

Efficient Perceptual Image Super Resolution: AIM 2025 Study and Benchmark

Bruno Longarela $\diamond^*$ Marcos V. Conde $\diamond^\dagger$ Alvaro Garcia $\diamond$ Radu Timofte †

$\diamond$ Cidaut AI † University of Würzburg, Computer Vision Lab

Abstract

This paper presents a comprehensive study and benchmark on Efficient Perceptual Super-Resolution (EPSR). While significant progress has been made in efficient PSNR-oriented super resolution, approaches focusing on perceptual quality metrics remain relatively inefficient. Motivated by this gap, we aim to replicate or improve the perceptual results of Real-ESRGAN while meeting strict efficiency constraints: a maximum of 5M parameters and 2000 GFLOPs, calculated for an input size of 960×540 pixels. The proposed solutions were evaluated on a novel dataset consisting of 500 test images of 4K resolution, each degraded using multiple degradation types, without providing the original high-quality counterparts. This design aims to reflect realistic deployment conditions and serves as a diverse and challenging benchmark. The top-performing approach manages to outperform Real-ESRGAN across all benchmark datasets, demonstrating the potential of efficient methods in the perceptual domain. This paper establishes the modern baselines for efficient perceptual super resolution.

1. Introduction

Single-image super-resolution (SR) aims to reconstruct a high-resolution image from a low-resolution input, which is a fundamentally ill-posed inverse problem. Traditionally, bicubic down-sampling has been the standard degradation model due to its simplicity and reproducibility. Models trained solely on bicubic degradation, however, perform poorly when confronted with complex real-world degradations such as noise, JPEG compression artifacts, and various types of blur [7].

Optimizing exclusively for distortion metrics like PSNR

or SSIM tends to produce overly smooth results due to regression to the mean, resulting in outputs that lack high-frequency details and perceptual quality. This phenomenon is theoretically supported by the perception-distortion tradeoff, which establishes that improving both distortion and perceptual quality simultaneously is fundamentally limited [3, 4]. Consequently, approaches incorporating perceptual losses have been shown to significantly improve the naturalness of generated images, although often at the expense of traditional metrics (PSNR or SSIM) [19].

State-of-the-art methods in perceptual super-resolution have traditionally relied on generative adversarial networks (GANs), with notable models such as SRGAN [21], ESRGAN [46], Real-ESRGAN [47], and BSRGAN [53] demonstrating visual quality improvements. Recently, diffusion-based models have emerged as strong alternatives. Notably, SR3 [38] employs denoising diffusion probabilistic models (DDPMs) to iteratively refine images, while latent diffusion models (LDMs) [36] improve efficiency by operating in a compressed latent space. These score-based generative methods achieve superior perceptual quality but remain computationally demanding, limiting their suitability for real-time or resource-constrained applications.

While GAN-based methods are generally more efficient than diffusion models, their computational demands still pose challenges for deployment on mobile and edge devices, where low latency and limited hardware resources are critical. In contrast, PSNR-oriented super-resolution methods have benefited from extensive research and optimization [8, 24, 26, 34, 35, 41, 52], successfully pushing the boundaries of efficiency and performance.

Despite these advances, the development and benchmarking of efficient perceptual super-resolution methods remain largely unexplored. This study and benchmark aims to bridge the divide between visual quality and efficiency.

Related Challenges This challenge is one of the AIM 2025 ¹ workshop associated challenges on: high FPS non-uniform motion deblurring [6], rip current segmentation [11], inverse tone mapping [43], robust offline video

^{*}B. Longarela (brulon@cidaut.es, Cidaut AI, Spain) and M. V. Conde (marcos.conde@uni-wuerzburg.de, Cidaut AI and University of Würzburg) are the corresponding authors.

AIM 2025 webpage: <https://cvlai.net/aim/2025>.

Code: <https://github.com/brulonga/AIM-2025-EPSR-Challenge>

PSR4K: <https://drive.google.com/file/PSR4K-LR.tar.gz>

¹<https://www.cvlai.net/aim/2025/>

super-resolution [20], low-light raw video denoising [50], screen-content video quality assessment [37], real-world raw denoising [23], perceptual image super-resolution [31], efficient real-world deblurring [12], 4K super-resolution on mobile NPUs [15], efficient denoising on smartphone GPUs [17], efficient learned ISP on mobile GPUs [16], and stable diffusion for on-device inference [18]. Descriptions of the datasets, methods, and results can be found in the corresponding challenge reports.

2. Efficient Perceptual Super Resolution Benchmark

The goals of the proposed study and benchmark are: (i) to improve the state of the art in perceptual super-resolution by encouraging the development of models that balance high visual quality with computational efficiency, (ii) to provide a standardized benchmark and platform where diverse approaches can be rigorously compared under consistent efficiency and quality constraints, and (iii) to foster collaboration and knowledge exchange between academic researchers and industry professionals, accelerating progress toward deployable, real-time perceptual SR solutions. This section presents an in-depth description of the challenge.

2.1. Datasets

Training Datasets In this challenge, participants were free to choose their training datasets. The most commonly used datasets among participants were DIV2K [2], Flickr2K [28], LSDIR [25], and OST [45] (see training details in Section 4).

- DIV2K: 800 high-quality images at 1K-2K resolution.
- Flickr2K: 2,650 images at 1K-2K resolution, typically used alongside DIV2K for super-resolution training.
- LSDIR: 86,991 high-quality images at 1K-2K resolution.
- OutdoorSceneTrain (OST): 10,324 images at 1K or 2K resolution. Originally introduced as a segmentation dataset by Wang et al. (2017) [45], later repurposed for super-resolution in Real-ESRGAN [47].

The numbers above refer to the training splits. Some datasets include official validation sets (e.g., DIV2K with 100 images and LSDIR with 1,000 images), while others do not. For instance, Flickr2K is typically used only for training, and OST contains only a test split of 300 images.

The degradation pipelines applied during training were not fixed, each method uses slightly different variants based on Real-ESRGAN [47] degradation pipeline. The down-scaling factor was set to $\times 4$.

Testing Datasets We evaluate diverse methods using our novel dataset **PSR4K** and no-reference image quality assessment (NR-IQA). This dataset consists of 500 low-resolution (LR) images at 960 $\times$ 540 pixels, grouped into ten



Figure 1. We show two samples of high-resolution (HR) and low-resolution (LR) images. For memory and layout considerations in this document, the HR images have been down-scaled to match the spatial dimensions of the LR images. Consequently, the visual differences between HR and LR examples may appear less pronounced in this figure. It should be noted, however, that the original LR images in the dataset are relatively large in resolution.

categories: animals, architecture, art, food, nature, objects, portraits, sports, text, and urban scenes. For each category, five different degradations were applied, involving various down-sampling methods, blurs, and JPEG compressions. The exact degradation pipeline remains private to ensure the integrity of the benchmark. The chosen input resolution produces ultra-high-definition (UHD) outputs (3840 $\times$ 2160 pixels) at $\times 4$ scaling. You can see some examples of our dataset and degradations in Figure 1.

Additionally, methods were tested on existing perceptual SR NR-IQA benchmark datasets:

- PIPAL validation dataset: 1,000 images (288 $\times$ 288 pixels) with 40 types of degradations, including GAN-based degradations. [13].
- DIV2K-LSDIR validation dataset: the 100 DIV2K validation images combined with 100 LSDIR validation images, degraded using only bicubic down-sampling [2, 25].
- RealSR validation dataset which consists of 100 images exhibiting real-world degradations commonly used

in perceptual super-resolution benchmarks [5].

- RealSRSet: 20 images with complex degradations. [53].
- Real47: 47 images with complex degradations.[29].

2.2. Preliminaries

Baseline The Real Enhanced Super-Resolution GAN (Real-ESRGAN) [47] is adopted as the baseline. It models complex real-world degradations using a second-order pipeline that repeatedly applies blur, noise, resizing, and compression, including sinc filtering. The generator uses ESRGAN’s residual-in-residual dense blocks (RRDB) with pixel-unshuffle for efficiency, while a U-Net discriminator with spectral normalization stabilizes GAN training and improves per-pixel feedback. Training is performed in two stages: PSNR-oriented pretraining with L1 loss, followed by fine-tuning with L1, perceptual, and adversarial losses. Ground-truth sharpening is applied during training to balance sharpness and artifact suppression.

