structures, yet spurious details or grainy artifacts "leaking" into flat backgrounds, thereby hurting both quantitative fidelity and visual comfort, as shown in Fig.1. In this work, we introduce a hybrid guidance design that combines state-space sequence modeling with Transformer attention, leveraging their complementary strengths. Specifically, the SSM branch provides efficient global context modeling that scales to high-resolution inputs, while the attention branch offers content-adaptive selectivity to focus guidance on salient structures. A learned gating mechanism further modulates the guidance strength across spatial locations to produce refined fused features. To strictly enforce texture fidelity, we incorporate a Variance-Guided Adversarial Training strategy. In this paradigm, pixel-wise variance serves as distinct *direct guidance*, spatially constraining the discriminator's focus strictly to texture-rich regions. Simultaneously, the feature maps fused by the gating mechanism serve as the primary subject for an auxiliary consistency loss, providing indirect corrective guidance to ensure structural alignment. This selective, globally aware guidance yields more faithful reconstructions without introducing spurious textures, achieving a better perception–distortion trade-off.

Building on the above discussions, we propose **Structure-Constrained GAN (SCGAN)**, a unified framework that bridges efficient global modeling and high fidelity texture synthesis via selective, structure guided adversarial restoration. Unlike prior approaches that treat guidance as an auxiliary component or separate architectural design from optimization objectives, we argue for joint design. Reducing pixel wise inconsistencies requires an objective that enforces explicit structural constraints, together with a network that can inherently capture and propagate structure aware signals at scale. Concretely, SCGAN combines a global modeling branch based on SSMs with a selective guidance branch based on attention, and introduces a learned gating mechanism that modulates the guidance strength across spatial locations. It strengthens guidance in structure rich regions while weakening it in intrinsically smooth areas, which helps prevent texture leakage and reduce spurious artifacts. As illustrated in the visualized guidance map in Fig. 1(d), our strategy successfully differentiates smooth regions such as the sky (highlighted in blue), effectively suppressing unnecessary guidance. Consequently, as shown in the reconstruction result in Fig. 1(a), spurious artifacts are avoided. By integrating objective level structural constraints with an architecture capturing and propagating them effectively, SCGAN produces reconstructions that are visually convincing while remaining consistent with the ground truth.

The main contributions are summarized as follows:

- We propose SCGAN, a hybrid architecture combining dynamic local sparsity with global state space modeling. By balancing pruning with long-range context, it eliminates background redundancy while preserving structural coherence, addressing the perception-distortion trade-off.
- We introduce the SPAM and AGFM modules for feature purification. SPAM implements progressive pruning to remove noise, while AGFM integrates heterogeneous features. This synergy blocks artifact propagation and prevents noise amplification during adversarial training.
- We propose a Variance-Guided Adversarial Fine-tuning strategy. By strictly constraining the discriminator to texture-rich regions, this strategy effectively mitigates pixel misalignment in smooth areas, reconciling high perceptual quality with objective fidelity.
- We conduct extensive experiments on five benchmark datasets.Evaluations demonstrate that SCGAN achieves state-of-the-art performance, showing superior robustness and generalization compared to existing methods.

II. Related Work

A. State Space Models

Originating in control theory and time-series analysis—most notably within linear Gaussian systems and Kalman filtering—State Space Models (SSMs) have experienced a resurgence in deep learning. This renewed interest is driven by their efficiency in modeling long-range dependencies with linear time and space complexity. Vim [12] and VMamba [13] introduced efficient fixed selective scanning pattern for visual data, a technique MambaIR [14] subsequently adapted for global modeling in image restoration. [15] SSMs map a one-dimensional input function $x(t) \in \mathbb{R}$ to an output $y(t)$ through

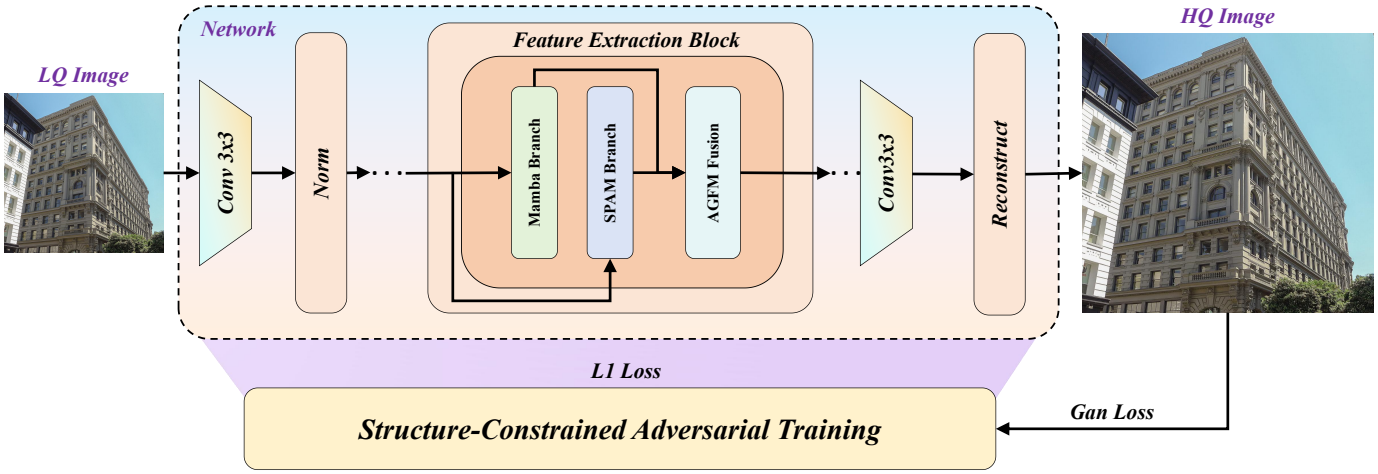

Fig. 2. Overview of the proposed SCGAN framework. The framework is composed of a hierarchical generation network and a structure-constrained optimization strategy. In terms of Network Architecture, deep features are extracted via stacked Feature Extraction Blocks. Inside each block, the Mamba Branch and SPAM Branch work in parallel to model global context and local details, respectively, followed by AGFM for adaptive fusion. The entire model is optimized using a hybrid strategy that combines standard L1 Loss with Structure-Constrained Adversarial Training to balance perceptual quality and objective fidelity.

an implicit latent state $h(t) \in \mathbb{R}^N$. This process is governed by the following Ordinary Differential Equations (ODEs):

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t) \\ y(t) &= \mathbf{C}h(t), \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ serves as the state evolution matrix, while $\mathbf{B} \in \mathbb{R}^{N\times 1}$ and $\mathbf{C} \in \mathbb{R}^{N\times 1}$ act as projection parameters controlling the input and output transformations.

To adapt this continuous framework to discrete sequence data, a timescale parameter $\Delta \in \mathbb{R}_{>0}$ is introduced. By applying the Zero-Order Hold (ZOH) discretization method, the continuous parameters $(\mathbf{A}, \mathbf{B})$ are transformed into their discrete counterparts $(\overline{\mathbf{A}}, \overline{\mathbf{B}})$:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}) \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}\left(\exp(\Delta\mathbf{A}) - \mathbf{I}\right) \cdot \Delta\mathbf{B}. \end{aligned} \tag{2}$$

Consequently, the system can be formulated as a discrete-time recurrence:

$$\begin{aligned} h_t &= \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \\ y_t &= \mathbf{C}h_t. \end{aligned} \tag{3}$$

For enhanced computational efficiency, particularly during parallel training, this recurrent process can be equivalently reformulated as a global convolution operation. Let $\circledast$ denote the convolution operator and $\overline{\mathbf{K}}$ represent the structured kernel:

$$\begin{aligned} y &= x \circledast \overline{\mathbf{K}}, \\ \overline{\mathbf{K}} &= (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}). \end{aligned} \tag{4}$$

B. Sparse Attention and Dynamic Feature Selection

Sparsity and efficiency have become central themes in minimizing computational costs. SwinIR [8] introduces sparsity via Window-based Multi-head Self-Attention (W-MSA) to limit interactions within local regions. Complementarily, HAT [16] employs channel attention to emphasize highly activated features. Other approaches leverage Deformable Attention to flexibly sample sparse key points across the spatial domain.

However, the sparsity in these methods is typically static or spatially uniform. [17] For instance, W-MSA incurs redundant computation on smooth background windows, while Deformable Attention remains computationally intensive despite its flexibility. Conversely, our SPAM exploits inter-layer correlations to establish a "progressive pruning" mechanism. This implements a data-dependent hard sparsity strategy that explicitly excludes background noise from processing, rather than merely suppressing it via soft weights.

C. Structure-Constrained Adversarial Training Strategies

The design of the discriminator plays an instrumental role in perception-driven restoration. The research trajectory has evolved from early Global GANs to spatially localized architectures, such as PatchGAN [18] and Relativistic GANs (RaGAN) [19], with the primary aim of refining local discriminative precision. These architectural advancements are often supplemented by recent works that incorporate frequency or edge-based losses to further sharpen textural details. [20] [21]

Yet, most methods remain global soft constraints, often forcing the generator to fill smooth background areas with artificial noise to satisfy the discriminator. Our Variance-Guided Fine-tuning resolves this via hard masking, which explicitly excludes smooth regions from the adversarial feedback loop. This ensures that texture optimization is physically restricted to relevant areas, a precision unattained by previous works.