Efficiency constraints The efficiency limits are fixed at 5M parameters ($\approx 30\%$ of Real-ESRGAN) and 2000 GFLOPs ($\approx 22\%$ of Real-ESRGAN), measured for an input size of 960×540 pixels. All proposed methods must meet these computational constraints. No restrictions are imposed on memory footprint or inference time.

Metric For the ranking we adopted a scoring methodology inspired by the approach used in the NTIRE 2024 ESR and NTIRE 2025 ESR Challenges [34, 35]. Our formulation aims to aggregate multiple perceptual metrics into a single score relative to the baseline. For evaluation metrics where lower values indicate better performance, the score is computed as:

$$Score = \sum \lambda_i \cdot e^{\left(\frac{Metric_i}{Metric_{baseline}^i}\right)}$$

Conversely, for metrics where higher values indicate better performance, the score is defined as:

$$Score = \sum \lambda_i \cdot e^{\left(\frac{Metric_{baseline}^i}{Metric_i}\right)}$$

The evaluation metrics used are Perceptual Index (PI) [4], CLIP Image Quality Assessment (CLIQQA) [44], and Multi-Dimension Attention Network for No-Reference Image Quality Assessment (MANIQA) [51]. Among commonly used metrics, we selected those exhibiting the highest Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank-Order Correlation Coefficient (SRCC) [39]. The weighting coefficients were set as $\lambda_{PI} = 0.5$, $\lambda_{CLIQQA} = 0.25$, and $\lambda_{MANIQA} = 0.25$, reflecting a balanced contribution between traditional non-deep learning NR-IQA metrics and deep learning-based metrics. All

metrics were computed using the PIQA package².

While no standard quantitative measure for hallucinations or artifacts is available, a qualitative analysis will be conducted in a subsequent stage.

Training and validation phase The training phase lasted eight weeks. Due to Codabench’s lack of GPU support for NR-IQA metrics, participants were provided with the validation code to assess their progress locally. Validation was conducted on the RealSRSet and Real47 datasets.

Testing phase The test phase lasted one day. Participants were required to submit their code, a factsheet, and the output images for RealSRSet and Real47 to the organizers. The organizers verified the results by executing the submitted code under controlled conditions.

3. Experimental Results

We show in Table 1 the initial results of the study.

Besides the methods proposed in this paper, we included (i) BSRGAN [53] to assess the influence of training-time degradation modeling on test performance, and (ii) the top two entries from the NTIRE 2024 Efficient Super-Resolution Challenge [34, 42]—SPAN (XiomiMM) and R2NET (Cao Group)—as strong PSNR-oriented, efficiency-focused baselines.

Table 1 reports both efficiency statistics (FLOPs and parameter counts) and perceptual quality indicators. Perceptual performance is evaluated from four complementary perspectives: PI (lower is better), CLIPIQA (higher is better), MANIQA (higher is better), and a scalar *Score*, computed relatively to the Real-ESRGAN baseline (lower is better). This score is not a normalized metric, but rather a direct comparative measure of overall perceptual performance against the baseline.

VPEG achieved the highest overall performance, ranking first in all three perceptual metrics. Compared to Real-ESRGAN, it reduced PI by 24.7%, increased CLIPIQA by 23.4%, and increased MANIQA by 19.4%, while using only $\sim 19.0\%$ of the parameters and $\sim 17.6\%$ of the FLOPs. These results demonstrate that substantial gains in perceptual quality can be achieved within a highly constrained efficiency budget.

MiAlgo ranked second, delivering perceptual improvements comparable to VPEG: PI reduced by 9.7%, CLIPIQA increased by 13.2%, and MANIQA increased by 11.5% over the baseline, with $\sim 21.1\%$ of the parameters and $\sim 21.4\%$ of the FLOPs. The final scores for VPEG and MiAlgo were 2.2015 and 2.4512, respectively, indicating closely matched performance.

²<https://github.com/francois-rozet/piqa>

Table 1. **Summary of results for EPSR**. The best and second-best results are highlighted in **bold** and underlined, respectively. All metrics were obtained using the official evaluation code available at <https://github.com/brulonga/AIM-2025-EPSR-Challenge>.

Team Name	Params↓ (M)	FLOPs↓ (G)	PI↓	CLIPQA↑	MANIQA↑	Score	Rank
Real-ESRGAN (baseline)	16.6980	9293.9416	4.1442	0.5302	0.3283	2.7182	-
VPEG	3.1684	1631.0842	3.1205	0.6544	0.3919	2.2015	1
MiAlgo	3.5214	1987.3922	<u>3.7420</u>	<u>0.5999</u>	<u>0.3662</u>	<u>2.4512</u>	2
IPIU	0.2762	132.1431	6.0676	0.3951	0.2722	3.9536	3
BSRGAN	16.6980	9293.9416	4.2112	0.5779	0.3350	2.6731	-
SPAN	0.1507	77.7870	6.1198	0.3996	0.2748	3.9571	-
R2NET	0.2148	103.2455	6.6837	0.3750	0.2811	4.3401	-

The third-ranked method, **IPIU** (EFDN; winner of the NTIRE 2023 ESR Challenge [25]), is extremely lightweight ($\sim 1.65\%$ of the baseline parameters and $\sim 1.42\%$ of the FLOPs). However, as a distortion-oriented architecture primarily optimized for PSNR, its design is not fully aligned with the perceptual objectives of this track. We nonetheless consider it a relevant comparison point, as it illustrates the trade-off between distortion-focused optimization and perceptual quality under strict efficiency constraints.

Among the additional baselines, **SPAN** and **R2NET** [35] show the high efficiency and PSNR performance characteristic of their original challenge context, but obtain comparatively low perceptual scores. These methods illustrate the extreme efficiency achievable in ESR. For efficient perceptual SR, this can be seen as a practical upper bound in FLOPs and parameter count, which is difficult to match while also maximizing perceptual quality.

Finally, **BSRGAN** [53] slightly outperforms Real-ESRGAN on the PSR4K dataset, both in PI and CLIPQA, while maintaining identical efficiency. This reinforces the importance of degradation modeling choices in determining perceptual outcomes, even when the network complexity remains constant.

3.1. Extended Evaluation on Standard Perceptual SR Benchmarks

To assess the generalization capabilities of the proposed and reference methods beyond the proposed PSR4K dataset, we evaluated all models on five widely adopted benchmarks for perceptual super-resolution. These datasets vary in content diversity, degradation characteristics, and difficulty, providing a comprehensive view of cross-dataset performance. Both perceptual quality metrics and runtime measurements are reported, enabling a joint analysis of visual fidelity and computational efficiency under diverse conditions.

As described in Section 2.1, we tested all methods on the following datasets: DIV2K-LSDIR validation (Table 2), PIPAL validation (Table 3), RealSR validation (Table 4),

Table 2. Results on the DIV2K-LSDIR validation dataset.

DIV2K-LSDIR Validation Dataset				
Team Name	PI↓	CLIPQA↑	MANIQA↑	Runtime ¹ (ms)
Real-ESRGAN (baseline)	3.4401	0.5919	0.4082	118.6400
VPEG	3.0813	<u>0.6426</u>	<u>0.4273</u>	35.7096
MiAlgo	<u>3.3829</u>	0.6790	0.4629	58.2874
IPIU	5.3896	0.5000	0.3457	11.1416
BSRGAN	3.5726	0.5963	0.4002	98.0291
SPAN	5.4637	0.5054	0.3505	4.2781
R2NET	6.1687	0.4935	0.3308	<u>5.1796</u>

Real47 (Table 5), and RealSRSet (Table 6).

Across all datasets, VPEG consistently achieves the best PI values, significantly improving over the Real-ESRGAN baseline. Specifically, VPEG reduces the PI by approximately 26.5% and 30% on the PIPAL and RealSR datasets, respectively, demonstrating substantial gains under real-world and GAN-based degradations. In contrast, MiAlgo leads in CLIPQA and MANIQA (with the exception of RealSRSet), achieving improvements of roughly 34% and 28%, respectively, on the PIPAL dataset.

Traditional ESR methods generally underperform compared to perceptual SR solutions, particularly on PIPAL and RealSR, which contain more challenging degradations. These methods perform best on the DIV2K-LSDIR dataset, which features controlled bicubic downsampling (the degradation they were trained on). Among ESR methods, SPAN and EFDN show comparatively better results, surpassing R2NET.

For traditional perceptual SR methods such as Real-ESRGAN and BSRGAN, performance is largely consistent with expectations. Most datasets show similar results,

¹The reported runtimes correspond to the average execution time obtained by running the provided evaluation code on an NVIDIA H100 80GB HBM3 GPU. Performance differences between Real-ESRGAN and BSRGAN can be attributed to variations in their original implementations (Real-ESRGAN project and BSRGAN project).

Table 3. Results on the PIPAL validation dataset.

PIPAL Validation Dataset				
Team Name	PI↓	CLIPQA↑	MANIQA↑	Runtime ¹ (ms)
Real-ESRGAN (baseline)	4.1254	0.4576	0.2783	69.0464
VPEG	3.0366	0.6125	<u>0.3467</u>	21.7737
MiAlgo	<u>3.4911</u>	<u>0.5885</u>	0.3563	36.5912
IPIU	6.5827	0.4199	0.2317	3.4071
BSRGAN	3.8208	0.5154	0.2886	64.9269
SPAN	6.5316	0.4019	0.2342	1.5343
R2NET	6.7858	0.3806	0.2113	<u>1.8262</u>

Table 4. Evaluation results were obtained on the RealSR validation dataset. Due to the large size of some images, certain images could not be processed. In total, only 58 images were successfully processed by all models; therefore, caution should be exercised when interpreting these results. It should be noted that the input images in this dataset are of a resolution ranging from 1K to 2K.

RealSR Validation Dataset				
Team Name	PI↓	CLIPQA↑	MANIQA↑	Runtime ¹ (ms)
Real-ESRGAN (baseline)	4.6645	<u>0.6479</u>	<u>0.4050</u>	10555.9394
VPEG	3.2666	0.6115	0.3906	8638.6934
MiAlgo	<u>4.3695</u>	0.6932	0.4098	8619.2845
IPIU	10.3502	0.5230	0.2974	8374.7927
BSRGAN	5.7443	0.5288	0.3306	10465.1627
SPAN	10.3741	0.5288	0.2983	<u>8269.8468</u>
R2NET	10.1132	0.4700	0.2990	8074.7931

Table 5. Evaluation results on the Real47 dataset.

Real47 Dataset				
Team Name	PI↓	CLIPQA↑	MANIQA↑	Runtime ¹ (ms)
Real-ESRGAN (baseline)	3.5294	0.5999	0.3968	53.9262
VPEG	3.0307	<u>0.6444</u>	0.4107	17.2070
MiAlgo	<u>3.4506</u>	0.6771	0.4488	29.0713
IPIU	5.6026	0.5315	0.2809	10.6158
BSRGAN	3.5734	0.6042	0.3958	41.7053
SPAN	5.6719	0.5233	0.2816	3.6206
R2NET	6.2918	0.4599	0.3033	<u>7.0637</u>

Table 6. Results on the RealSRSet dataset.

RealSRSet Dataset				
Team Name	PI↓	CLIPQA↑	MANIQA↑	Runtime ¹ (ms)
Real-ESRGAN (baseline)	4.8358	0.5875	0.3807	78.4476
VPEG	4.0995	0.6635	0.4336	27.6264
MiAlgo	<u>4.3723</u>	0.6255	<u>0.4317</u>	40.4732
IPIU	6.0329	0.5397	0.3008	20.8097
BSRGAN	4.6087	<u>0.6388</u>	0.4110	51.1416
SPAN	6.0452	0.5166	0.3025	3.9676
R2NET	6.6692	0.5264	0.3027	<u>9.2148</u>

except for PIPAL (where BSRGAN benefits from GAN-degradation training) and RealSR, where Real-ESRGAN

Table 7. Comparison of VPEG and MiAlgo performance relative to Real-ESRGAN metrics for each dataset, including the score differences between the two methods.

Team Name	DIV2K-LSDIR	PIPAL	RealSR	Real47	RealSRSet
VPEG	2.5024	2.1294	2.4335	2.4712	2.3747
MiAlgo	2.5383	2.2554	2.5840	2.5407	2.4783
Difference	0.0359	0.1260	0.1505	0.0695	0.1036

demonstrates greater robustness to complex real-world degradations.

Table 7 reports a score relative to Real-ESRGAN for each dataset, providing a direct comparison between VPEG and MiAlgo. VPEG outperforms MiAlgo across all datasets, although the margin is smaller than on the PSR4K dataset. The average difference between VPEG and MiAlgo on these benchmarks is 0.0971, compared to 0.2497 on PSR4K. On DIV2K-LSDIR, the gap is minimal (0.0359). This analysis shows that both solutions achieve the largest improvement over Real-ESRGAN on the PIPAL dataset.

In terms of runtime performance, although these measurements should be interpreted with caution¹, VPEG consistently demonstrates significantly greater efficiency, requiring less than half the runtime of Real-ESRGAN across all evaluated datasets, with the exception of RealSR. This underscores the substantial improvement in computational efficiency.

3.2. Per-Class Performance on the PSR4K Dataset

The PSR4K dataset comprises ten distinct semantic categories. Analyzing performance at the class level provides valuable insights into the strengths and limitations of each method. Class-wise evaluation reveals perceptual quality patterns that may be obscured in aggregated metrics, such as performance degradation in texture rich scenes or improvements in structured environments.

As shown in Table 8, several consistent trends emerge. The architecture category is the most favorable across all methods, likely due to the structured nature of these scenes and their prevalence in training datasets such as DIV2K, FLICKR2K, and LSDIR. Similarly, the animals and nature categories exhibit above-average performance, which may also be attributed to their frequent appearance in training data.

In contrast, the food category consistently yields the lowest scores. This under-performance is likely due to the rich textures and fine details of food imagery, combined with its under-representation in existing datasets. Urban and sports scenes also pose challenges, particularly for PSNR-oriented methods. Interestingly, metric correlations within these categories are less consistent, with some metrics favoring certain methods and classes while others penalize them.

Table 8. Results obtained for each class in the PSR4K test set. The best class for each model is marked in blue and the worst in red.