III. Methodology

A. Overall Architecture

SCGAN orchestrates a sophisticated hierarchical residual paradigm, tailored to facilitate robust information flow for high-fidelity image restoration, as shown in Fig.2. The proposed network comprises three pivotal phases: initial feature embedding, deep structural representation learning, and high-fidelity image reconstruction. Given a degraded input image $I_{LQ} \in \mathbb{R}^{H\times W\times 3}$, we first apply a 3×3 convolution to extract shallow features, transforming I_{LQ} into $I_{SL} \in \mathbb{R}^{H\times W\times C}$.

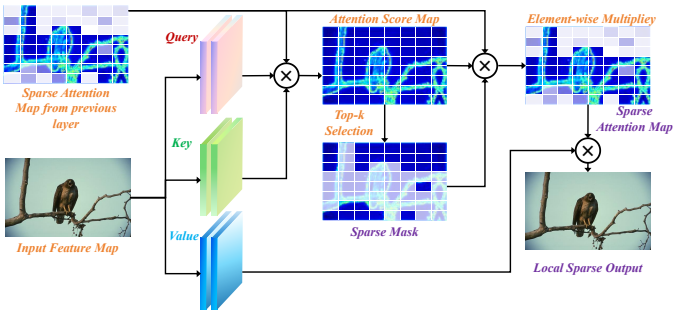

Fig. 3. Illustration of the Selective Progressive Attention Module (SPAM). By applying Top-k Selection to the attention scores, the module generates a binary Sparse Mask to physically prune background redundancy. This ensures the attention mechanism focuses exclusively on informative high-frequency textures, producing a high-purity local sparse output.

Here, H and W represent the height and width of the input image, respectively, and C denotes the number of feature channels. The shallow features are subsequently propagated through a series of **Feature Extraction Blocks (FEB)** to extract deep hierarchical features. Let $F_D^l \in \mathbb{R}^{H\times W\times C}$ denote the feature map at the l-th layer, where $l \in \{1, 2, \ldots, L\}$. Each FEB is meticulously constructed as a hybrid unit, synergistically integrating the **Selective Progressive Attention Module (SPAM)** and the **Attentive State Space Module (ASSM)**, simultaneously modeling local discriminative features while characterizing global structural dependencies. Finally, a 3×3 convolution is applied to project the refined features back to the image space, yielding the final output I_{SR}. The entire network is optimized via a composite objective function $\mathcal{L}_{total}$, which integrates the pixel-wise reconstruction loss $\mathcal{L}_{rec}$ for signal fidelity with a variance-guided adversarial loss $\mathcal{L}_{adv}$. This guidance mechanism explicitly compels the generator to prioritize the restoration of realistic textures in high-frequency regions, ultimately yielding the high-quality output image I_{HQ}.

B. Selective Progressive Attention Module

Let F_D^l be the input features of the l-th layer. SPAM operates strictly on a sparse index set $\mathcal{I}_{l-1}$ inherited from the previous stage to reduce redundancy. Query, key, and value embeddings are generated via linear projections incorporating relative positional bias $\mathbf{B} \in \mathbb{R}^{M^2\times M^2}$:

$$\mathbf{Q} = F_D^l \mathbf{W}_Q, \quad \mathbf{K} = F_D^l \mathbf{W}_K, \quad \mathbf{V} = F_D^l \mathbf{W}_V, \tag{5}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable projection matrices.

Instead of computing a dense affinity matrix, we estimate the raw attention logits $\mathbf{E}^{(l)}$ solely on the active connections defined by $\mathcal{I}_{l-1}$, which corresponds to the generation of the Attention Score Map as illustrated in Fig. 3. The sparse affinity computation is formulated as a conditional mapping:

$$\mathbf{E}_{i,j}^{(l)} = \begin{cases} \left(\mathbf{q}_i \mathbf{k}_j^T\right) \cdot \tau + \mathbf{B}_{i,j}, & \text{if } (i,j) \in \mathcal{I}_{l-1} \\ -\infty, & \text{otherwise} \end{cases}, \tag{6}$$

where $\mathbf{q}_i, \mathbf{k}_j \in \mathbb{R}^{d_k}$ represent the feature vectors of the i-th query and j-th key, and τ is the temperature scaling factor. This step is implemented using efficient Sparse Matrix Multiplication (SMM) to bypass zero-value computations.

Targeting the suppression of background noise and the maintenance of temporal structural consistency, we incorporate a modulation mechanism. The raw logits are normalized via Softmax and explicitly gated by the historical prior $\mathbf{A}^{(l-1)}$ to yield the modulated attention map $\mathbf{M}^{(l)}$:

$$\mathbf{M}_{i,j}^{(l)} = \mathbf{A}_{i,j}^{(l-1)} \cdot \frac{\exp(\mathbf{E}_{i,j}^{(l)})}{\sum_{k\in\mathcal{N}_i} \exp(\mathbf{E}_{i,k}^{(l)})}, \tag{7}$$

where $\mathcal{N}_i = \{k \mid (i,k) \in \mathcal{I}_{l-1}\}$ denotes the sparse neighborhood (i.e., the set of valid key indices) for the i-th query token, strictly constrained by the inherited index set.