Team Name	Animals			Architecture			Art			Food			Nature		
	PI↓	CLIPQA↑	MANIQA↑	PI↓	CLIPQA↑	MANIQA↑	PI↓	CLIPQA↑	MANIQA↑	PI↓	CLIPQA↑	MANIQA↑	PI↓	CLIPQA↑	MANIQA↑
Real-ESRGAN	4.1044	0.5387	0.3254	3.4564	0.5727	0.3791	4.3428	0.5184	0.3009	4.9594	0.4307	0.2788	3.4804	0.5560	0.3139
VPEG	2.8712	0.6507	0.3635	3.1070	0.6550	0.4342	3.1056	0.6779	0.3852	3.4790	0.6187	0.3403	2.9538	0.6512	0.3702
MiAlgo	3.5588	0.5981	0.3567	3.3224	0.6604	0.4324	3.9396	0.5614	0.3347	4.3602	0.4829	0.2874	3.3214	0.6586	0.3707
IPIU	6.3996	0.4199	0.2923	5.4120	0.3815	0.2952	6.3712	0.3925	0.2690	6.1494	0.3894	0.2369	5.7130	0.4077	0.2850
BSRGAN	4.1098	0.5993	0.3199	3.7090	0.5750	0.3856	4.4004	0.5882	0.3132	4.5946	0.5495	0.2912	3.6012	0.5904	0.3101
SPAN	6.3950	0.4180	0.2952	5.4864	0.3943	0.3002	6.3970	0.3910	0.2703	6.2188	0.3908	0.2380	5.7620	0.4040	0.2864
R2NET	6.9088	0.3869	0.3033	6.1674	0.3935	0.3098	6.8458	0.3638	0.2682	6.8612	0.3361	0.2476	6.5912	0.3761	0.2855
	Objects			Portraits			Sports			Text			Urban		
	PI↓	CLIPQA↑	MANIQA↑	PI↓	CLIPQA↑	MANIQA↑	PI↓	CLIPQA↑	MANIQA↑	PI↓	CLIPQA↑	MANIQA↑	PI↓	CLIPQA↑	MANIQA↑
Real-ESRGAN	3.9726	0.5359	0.3353	4.3434	0.5189	0.3135	4.6794	0.5310	0.3252	4.5560	0.5576	0.3610	3.5492	0.5423	0.3495
VPEG	3.0282	0.6523	0.4038	3.0018	0.6726	0.3774	3.1226	0.6665	0.3931	3.6368	0.6566	0.4300	2.9012	0.6425	0.4221
MiAlgo	3.6186	0.6190	0.3786	3.7072	0.5855	0.3431	4.1018	0.5743	0.3513	4.2674	0.6326	0.4019	3.2238	0.6262	0.4051
IPIU	5.9424	0.4136	0.2724	6.2422	0.4062	0.2662	6.5130	0.3974	0.2570	6.2766	0.4132	0.2886	5.6584	0.3296	0.2593
BSRGAN	4.0920	0.5668	0.3415	4.2946	0.5970	0.3256	4.7294	0.5897	0.3397	4.8254	0.5795	0.3674	3.7566	0.5440	0.3566
SPAN	6.0010	0.4176	0.2748	6.3114	0.4147	0.2690	6.5666	0.4007	0.2599	6.3116	0.4257	0.2906	5.7486	0.3392	0.2633
R2NET	6.5486	0.3880	0.2831	6.8230	0.3829	0.2831	6.9994	0.3692	0.2659	6.7734	0.4265	0.2996	6.3162	0.3268	0.2646

Table 9. The mean, median, and standard deviation were computed for the set of metric values obtained for each class in the PSR4K test set. The best overall results per class are highlighted in blue, and the worst in red.

Team Name	PI↓			CLIPQA↑			MANIQA↑		
	Mean	Median	Standard Deviation	Mean	Median	Standard Deviation	Mean	Median	Standard Deviation
Real-ESRGAN	4.1444	4.2236	0.5269	0.5302	0.5373	0.0389	0.3283	0.3253	0.0294
VPEG	3.1207	3.0669	0.2486	0.6544	0.6537	0.0166	0.3920	0.3891	0.0307
MiAlgo	3.7421	3.6629	0.4080	0.5999	0.6085	0.0530	0.3662	0.3637	0.0414
IPIU	6.0678	6.1958	0.3682	0.3951	0.4018	0.0260	0.2722	0.2707	0.0184
BSRGAN	4.2113	4.2022	0.4335	0.5779	0.5838	0.0192	0.3351	0.3327	0.0288
SPAN	6.1198	6.2651	0.3522	0.3996	0.4023	0.0245	0.2748	0.2725	0.0189
R2NET	6.6835	6.7982	0.2716	0.3750	0.3795	0.0286	0.2811	0.2831	0.0197

Surprisingly, most models perform reasonably well on the text category. While PI tends to penalize blurry or imprecise lettering, CLIPQA and MANIQA achieve above-average scores, comparable to those in the architecture category. This suggests that these metrics prioritize global perceptual quality over fine-grained textual details.

The art, portraits, and objects categories tend to align closely with the overall average performance, showing neither significant advantage nor disadvantage.

In summary, the poor performance on food images can be attributed to both their complex textures and lack of representation in training datasets. Meanwhile, categories such as sports, urban scenes, and text present more intrinsic challenges for perceptual super-resolution, as they expose inconsistencies in metric behavior and highlight the limitations of current evaluation frameworks.

In addition to reporting the metric values per class, we

computed descriptive statistics (mean, median, and standard deviation) across the set of metrics obtained for each semantic category, as shown in Table 9. This analysis provides insight into the stability of model performance under varying content conditions. A lower standard deviation reflects greater robustness to scene variability, whereas higher variability may indicate sensitivity to specific visual patterns.

Once again, the VPEG solution emerges as the most robust method. It achieves a standard deviation of 0.2486 for the PI metric, significantly outperforming Real-ESRGAN (0.5269). Similarly, for CLIPQA, VPEG attains a standard deviation of 0.0166, compared to 0.0389 for Real-ESRGAN. Although VPEG’s standard deviation on the MANIQA metric is slightly higher than that of Real-ESRGAN and BSRGAN, it remains competitive.

Conversely, while the MiAlgo team delivers results comparable to VPEG in overall performance, their model ex-

hibits greater variability across categories. This suggests reduced robustness and increased sensitivity to diverse textures and semantic content. Overall, this analysis further reinforces the strength of the VPEG solution, not only as a top-performing approach, but also as a robust alternative that surpasses traditional methods such as Real-ESRGAN and BSRGAN

3.3. Qualitative Comparison Across Benchmarks

To complement the quantitative results, we present a visual comparison of representative image crops from each benchmark, processed by all evaluated models. Each crop is accompanied by its corresponding perceptual metric scores, enabling a direct correlation between numerical performance and perceived image quality. This comparison is illustrated in Figure 2.

It is important to note that the images have been rescaled and, in some cases, compressed for visualization purposes within the figure. As a result, certain visual enhancements introduced by the models may be partially lost. With this in mind, it becomes evident that non-perceptual methods, such as SPAN, R2NET, and EFDN, fail to deliver meaningful perceptual improvements. These models tend to produce results that are only marginally better than bicubic upsampling in terms of PI, while CLIPIQA and MANIQA scores often deteriorate. In many cases, these methods do not clearly surpass bicubic interpolation, which should be considered a baseline requirement.

In contrast, perceptual methods demonstrate clear qualitative improvements. Models such as Real-ESRGAN, BSRGAN, MiAlgo, and VPEG not only outperform bicubic upsampling visually, but also achieve significantly better scores across perceptual metrics. The VPEG solution continues to stand out as the top-performing approach, closely followed by MiAlgo, with Real-ESRGAN and BSRGAN occasionally surpassing them in specific samples (Real47 and RealSRSet samples, respectively).

However, this qualitative comparison also underscores a critical limitation of current perceptual metrics: their inability to effectively penalize hallucinations or artifacts. As observed in the RealSR and Real47 crops, both VPEG and MiAlgo introduce noticeable artifacts. In particular, the VPEG solution severely degrades the RealSR crop, producing results that are clearly flawed to the human eye. Yet, despite these issues, the perceptual metrics fail to reflect the degradation, assigning top-performing scores to these outputs. In contrast, traditional models such as Real-ESRGAN and BSRGAN appear more robust in these cases, with Real-ESRGAN delivering an exceptionally accurate reconstruction in the Real47 sample, free of hallucinations (or artifacts) and accompanied by the highest metric scores.

It is also worth highlighting that the VPEG and MiAlgo solutions excel in the PIPAL dataset, producing high-quality

reconstructions that align well with perceptual metrics. This observation supports the analysis presented in Section 3.1, where PIPAL was identified as the benchmark most responsive to perceptual improvements.

4. Methods

In the following section, we outline each contributor’s solution, all designed to satisfy our efficiency constraints, while maximizing perceptual super-resolution performance.

Note that the method descriptions were provided by each team as their contribution to this report.

4.1. VPEG

Spatially-Adaptive Feature Modulation for Efficient Perceptual Image Super-Resolution

Ke Wu, Long Sun, Lingshun Kong, Jinshan Pan, Jiangxin Dong, Jinhui Tang

Nanjing University of Science and Technology

Method The VPEG team uses SAFMN architecture [40] as the baseline model in their work. The original SAFMN open-source model exceeded the challenge efficiency constraints using 2888.23 GFLOPs.

To meet efficiency requirements without significantly compromising quality, the VPEG team proposed a reduced-complexity variant, SAFMN-L, which maintained 16 blocks but reduced the channel dimension from 128 to 96. The overall architecture is shown in Figure 3. They successfully created a perceptual version of SAFMN by incorporating Perceptual [19], LDL [27], GAN [47] and AESOP losses [22]. No pre-trained SR weights were used for fine-tuning; however, the AESOP pre-trained autoencoder was employed within the loss computation.

Training Details A three-stage training strategy was employed to progressively enhance performance:

1. Stage I: SAFMN-L was trained with 192×192 input patches, a batch size of 64, an initial learning rate of 3×10^{-4} decayed to 1×10^{-6} via cosine annealing, using a weighted combination of L1 loss (1.0) and FFT-based L1 loss (0.05) for 300k iterations using Adam.
2. Stage II: Used the same patch size, a batch size of 36, a learning rate schedule of 1×10^{-4} to 1×10^{-6} , and minimized L1 (1.0), Perceptual (0.1), LDL (1.0), and GAN (0.1) losses over 300k iterations.
3. Stage III: Retained the patch size, used a batch size of 16, and applied the same learning rate schedule while optimizing a combination of AESOP (1.0), Perceptual (0.1), LDL (1.0), and GAN (0.1) losses for 100k iterations.

Perceptual loss was defined using pre-activation conv1–conv5 feature maps from VGG19 (weights 0.1, 0.1,

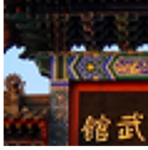
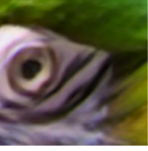
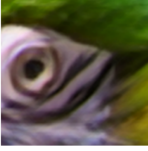
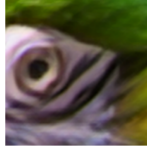
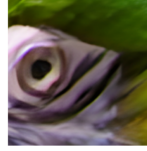
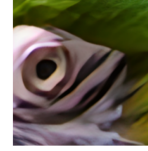
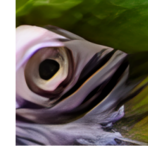
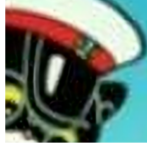
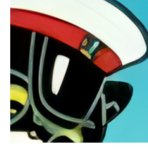
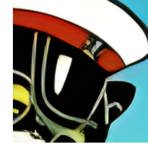
							
BICUBIC PI: 8.98 CLIPQA: 0.3193 MANIQA: 0.2096	SPAN PI: 7.62 CLIPQA: 0.1982 MANIQA: 0.1634	R2NET PI: 7.58 CLIPQA: 0.1816 MANIQA: 0.2187	IPIU PI: 7.43 CLIPQA: 0.1830 MANIQA: 0.1607	ESRGAN PI: 3.79 CLIPQA: 0.5836 MANIQA: 0.5063	BSRGAN PI: 4.57 CLIPQA: 0.6173, MANIQA: 0.4933	MiAlgo PI: 3.76 CLIPQA: 0.5810 MANIQA: 0.5162	VPEG PI: 3.33 CLIPQA: 0.6018 MANIQA: 0.5909
							
BICUBIC PI: 6.99 CLIPQA: 0.3342 MANIQA: 0.2731	SPAN PI: 5.39 ClipQA: 0.2357 MANIQA: 0.3554	R2NET PI: 5.89 ClipQA: 0.3052 MANIQA: 0.3393	IPIU PI: 5.77 ClipQA: 0.2869 MANIQA: 0.3628	ESRGAN PI: 3.14 ClipQA: 0.5548 MANIQA: 0.4531	BSRGAN PI: 2.91 ClipQA: 0.4773 MANIQA: 0.4522	MiAlgo PI: 2.68 ClipQA: 0.5875 MANIQA: 0.4872	VPEG PI: 2.50 ClipQA: 0.5863 MANIQA: 0.5245
							
BICUBIC PI: 9.61 CLIPQA: 0.2957 MANIQA: 0.1398	SPAN PI: 9.21 ClipQA: 0.2941 MANIQA: 0.1363	R2NET PI: 8.91 ClipQA: 0.2681 MANIQA: 0.1344	IPIU PI: 8.96 ClipQA: 0.3001 MANIQA: 0.1360	ESRGAN PI: 7.07 ClipQA: 0.2514 MANIQA: 0.1860	BSRGAN PI: 5.51 ClipQA: 0.3761 MANIQA: 0.1882	MiAlgo PI: 6.72 ClipQA: 0.3249 MANIQA: 0.2242	VPEG PI: 2.96 ClipQA: 0.6889 MANIQA: 0.4608
							
BICUBIC PI: 10.42 CLIPQA: 0.3742 MANIQA: 0.2545	SPAN PI: 10.05 ClipQA: 0.3773 MANIQA: 0.2583	R2NET PI: 9.90 ClipQA: 0.3153 MANIQA: 0.2596	IPIU PI: 10.02 ClipQA: 0.3761 MANIQA: 0.2573	ESRGAN PI: 4.24 ClipQA: 0.4884 MANIQA: 0.3322	BSRGAN PI: 6.26 ClipQA: 0.4188 MANIQA: 0.2993	MiAlgo PI: 3.81 ClipQA: 0.5943 MANIQA: 0.3572	VPEG PI: 4.11 ClipQA: 0.3658 MANIQA: 0.2349
							
BICUBIC PI: 7.16 CLIPQA: 0.5188 MANIQA: 0.3168	SPAN PI: 6.22 ClipQA: 0.4033 MANIQA: 0.3073	R2NET PI: 6.13 ClipQA: 0.4485 MANIQA: 0.3598	IPIU PI: 6.46 ClipQA: 0.3715 MANIQA: 0.2979	ESRGAN PI: 2.90 ClipQA: 0.6805 MANIQA: 0.5580	BSRGAN PI: 3.64 ClipQA: 0.5616 MANIQA: 0.4524	MiAlgo PI: 3.14 ClipQA: 0.6266 MANIQA: 0.5733	VPEG PI: 2.98 ClipQA: 0.6117 MANIQA: 0.5409
							
BICUBIC PI: 7.89 CLIPQA: 0.4879 MANIQA: 0.2837	SPAN PI: 6.85 ClipQA: 0.4264 MANIQA: 0.2089	R2NET PI: 6.99 ClipQA: 0.3409 MANIQA: 0.2806	IPIU PI: 6.86 ClipQA: 0.4342 MANIQA: 0.2070	ESRGAN PI: 6.81 ClipQA: 0.7717 MANIQA: 0.5170	BSRGAN PI: 5.45 ClipQA: 0.8285 MANIQA: 0.5228	MiAlgo PI: 6.06 ClipQA: 0.8060 MANIQA: 0.5709	VPEG PI: 6.72 ClipQA: 0.7968 MANIQA: 0.5460

Figure 2. **Qualitative comparison of super-resolution results across multiple datasets.** For each dataset, cropped image regions are arranged from left to right in approximate order of increasing perceptual quality, as indicated by the corresponding quantitative metrics. The corresponding datasets for each crop, from top to bottom, are: PSR4K, DIV2K-LSDIR validation set, RealSR validation set, PIPAL validation set, Real47, and RealSRSet. Images have been downsampled from their original upscaled resolution and compressed to meet compilation constraints, which may slightly affect their visual fidelity. Although the image displays 'ESRGAN', we are in fact referring to 'Real-ESRGAN'; the omission of 'Real' is purely for aesthetic purposes.

1, 1, 1). GAN loss employed a Spectral UNet discriminator, optimized with a CosineAnnealing scheduler (min LR $1e-6$). EMA strategy was applied throughout training.

All experiments were conducted with PyTorch on NVIDIA RTX 3090 GPUs, with a memory footprint of approximately 20–23 GB during training. Data preprocessing, augmentation, and training procedures followed BasicSR [48], and the total training duration was about eight days. The training datasets used were DIV2K and LSDIR, and degraded images were obtained following the Real-ESRGAN degradation pipeline.