To guarantee probabilistic validity and numerical stability, the modulated map is re-normalized. The final output $\mathbf{X}_l$ is then generated by aggregating value features enriched with Locally-enhanced Positional Encoding (LePE):

$$\mathbf{x}_i^{(l)} = \left(\sum_{j\in\mathcal{N}_i} \frac{\mathbf{M}_{i,j}^{(l)} + \epsilon}{\sum_{k\in\mathcal{N}_i} (\mathbf{M}_{i,k}^{(l)} + \epsilon)} \mathbf{v}_j + \mathrm{LePE}(\mathbf{v}_i) \right) \mathbf{W}_O, \tag{8}$$

where ϵ is a stabilizer, $\mathbf{v}_j$ denotes the value vector of the j-th token, and $\mathbf{W}_O$ is the output projection matrix.

Finally, to facilitate progressive focusing, we evolve the sparse computational manifold for the subsequent layer. We prune the index set by strictly retaining the most discriminative connections based on a decaying budget:

$$\mathcal{I}_l = \mathrm{argtopk}_{(i,j)} \left(\mathbf{A}^{(l)}, \lfloor \alpha_l |\mathcal{I}_{l-1}| \rfloor \right), \text{ with } \alpha_l = \gamma^{\lfloor l/5 \rfloor}. \tag{9}$$

where $\gamma \in (0, 1)$ controls the stepwise sparsification process. Unlike aggressive layer-wise pruning, we apply the decay factor γ at intervals of 5 layers. This interval-based strategy maintains sufficient context for feature aggregation between pruning stages, ensuring that while redundant backgrounds are eventually removed, the intricate high-frequency structural details are robustly preserved.

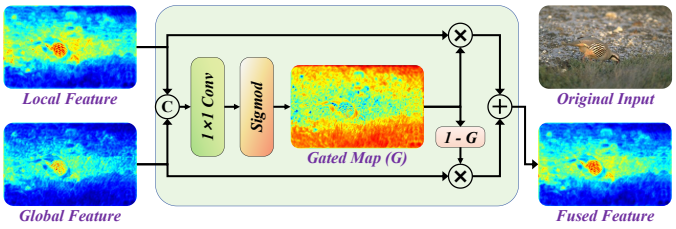

Fig. 4. Adaptive Gating Fusion Module (AGFM). A pixel-wise Gated Map (G) is generated to dynamically assign complementary weights (G and $1-G$) to local and global features, enabling adaptive fusion of structural details and global context.

C. Adaptive Gating Fusion

To harmonize the restoration of microscopic textures and macroscopic structures, we engineer a dual-branch synergy. The Adaptive Semantic State Mamba (ASSM) operates in

parallel with SPAM to capture long-range dependencies, complementing local details. This feature dichotomy is ultimately resolved via a dynamic fusion mechanism.

Global Modeling via ASSM. While SPAM targets local sparsity, ASSM captures long-range dependencies via semantic reorganization. We derive a permutation index π by sorting tokens based on a learnable policy to cluster similar textures:

$$\pi = \text{argsort}\left(\text{argmax}(\text{Gumbel}(\mathbf{W}_{route}\mathbf{X}))\right), \tag{10}$$

where $\mathbf{W}_{route}$ denotes the learnable weights and $\text{Gumbel}(\cdot)$ represents the Gumbel-Softmax relaxation, facilitating the shuffling of inputs into a content-coherent sequence $\mathbf{X}_\pi = \mathbf{X}[\pi]$. The sequence undergoes Selective Scan (SSM) and inverse permutation π^{-1} to yield the global feature:

$$\mathbf{F}_{global} = \text{SSM}(\mathbf{X}_\pi)[\pi^{-1}]. \tag{11}$$

This effectively bridges spatially dispersed yet texturally coherent regions, capturing non-local correlations.

Feature Integration. Addressing artifacts caused by naive addition, we generate a pixel-wise selection mask $\mathbf{G}$ (visualized as the Gating Map in Fig. 4) by projecting the concatenated inputs:

$$\mathbf{G} = \sigma\left(\mathbf{W}_{gate} * [\mathbf{F}_{local}, \mathbf{F}_{global}] + \mathbf{b}\right), \tag{12}$$

where $[\cdot,\cdot]$ denotes channel concatenation, $\mathbf{W}_{gate}$ is a 1×1 convolution for channel mixing, and σ is Sigmoid. $\mathbf{G} \in (0, 1)$ acts as a dynamic switch. The final output is synthesized via a complementary weighted aggregation:

$$\mathbf{F}_{out} = \mathbf{G} \odot \mathbf{F}_{local} + (1 - \mathbf{G}) \odot \mathbf{F}_{global}. \tag{13}$$

This gating mechanism enables the network to spatially adapt its focus: prioritizing $\mathbf{F}_{local}$ in high-frequency texture regions (where $\mathbf{G} \to 1$) while leveraging $\mathbf{F}_{global}$ in smooth backgrounds or repetitive patterns (where $\mathbf{G} \to 0$), ensuring a seamless integration of dual-branch information.