4.2. MiAlgo

TinyESRGAN: A Lightweight ESRGAN Variant for Real-World Image Super-Resolution

Tianyu Hao, Yuxuan Qiu, Yueqi Yang,
Chaoyu Feng, Na Jiang, Dongqing Zou, Lei Lei

Xiaomi Inc.
Capital Normal University

Method The overall architecture of the MiAlgo team’s solution, illustrated in Figure 4, is based on the ESRGAN framework [46] and redesigned as a lightweight variant, named TinyESRGAN, to achieve efficient 4× image super-resolution. Several structural modifications were introduced by the MiAlgo team to reduce computational complexity while maintaining comparable perceptual performance to the baseline. Specifically, the number of Residual-in-Residual Dense Blocks (RRDBs) was reduced to 17, the number of intermediate feature channels in each RRDB was set to 32, and the channel growth rate within each dense block was set to 18. These adjustments resulted in an approximate 79% reduction in computational cost relative to the original ESRGAN architecture, thereby improving its suitability for deployment on resource-constrained platforms such as mobile and embedded devices.

Training Details The TinyESRGAN model was implemented in the PyTorch framework and trained on a single NVIDIA H20 GPU following a multi-stage strategy:

1. **Stage I:** Optimization with a combination of MSE loss (1.0) and LPIPS [54] perceptual loss (1.0) over 500,000 iterations, with a batch size of 32 and an initial learning rate of 3×10^{-4} , decayed via cosine annealing.
2. **Stage II:** Addition of a GAN loss (0.1) [47], with training continuing for an additional 250,000 iterations, keeping the batch size unchanged and reducing the learning rate to 1×10^{-4} .

The network was trained for 4× super-resolution, mapping 128×128 low-resolution inputs to 512×512 high-resolution outputs. Adam was used for optimization in

both stages. For GAN loss, the MiAlgo team followed the Real-ESRGAN setup with a Spectral-UNet discriminator, but without applying a learning-rate scheduler. Training employed high-resolution images from DIV2K, Flickr2K, and OST as ground truth, with realistic low-resolution counterparts generated via the Real-ESRGAN degradation pipeline. Training procedures relied on BasicSR [48], and EMA was applied throughout.

4.3. IPIU

Data Augmented Edge Distillation for Resource Efficient Image Super-Resolution

Lianping Lu, Heng Yang, Meilin Gao

Intelligent Perception and Image Understanding Lab,
Xidian University

Method The proposed solution is built upon the Edge-enhanced Feature Distillation Network (EFDN) [49], a lightweight yet high-performing super-resolution model that combines block design, neural architecture search, and tailored loss functions to achieve an optimal balance between reconstruction quality and computational efficiency. The architecture of the EFDN is presented in Figure 5. At its core, EFDN employs an Edge-Enhanced Diverse Branch Block (EDBB), which consolidates and extends existing reparameterization techniques into a versatile, multi-branch module that enhances both structural and high-frequency edge feature extraction [9, 10, 55]. Multiple reparameterizable branches capture complementary edge and texture cues, which are then fused into a standard convolution to preserve inference efficiency.

Training Details The model was implemented in PyTorch [33] and trained on a single NVIDIA RTX 3090 GPU using the Flickr2K dataset [28] as ground truth. For this challenge training was performed for 15 hours with a batch size of 64, optimizing an L1 loss with the Adam optimizer [1] and an initial learning rate of 1×10^{-3} , decayed using a cosine annealing schedule. Horizontal and vertical flipping were applied for data augmentation.

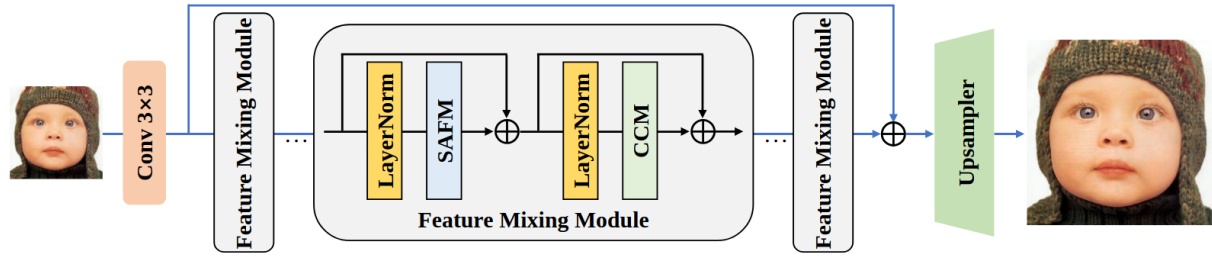


Figure 3. An overview of the proposed SAFMN-L in Section 4.1 by the VPEG team. SAFMN employs a series of feature mixing modules (FMMs) to process deep-level features. The FMM block is composed of a spatially-adaptive feature modulation (SAFM) and a convolutional channel mixer (CCM).

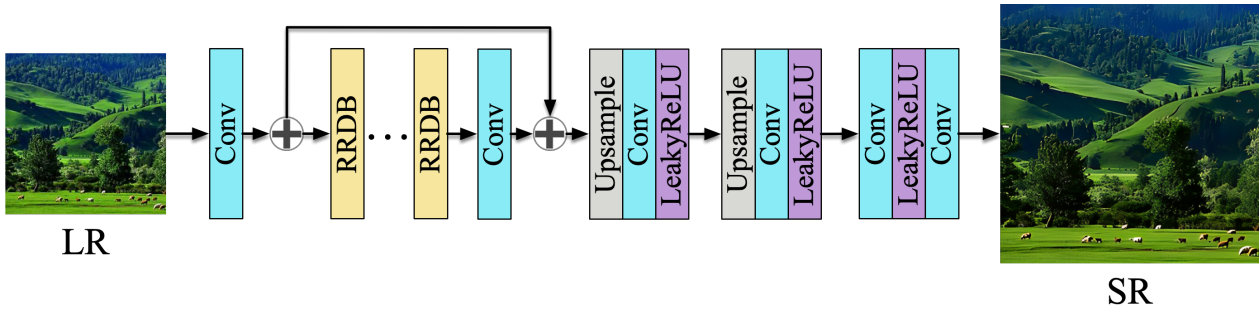


Figure 4. Overview of TinyESRGAN proposed in Section 4.2 by MiAlgo team.

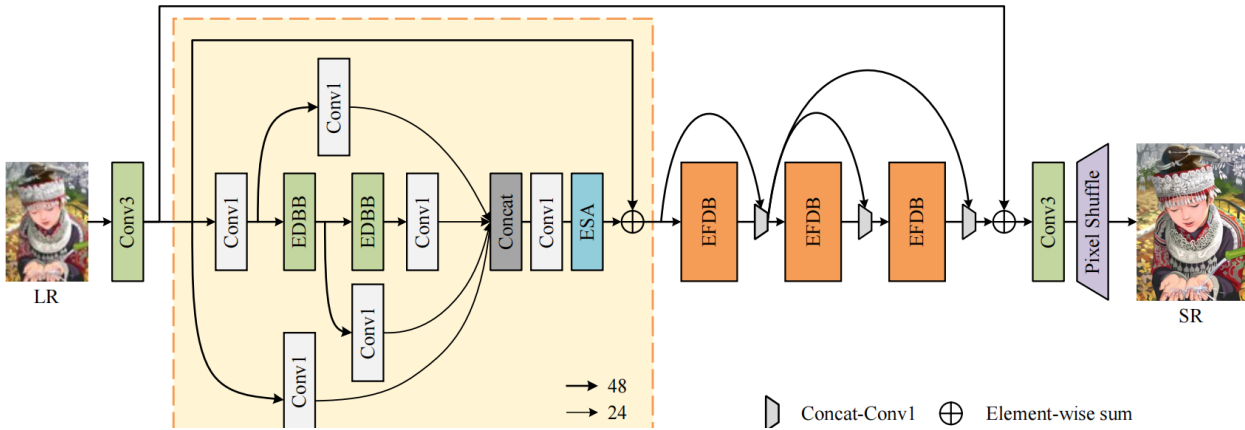


Figure 5. Overview of EFDN architecture presented by IPIU team in Section 4.3.