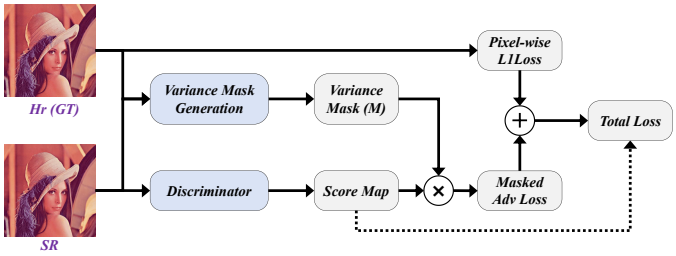

Fig. 5. Variance-Guided Adversarial Training. A binary Variance Mask (M) derived from the ground truth is utilized to spatially filter the discriminator's Score Map. This explicitly restricts the Masked Adversarial Loss to texture-rich regions, preventing the generation of artifacts in smooth backgrounds while enhancing high-frequency details.

D. Variance-Guided Optimization

To align optimization with perceptual relevance, as depicted in the pipeline shown in Fig. 5, we first identify high-frequency texture regions using the local variance of the ground truth I_{GT}. We generate a binary guide mask $\mathcal{M}$ by thresholding the variance map Σ:

$$\mathcal{M}_{x,y} = \mathbb{I}(\Sigma_{x,y} > \delta), \quad \text{where } \Sigma_{x,y} = \text{Var}(I_{GT}). \tag{14}$$

Here, δ is a predefined variance threshold. This mask $\mathcal{M}$ serves as a spatial anchor for the subsequent adversarial training.

Reconstruction and Spectral Consistency. To ensure robust pixel-wise fidelity and bridge the frequency gap, we employ a composite reconstruction objective consisting of the Charbonnier loss $\mathcal{L}_{char}$ and the Frequency loss $\mathcal{L}_{freq}$:

$$\mathcal{L}_{rec} = \sqrt{\|I_{SR} - I_{GT}\|^2 + \epsilon^2} + \lambda_{freq}\|\mathcal{F}(I_{SR}) - \mathcal{F}(I_{GT})\|_1, \tag{15}$$

where $\mathcal{F}(\cdot)$ denotes the Fast Fourier Transform, enforcing consistency in the spectral domain to prevent blurring.

Masked Adversarial Alignment. Standard adversarial training often induces hallucinated artifacts in smooth regions. To address this, we constrain the discriminator to focus *exclusively* on the texture-rich areas defined by $\mathcal{M}$. The masked adversarial loss is formulated as:

$$\mathcal{L}_{adv}^{\mathcal{M}} = -\mathbb{E}\left[\mathcal{M} \odot \log(\sigma(D(I_{SR})))\right]. \tag{16}$$

By applying the mask $\mathcal{M}$ via element-wise multiplication $\odot$, we penalize unrealistic textures only where high-frequency details naturally exist.

The final objective function integrates these terms to balance structural fidelity and texture realism:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv}^{\mathcal{M}}. \tag{17}$$

Minimizing this composite objective allows the network to recover intricate high-frequency details while effectively suppressing artifacts in smooth background regions.

IV. Experiments

In this section, we conduct extensive experiments to evaluate the proposed method, and the results show that it achieves state-of-the-art performance across five benchmark datasets, consistently outperforming prior approaches in both quantitative metrics and perceptual quality, particularly in recovering fine details while maintaining high structural fidelity.

A. Experiment Settings

The model is trained on the DIV2K dataset [27], which contains 800 high-quality training images. Standard data augmentation is employed to enrich the training set and enhance robustness, using random horizontal flips and 90-degree rotations [28]. For evaluation, five standard super-resolution benchmarks are used: Set5 [29], Set14 [30], BSDS100 [31], Urban100 [32], and Manga109 [33]. These datasets cover a wide range of content, from natural images to urban scenes and manga-style illustrations, with varying texture richness and structural complexity. Restoration performance is evaluated using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [34], computed on the luminance (Y) channel in the YCbCr color space between the restored images and the

TABLE I
QUANTITATIVE COMPARISON (PSNR/SSIM) FOR LIGHTWEIGHT IMAGE SR WITH STATE-OF-THE-ART METHODS ON BENCHMARK DATASETS. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE.

Method	scale	Set5		Set14		BSDS100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CARN2018 [22]	2×	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
LatticeNet2020 [23]	2×	38.13	0.9610	33.78	0.9193	32.25	0.9005	32.43	0.9302	-	-
SwinIR-light2021 [8]	2×	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
ELAN2022 [24]	2×	38.17	0.9611	33.94	0.9207	32.30	0.9012	32.76	0.9340	39.11	0.9782
SRFormer-light2023 [25]	2×	38.23	0.9613	33.94	0.9209	32.36	0.9019	32.91	0.9353	39.28	0.9785
MambaIRv2-light2025 [26]	2×	38.26	0.9615	34.09	0.9221	32.36	0.9019	33.26	0.9378	39.35	0.9785
SCGAN (ours)	2×	38.33	0.9619	34.15	0.9225	32.39	0.9024	33.39	0.9392	39.56	0.9790
CARN2018 [22]	3×	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
LatticeNet2020 [23]	3×	34.53	0.9281	30.39	0.8424	29.15	0.8059	28.33	0.8538	-	-
SwinIR-light2021 [8]	3×	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
ELAN2022 [24]	3×	34.61	0.9288	30.55	0.8463	29.21	0.8081	28.69	0.8624	34.00	0.9478
SRFormer-light2023 [25]	3×	34.67	0.9296	30.57	0.8469	29.26	0.8099	28.81	0.8655	34.19	0.9489
MambaIRv2-light2025 [26]	3×	34.71	0.9298	30.68	0.8483	29.26	0.8098	29.01	0.8689	34.41	0.9497
SCGAN (ours)	3×	34.79	0.9303	30.72	0.8489	29.32	0.8112	29.24	0.8717	34.64	0.9505
CARN2018 [22]	4×	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
LatticeNet2020 [23]	4×	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873	-	-
SwinIR-light2021 [8]	4×	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
ELAN2022 [24]	4×	32.43	0.8975	28.78	0.7858	27.69	0.7406	26.54	0.7982	30.92	0.9150
SRFormer-light2023 [25]	4×	32.51	0.8988	28.82	0.7872	27.73	0.7422	26.67	0.8032	31.17	0.9165
MambaIRv2-light2025 [26]	4×	32.51	0.8992	28.84	0.7878	27.75	0.7426	26.82	0.8079	31.24	0.9182
SCGAN (ours)	4×	32.62	0.9000	28.93	0.7888	27.80	0.7437	26.98	0.8100	31.50	0.9192

corresponding ground-truth images. In addition, the number of parameters (Params) and multiply–accumulate operations (MACs) are reported to assess computational efficiency, where MACs are measured with an input resolution of 1280×720. SCGAN is implemented in PyTorch and trained on 4 NVIDIA RTX A6000 GPUs. For fair comparison, the architecture is configured with $K = 4$ groups, $M = 5$ blocks, and channel dimension $C = 48$. Training uses 64×64 patches with a batch size of 32 for 500,000 iterations, optimized with Adam and a multi-step learning rate schedule ($4 \times 10^{-4} \rightarrow 10^{-6}$). A two-stage training strategy is adopted, consisting of L_1 pre-training followed by variance-guided fine-tuning, with loss weights set to $\lambda_{pix} = \lambda_{per} = 1.0$ and $\lambda_{adv} = 0.01$.

B. Comparison with State-of-the-Art Methods

The proposed method is compared with representative lightweight approaches, including CNN-based methods CARN [22] and LatticeNet [23], as well as Transformer and SSM-based models ELAN [24], SwinIR-light [8], SRFormer-light [25], and MambaIRv2-light [14]. As reported in Tab. I, SCGAN consistently achieves the best PSNR and SSIM on Set5, Set14, BSDS100, Urban100, and Manga109 under different scale settings. The advantage becomes more evident on structure-rich benchmarks. For example, under $\times 4$ on Urban100, SCGAN reaches 26.98 dB and 0.8100 SSIM, improving upon the strongest baseline MambaIRv2-light by 0.16 dB. Similar trends are observed on Manga109, where SCGAN attains 31.50 dB at $\times 4$, and remains the top-performing method at $\times 2$ and $\times 3$. Overall, these results indicate that SCGAN offers more accurate reconstruction in challenging scenes with abundant high-frequency textures and geometric structures. Qualitative comparisons are presented in Fig. 6. On Urban100 and Manga109, existing methods often exhibit slight spatial misalignment or over-smoothed structures in regions with complex textures. By contrast, the proposed approach reconstructs sharper edges and more coherent backgrounds while effectively suppressing spurious high-frequency patterns commonly introduced by adversarial training. Notably, in the PsychoStaff example in Fig. 6, only the proposed method faithfully reconstructs the thin eyebrow strokes. In img056, it is also the only method that fully recovers the subtle sky textures. Furthermore, in img012 and img076, the proposed method not only restores fine details but also preserves accurate building geometry, whereas other methods exhibit structural distortions and noticeable detail loss. We further evaluate the computational efficiency in Table II. Despite improved performance, our model maintains high efficiency. Compared to MambaIRv2-light, which requires 286G FLOPs, our SCGAN reduces the computational cost to 233G FLOPs