5. Conclusions

We conclude the following points from this study and the proposed benchmarks:

- The challenge introduced a new test set, **PSR4K**, divided into ten semantic categories to facilitate more fine-grained analysis. This dataset has proven valuable for in-depth research and, as its name suggests, consists of 4K-resolution images, aiming to establish a benchmark for 4K SR.
- The results corroborate that improving perceptual image quality while adhering to strict efficiency constraints is indeed possible, thereby opening the door to a relatively unexplored research direction.
- Despite improvements in perceptual quality metrics, the studied methods tend to produce visual artifacts, which raises questions about a possible efficiency-perception trade-off. Moreover, this emphasizes the need for new perceptual quality metrics that remain robust in the presence of artifacts or hallucinations.
- Notably, several widely used techniques for efficient super-resolution, such as knowledge distillation [14], re-parameterization (applied only by the IPIU team) [9, 10, 55], and pruning [30, 32], were not employed in this analysis, thereby leaving untapped potential for future exploration.

Acknowledgments

This work was partially supported by the Alexander von Humboldt Foundation. We thank the AIM 2025 sponsors: AI Witchlabs and University of Würzburg (Computer Vision Lab).

Cidaut AI thank Supercomputing of Castile and Leon (SCAYLE. Leon, Spain) for assistance with the model training and GPU resources.

References

- [1] Kingma DP, Ba J, Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6), 2014. 9
- [2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1122–1131, 2017. 2
- [3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 6228–6237. IEEE, 2018. 1
- [4] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018. 1, 3
- [5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model, 2019. 3
- [6] George Ciubotariu, Florin-Alexandru Vasluianu, Zhuyun Zhou, Nancy Mehta, Radu Timofte, et al. AIM 2025 high FPS non-uniform motion deblurring challenge report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 1
- [7] Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2SR: SwinV2 transformer for compressed image super-resolution and restoration. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2022. 1
- [8] Marcos V. Conde, Eduard Zamfir, Radu Timofte, Daniel Motilla, et al. Efficient deep models for real-time 4k image super-resolution. ntire 2023 benchmark and report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1495–1521, 2023. 1
- [9] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks, 2019. 9, 11
- [10] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Diverse branch block: Building a convolution as an inception-like unit, 2021. 9, 11
- [11] Andrei Dumitriu, Florin Miron, Florin Tatui, Radu Tudor Ionescu, Radu Timofte, Aakash Ralhan, Florin-Alexandru Vasluianu, et al. AIM 2025 challenge on rip current segmentation (RipSeg). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 1
- [12] Daniel Feijoo, Paula Garrido, Marcos Conde, Jaesung Rim, Alvaro Garcia, Sunghyun Cho, Radu Timofte, et al. Efficient real-world deblurring using single images: AIM 2025 challenge report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 2
- [13] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. Pipal: a large-scale image quality assessment dataset for perceptual image restoration, 2020. 2
- [14] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, 2020. 11
- [15] Andrey Ignatov, Georgy Perevozchikov, Radu Timofte, et al. 4K image super-resolution on mobile NPUs: Mobile AI & AIM 2025 challenge report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 2
- [16] Andrey Ignatov, Georgy Perevozchikov, Radu Timofte, et al. Efficient learned smartphone ISP on mobile GPUs: Mobile AI & AIM 2025 challenge report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 2
- [17] Andrey Ignatov, Georgy Perevozchikov, Radu Timofte, et al. Efficient image denoising on smartphone GPUs: Mobile AI & AIM 2025 challenge report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 2

- [18] Andrey Ignatov, Georgy Perevozchikov, Radu Timofte, et al. Adapting stable diffusion for on-device inference: Mobile AI & AIM 2025 challenge report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 2
- [19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 1, 7
- [20] Nikolai Karetin, Ivan Molodetskikh, Dmitry Vatolin, Radu Timofte, et al. AIM 2025 challenge on robust offline video super-resolution: Dataset, methods and results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 2
- [21] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network, 2017. 1
- [22] MinKyu Lee, Sangeek Hyun, Woojin Jun, and Jae-Pil Heo. Auto-encoded supervision for perceptual image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17958–17968, 2025. 7
- [23] Feiran Li, Jiacheng Li, Marcos Conde, Beril Besbinar, Vlad Hosu, Daisuke Iso, Radu Timofte, et al. Real-world raw denoising using diverse cameras: AIM 2025 challenge report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 2
- [24] Yawei Li, Kai Zhang, Radu Timofte, Luc Van Gool, Fangyuan Kong, Mingxi Li, Songwei Liu, Zongcai Du, Ding Liu, Chenhui Zhou, Jingyi Chen, Qingrui Han, Zheyuan Li, Yingqi Liu, Xiangyu Chen, Haoming Cai, Yu Qiao, Chao Dong, Long Sun, Jinshan Pan, Yi Zhu, Zhikai Zong, Xiaoxiao Liu, Zheng Hui, Tao Yang, Peiran Ren, Xuansong Xie, Xian-Sheng Hua, Yanbo Wang, Xiaozhong Ji, Chuming Lin, Donghao Luo, Ying Tai, Chengjie Wang, Zhizhong Zhang, Yuan Xie, Shen Cheng, Ziwei Luo, Lei Yu, Zhihong Wen, Qi Wu, Youwei Li, Haoqiang Fan, Jian Sun, Shuaicheng Liu, Yuanfei Huang, Meiguang Jin, Hua Huang, Jing Liu, Xinjian Zhang, Yan Wang, Lingshun Long, Gen Li, Yuanfan Zhang, Zuowei Cao, Lei Sun, Panaetov Alexander, Yucong Wang, Minjie Cai, Li Wang, Lu Tian, Zheyuan Wang, Hongbing Ma, Jie Liu, Chao Chen, Yidong Cai, Jie Tang, Gangshan Wu, Weiran Wang, Shirui Huang, Honglei Lu, Huan Liu, Keyan Wang, Jun Chen, Shi Chen, Yuchun Miao, Zimo Huang, Lefei Zhang, Mustafa Ayazoğlu, Wei Xiong, Chengyi Xiong, Fei Wang, Hao Li, Ruimian Wen, Zhi-jing Yang, Wenbin Zou, Weixin Zheng, Tian Ye, Yuncheng Zhang, Xiangzhen Kong, Aditya Arora, Syed Waqas Zamin, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Dandan Gao and Dengwen Zhou and Qian Ning, Jingzhu Tang, Han Huang, Yufei Wang, Zhangheng Peng, Haobo Li, Wenxue Guan, Shenghua Gong, Xin Li, Jun Liu, Wan-jun Wang, Dengwen Zhou, Kun Zeng, Hanjiang Lin, Xinyu Chen, and Jinsheng Fang. Ntire 2022 challenge on efficient super-resolution: Methods and results, 2022. 1
- [25] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhong Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Deman-dolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Ls-dir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1775–1787, 2023. 2, 4
- [26] Yawei Li, Yulun Zhang, Radu Timofte, Luc Gool, Lei Yu, Youwei Li, Li Xinpeng, Ting Jiang, Qi Wu, Mingyan Han, Wenjie Lin, Chengzhi Jiang, Jinting Luo, Haoqiang Fan, Shuaicheng Liu, Yucong Wang, Minjie Cai, Mingxi Li, Yuhang Zhang, and Xin Wang. Ntire 2023 challenge on efficient super-resolution: Methods and results, 2023. 1
- [27] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 7
- [28] 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. 2, 9
- [29] Xinqi Lin, Jingwen He, Ziyang Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diff-bir: Towards blind image restoration with generative diffusion prior, 2024. 3
- [30] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming, 2017. 11
- [31] Bruno Longarela, Marcos Conde, Álvaro García, Radu Timofte, et al. AIM 2025 perceptual image super-resolution challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 2
- [32] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference, 2017. 11
- [33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 9
- [34] Bin Ren, Yawei Li, Nancy Mehta, Radu Timofte, Hongyuan Yu, Cheng Wan, Yuxin Hong, Bingnan Han, Zhuoyuan Wu, Yajun Zou, Yuqing Liu, Jizhe Li, Keji He, Chao Fan, Heng Zhang, Xiaolin Zhang, Xuanwu Yin, Kunlong Zuo, Bohao Liao, Peizhe Xia, Long Peng, Zhibo Du, Xin Di, Wangkai Li, Yang Wang, Wei Zhai, Renjing Pei, Jiaming Guo, Songcen Xu, Yang Cao, Zhengjun Zha, Yan Wang, Yi Liu, Qing Wang, Gang Zhang, Liou Zhang, Shijie Zhao, Long Sun, Jinshan Pan, Jiangxin Dong, Jinhui Tang, Xin Liu, Min Yan, Qian Wang, Menghan Zhou, Yiqiang Yan, Yixuan Liu, Wensong Chan, Dehua Tang, Dong Zhou, Li Wang, Lu Tian, Barsoum Emad, Bohan Jia, Junbo Qiao, Yunshuai Zhou, Yun Zhang, Wei Li, Shaohui Lin, Shenglong Zhou, Binbin Chen, Jincheng Liao, Suiyi Zhao, Zhao Zhang, Bo Wang, Yan Luo, Yanyan Wei, Feng Li, Mingshen Wang, Yawei Li, Jinhan Guan, Dehua Hu, Jiawei Yu, Qisheng Xu, Tao Sun, Long Lan, Kele Xu, Xin Lin, Jingtong Yue, Lehan Yang, Shiyi Du, Lu Qi, Chao Ren, Zeyu Han, Yuhang Wang, Chaolin Chen, Haobo Li, Mingjun Zheng, Zhongbao Yang, Lianhong

- Song, Xingzhuo Yan, Minghan Fu, Jingyi Zhang, Baiang Li, Qi Zhu, Xiaogang Xu, Dan Guo, Chunle Guo, Jiadi Chen, Huanhuan Long, Chunjiang Duanmu, Xiaoyan Lei, Jie Liu, Weilin Jia, Weifeng Cao, Wenlong Zhang, Yanyu Mao, Ruilong Guo, Nihao Zhang, Qian Wang, Manoj Pandey, Maksym Chernozhukov, Giang Le, Shuli Cheng, Hongyuan Wang, Ziyang Wei, Qingting Tang, Liejun Wang, Yongming Li, Yanhui Guo, Hao Xu, Akram Khatami-Rizi, Ahmad Mahmoudi-Aznaveh, Chih-Chung Hsu, Chia-Ming Lee, Yi-Shiuan Chou, Amogh Joshi, Nikhil Akalwadi, Sampada Malagi, Palani Yashaswini, Chaitra Desai, Ramesh Ashok Tabib, Ujwala Patil, and Uma Mudanagudi. The ninth ntire 2024 efficient super-resolution challenge report, 2024. 1, 3
- [35] Bin Ren, Hang Guo, Lei Sun, Zongwei Wu, Radu Timofte, Yawei Li, Yao Zhang, Xinning Chai, Zhengxue Cheng, Yingsheng Qin, Yucai Yang, Li Song, Hongyuan Yu, Pufan Xu, Cheng Wan, Zhijuan Huang, Peng Guo, Shuyuan Cui, Chenjun Li, Xuehai Hu, Pan Pan, Xin Zhang, Heng Zhang, Qing Luo, Linyan Jiang, Haibo Lei, Qifang Gao, Yaqing Li, Weihua Luo, Tsing Li, Qing Wang, Yi Liu, Yang Wang, Hongyu An, Liou Zhang, Shijie Zhao, Lianhong Song, Long Sun, Jinshan Pan, Jiangxin Dong, Jinhui Tang, Jing Wei, Mengyang Wang, Ruilong Guo, Qian Wang, Qingliang Liu, Yang Cheng, Davinci, Enxuan Gu, Pinxin Liu, Yongsheng Yu, Hang Hua, Yunlong Tang, Shihao Wang, Yukun Yang, Zhiyu Zhang, Yukun Yang, Jiyu Wu, Jiancheng Huang, Yifan Liu, Yi Huang, Shifeng Chen, Rui Chen, Yi Feng, Mingxi Li, Cailu Wan, Xiangji Wu, Zibin Liu, Jinyang Zhong, Kihwan Yoon, Ganzorig Gankhuyag, Shengyun Zhong, Mingyang Wu, Renjie Li, Yushen Zuo, Zhengzhong Tu, Zongang Gao, Guannan Chen, Yuan Tian, Wenhui Chen, Weijun Yuan, Zhan Li, Yihang Chen, Yifan Deng, Ruting Deng, Yilin Zhang, Huan Zheng, Yanyan Wei, Wenxuan Zhao, Suiyi Zhao, Fei Wang, Kun Li, Yinggan Tang, Mengjie Su, Jae hyeon Lee, Dong-Hyeop Son, Ui-Jin Choi, Tiancheng Shao, Yuqing Zhang, Mengcheng Ma, Donggeun Ko, Youngsang Kwak, Jiun Lee, Jaehwa Kwak, Yuxuan Jiang, Qiang Zhu, Siyue Teng, Fan Zhang, Shuyuan Zhu, Bing Zeng, David Bull, Jing Hu, Hui Deng, Xuan Zhang, Lin Zhu, Qinrui Fan, Weijian Deng, Junnan Wu, Wenqin Deng, Yuquan Liu, Zhaohong Xu, Jameer Babu Pinjari, Kuldeep Purohit, Zeyu Xiao, Zhuoyuan Li, Surya Vashisth, Akshay Dudhane, Praful Hambarde, Sachin Chaudhary, Satya Naryan Tazi, Prashant Patil, Santosh Kumar Vipparthi, Subrahmanyam Murala, Wei-Chen Shen, I-Hsiang Chen, Yunzhe Xu, Chen Zhao, Zhizhou Chen, Akram Khatami-Rizi, Ahmad Mahmoudi-Aznaveh, Alejandro Merino, Bruno Longarela, Javier Abad, Marcos V. Conde, Simone Bianco, Luca Cogo, and Gianmarco Corti. The tenth ntire 2025 efficient super-resolution challenge report, 2025. 1, 3, 4
- [36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1
- [37] Nickolay Safonov, Mikhail Rakhmanov, Dmitriy Vatolin, Radu Timofte, et al. AIM 2025 challenge on screen-content video quality assessment: Methods and results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 2
- [38] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement, 2021. 1
- [39] Shaolin Su, Josep M. Rocafor, Danna Xue, David Serrano-Lozano, Lei Sun, and Javier Vazquez-Corral. Rethinking image evaluation in super-resolution, 2025. 3
- [40] Long Sun, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Spatially-adaptive feature modulation for efficient image super-resolution. In *ICCV*, 2023. 7
- [41] 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 (CVPR) Workshops*, 2017. 1
- [42] Cheng Wan, Hongyuan Yu, Zhiqi Li, Yihang Chen, Yajun Zou, Yuqing Liu, Xuanwu Yin, and Kunlong Zuo. Swift parameter-free attention network for efficient super-resolution, 2024. 3
- [43] Chao Wang, Francesco Banterle, Bin Ren, Radu Timofte, et al. AIM 2025 challenge on inverse tone mapping report: Methods and results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 1
- [44] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images, 2022. 3
- [45] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform, 2018. 2
- [46] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks, 2018. 1, 9
- [47] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data, 2021. 1, 2, 3, 7, 9
- [48] Xintao Wang, Liangbin Xie, Ke Yu, Kelvin C.K. Chan, Chen Change Loy, and Chao Dong. BasicSR: Open source image and video restoration toolbox. <https://github.com/XPiPixelGroup/BasicSR>, 2022. 9
- [49] Yan Wang. Edge-enhanced feature distillation network for efficient super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 777–785, 2022. 9
- [50] Alexander Yakovenko, George Chakvetadze, Ilya Khrapov, Maksim Zhelezov, Dmitry Vatolin, Radu Timofte, et al. AIM 2025 low-light raw video denoising challenge: Dataset, methods and results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 2
- [51] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment, 2022. 3
- [52] Eduard Zamfir, Marcos V. Conde, and Radu Timofte. Towards real-time 4k image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1522–1532, 2023. 1

- [53] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution, 2021. [1](#), [3](#), [4](#)
- [54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. [9](#)
- [55] Xindong Zhang, Hui Zeng, and Lei Zhang. Edge-oriented convolution block for real-time super resolution on mobile devices. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4034–4043, 2021. [9](#), [11](#)