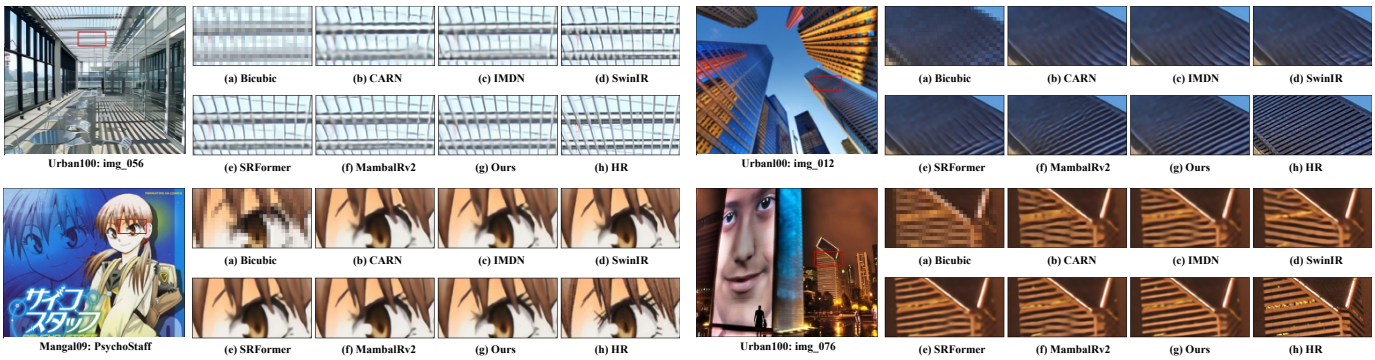

Fig. 6. Qualitative comparison of our SCGAN with different methods on 4× classic image SR.

TABLE II
COMPARISON WITH OTHER EFFICIENT SR METHODS. PERFORMANCE INDICATORS COMPARE TESTED ON SET5 (×2).

Model	SwinIR	SRFormer	MambaIRv2	Ours
Params (K)	910	853	774	835
FLOPs (G)	244	236	286	233
PSNR (dB)	38.14	38.23	38.26	**38.33**
SSIM	0.9611	0.9613	0.9615	**0.9619**

TABLE III
ABLATION STUDIES OF THREE COMPONENTS. THE MODELS ARE TRAINED ON DIV2K, AND TESTED ON SET5(×2).

SPAM	AGFM	Variance-Guided Adversarial Training	FLOPs	PSNR
✓			193.5G	38.29dB
✓	✓		205.3G	38.30dB
✓	✓	✓	232.6G	38.33dB

while achieving a higher PSNR . This efficiency gain (18.8%) is attributed to the SPAM mechanism, which effectively prunes redundant computations in smooth background regions.

C. Ablation Studies

Ablation experiments are conducted on Set5 under the ×2 setting to quantify the effect of each design choice. Starting from a common baseline, components are added one at a time while keeping the network configuration and training protocol fixed; computational cost is reported in terms of FLOPs together with PSNR to reflect the accuracy–efficiency trade-off. As summarized in Table III, introducing SPAM substantially reduces computation by eliminating redundant operations, while maintaining comparable reconstruction accuracy. Adding AGFM consistently improves PSNR, indicating more effective feature fusion and alleviation of conflicts between local detail cues and long-range contextual information. Finally, incorporating variance-guided training brings the largest additional gain in perceptual quality: by steering the adversarial constraint toward texture-dominant regions, it encourages the discriminator to focus on informative high-frequency content, thereby improving detail fidelity without over-amplifying artifacts.

D. Analysis and Visualization

The fused feature map produced by AGFM is visualized in Fig. 7(b). Compared with naive element-wise addition, which often propagates inconsistent responses and introduces spurious activations, AGFM yields a cleaner and more structured representation. Fine textures and edges are preserved with higher contrast, largely inherited from the SPAM branch,

Fig. 7. **Visualization of internal mechanisms.** (a) The **Original Image** from the dataset.(b) The **Fused Feature Map** output by AGFM, demonstrating clear structural details in texture regions while maintaining a clean, noise-free state in smooth areas. (c) The **Variance Mask** $\mathcal{M}$ derived from the ground truth, effectively distinguishing high-frequency textures from smooth backgrounds.

while homogeneous regions remain smooth and stable, benefiting from the Mamba branch. This indicates that AGFM performs content-adaptive fusion that suppresses conflicting cues and effectively combines the complementary strengths of the two pathways. Given the input image in Fig. 7(a), the corresponding variance mask in Fig. 7(c) highlights edge- and texture-dominant regions and stays inactive in flat areas. This spatial prior encourages the adversarial supervision to focus on informative high-frequency structures rather than uniform backgrounds. Consequently, background noise and undesired high-frequency artifacts are reduced, which helps alleviate pixel-level inconsistencies and improves structural alignment in the restored results.

V. CONCLUSION

In this paper, we presented SCGAN, a lightweight yet powerful framework that effectively bridges the gap between

efficient global modeling and high-fidelity texture restoration. Addressing the inherent "hallucination" and background noise issues in Mamba-based adversarial training, a holistic solution spanning from architectural design to optimization strategy is proposed. Specifically, the Selective Progressive Attention module (SPAM) is introduced to excise computational redundancy and distill high-purity local features, while the Adaptive Gating Fusion Module (AGFM) was designed to harmonize the semantic disparity between local and global representations. To further enhance texture fidelity, a Variance-Guided Adversarial Fine-tuning strategy is integrated to spatially constrain the discriminator's focus strictly to texture-rich regions. Extensive benchmarking confirms that SCGAN outperforms state-of-the-art lightweight models (e.g., MambaIRv2-light) in both quantitative accuracy and perceptual realism, producing images with sharper edges and artifact-free backgrounds. Ultimately, our method demonstrates a superior perception-distortion trade-off while maintaining high computational efficiency.

REFERENCES

[1] Y. Zhang, K. Li, K. Li, and Y. Fu, "Mr image super-resolution with squeeze and excitation reasoning attention network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 425–13 434.

[2] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[3] Y. Wang, Z. Shao, T. Lu, X. Huang, J. Wang, Z. Zhang, and X. Zuo, "Lightweight remote sensing super-resolution with multi-scale graph attention network," *Pattern Recognition*, vol. 160, p. 111178, 2025.

[4] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6228–6237.

[5] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

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

[7] 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.

[8] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.

[9] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[10] T. Dao and A. Gu, "Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024," *URL https://arxiv. org/abs/2405.21060*, p. 1, 2021.

[11] L. Peng, X. Di, Z. Feng, W. Li, R. Pei, Y. Wang, X. Fu, Y. Cao, and Z.-J. Zha, "Directing mamba to complex textures: An efficient texture-aware state space model for image restoration," *arXiv preprint arXiv:2501.16583*, 2025.

[12] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[13] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," *Advances in neural information processing systems*, vol. 37, pp. 103 031–103 063, 2024.

[14] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia, "Mambair: A simple baseline for image restoration with state-space model," in *European conference on computer vision*. Springer, 2024, pp. 222–241.

[15] J. Qiao, J. Liao, W. Li, Y. Zhang, Y. Guo, J. Xie, J. Hu, and S. Lin, "Hi-mamba: Hierarchical mamba for efficient image super-resolution," *IEEE Transactions on Image Processing*, vol. 34, 2025.

[16] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 367–22 377.

[17] Z. Ni, Y. Zhang, W. Yang, and S. Kwong, "Structural similarity-inspired unfolding for lightweight image super-resolution," *IEEE Transactions on Image Processing*, 2025.

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[19] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard gan," *arXiv preprint arXiv:1807.00734*, 2018.

[20] M. Aktar, K. S. Yadav, and R. H. Laskar, "A robust and lightweight generative adversarial network with zero-shot learning for image super-resolution," in *2025 National Conference on Communications (NCC)*. IEEE, 2025, pp. 1–6.

[21] P. Rao and T. Zeng, "Eatgan: An edge-attention guided generative adversarial network for single image super-resolution," *arXiv preprint arXiv:2509.14550*, 2025.

[22] 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.

[23] X. Luo, Y. Xie, Y. Zhang, Y. Qu, C. Li, and Y. Fu, "Latticenet: Towards lightweight image super-resolution with lattice block," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 272–289.

[24] X. Zhang, H. Zeng, S. Guo, and L. Zhang, "Efficient long-range attention network for image super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 649–667.

[25] D. Zhou, J. Shang, Z. Guo, Z. Li, and S. Yan, "Srformer: Permuted self-attention for single image super-resolution," *arXiv preprint arXiv:2303.09735*, 2023.

[26] H. Guo, Y. Guo, Y. Zha, Y. Zhang, W. Li, T. Dai, S.-T. Xia, and Y. Li, "Mambairv2: Attentive state space restoration," *arXiv preprint arXiv:2411.15269*, 2024.

[27] 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.

[28] B. Lim, S. Son, H. Kim, S. Nah, and K. M. 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.

[29] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proceedings of the British Machine Vision Conference*, 2012, pp. 135.1–135.10.

[30] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*. Springer, 2010, pp. 711–730.

[31] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, 2001, pp. 416–423.

[32] 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.

[33] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools and Applications*, vol. 76, pp. 21 811–21 838, 2017.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.