QuARF: Quality-Adaptive Receptive Fields for Degraded Image Perception

Fei Gao[1,*], Ying Zhou[2,*], Ziyun Li[3], Wenwang Han[1], Jiaqi Shi[1], Maoying Qiao[4], Jinlan Xu[2], Nannan Wang[1,†]

[1] Xidian University, Xi'an 710126, China
[2] Hangzhou Dianzi University, Hangzhou 310018, China
[3] KTH Royal Institute of Technology, Stockholm 100 44, Sweden
[4] The University of Technology, Sydney, NSW 2007, Australia
{fgao, nnwang}@xidian.edu.cn, {yyzhou, jlxu}@hdu.edu.cn, liziyun2014@gmail.com, Maoying.Qiao@uts.edu.au

Abstract

Advanced *Deep Neural Networks* (DNNs) perform well for high-quality images, but their performance dramatically decreases for degraded images. Data augmentation is commonly used to alleviate this problem, but using too much perturbed data might seriously decrease the performance on pristine images. To tackle this challenge, we take our cue from the assumption of spatial coincidence in human visual perception, i.e. *multiscale and varying receptive fields are required for understanding pristine and degraded images*. Correspondingly, we propose a novel plug-and-play network architecture, dubbed *Quality-Adaptive Receptive Fields* (QuARF), to automatically select the optimal receptive fields based on the quality of the input image. To this end, we first design a multi-kernel convolutional block, which comprises multiscale continuous receptive fields. Afterward, we design a quality-adaptive routing network to predict the significance of each kernel, based on the quality features extracted from the input image. In this way, QuARF automatically selects the optimal inference route for each image. To further boost efficiency and effectiveness, the input feature map is split into multiple groups, with each group independently learning its quality-adaptive routing parameters. We apply QuARF to a variety of DNNs and conduct experiments in both discriminative and generation tasks, including semantic segmentation, image translation, and restoration. Thorough experimental results show that QuARF significantly and robustly improves the performance for degraded images, and outperforms data augmentation in most cases.

Code — https://github.com/AiArt-Gao/QuARF

1 Introduction

Deep Neural Networks (DNNs) have led to remarkable progress in computer vision, from discriminative tasks, including image classification (He et al. 2016; Liu et al. 2022b), object detection (Redmon et al. 2016), and semantic segmentation (Ke et al. 2024), to generative tasks, e.g. unsupervised image generation (Zhu et al. 2017), text-to-image generation (Rombach et al. 2022) and multimodal conditioned generation (Isola et al. 2017; Zhang, Rao, and Agrawala 2023). Nevertheless, despite these advancements, a critical challenge remains: *the performance of DNNs tends to deteriorate when faced with degraded input images.* As illustrated in Fig. 1, the performance of advanced DNNs declines rapidly with increased Gaussian blur, in either image understanding or generation tasks. Such observations expose a critical weakness in models that excel only with high-quality inputs (Fan et al. 2023). To alleviate this problem, existing works commonly use data augmentation during training, by randomly perturbing an input image (Zhong et al. 2020; Xu et al. 2023). Although data augmentation boosts the robustness of DNNs on degraded samples, it inversely damages the performance on pristine images (Rebuffi et al. 2021). Since images obtained in practical application scenarios are spread in a wide range of degradations, it's significant to develop degradation-robust DNN architectures.

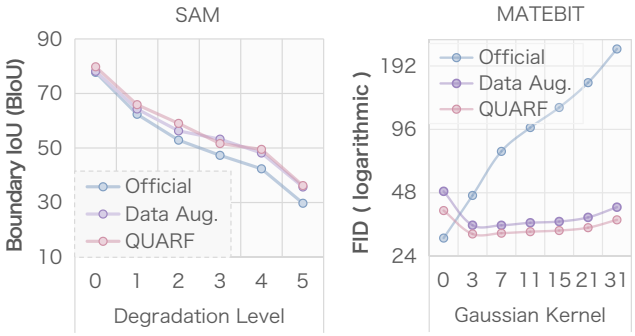

Figure 1: Performance of SAM (semantic segmentation) (Kirillov et al. 2023a) and MATEBIT (image-to-image translation) (Jiang et al. 2023), on degraded images. The x-axis indicates the degree of hybrid degradation, or Gaussian blurring kernel size; 0 indicates the pristine input; and larger values indicate lower quality. Higher BIoU or lower FID values indicate better performance.

To find the reasons why the quality of an input image heavily affects the performance of a network, inspired by *the spatial coincidence assumption* (Marr 2010), we use a group of *Difference-of-Gaussian* (DoG) kernels to filter a pristine and its degraded versions (Fig. 2). DoG can detect multiscale zero-crossings in an image, which approximate the observed psychophysical effects in human vision (Marr 2010). As shown in Fig. 2, the resulting *raw primal sketches* provide

*These authors contributed equally.
†Corresponding Author.

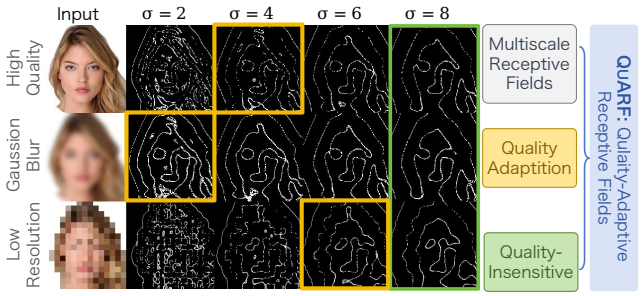

Figure 2: Illustration of the requirement of adaptive receptive fields. The input image is processed by the Difference-of-Gaussian (DoG) operator: $\mathrm{DoG}(\sigma, 2\sigma)$. Larger σ values indicate larger receptive fields. The yellow and green boxes present similar zero-cross segments, respectively.

essential and diverse information for understanding an image. The coincidence and diversity between adjacent scales indicate the presence of a specific physical phenomenon, e.g. the primary boundary, reflectance, or depth (Marr and Hildreth 1980). Thus, multiscale *continuous* receptive fields are necessary for visual understanding. In addition, to extract the same (or similar) visual information from images with different degradation, the required receptive fields may be diverse (yellow boxes). In other words, the receptive fields for inference should *adapt to the quality* of the input image. Besides, *large kernels* usually produce consistent representations, which are insensitive to the degradation in an image (green boxes). Thus, *a robust network should contain multiple kernels with continuous receptive fields (from small, medium, to large), and automatically select the optimal inference route based on the quality of input.*

Numerous advanced DNNs use multiscale kernels (as summarized in Fig. 3), and present performance or robustness improvement, in diverse tasks. However, they typically use a group of small kernels (Chollet 2017; Guo et al. 2022; Zhang et al. 2022), or small & large kernels (Yu and Koltun 2015; Li et al. 2023; Cui, Ren, and Knoll 2024). Few works propose using continuous kernels (He et al. 2022). In addition, researchers have proposed some degradation-aware networks for the image restoration or *super-resolution reconstruction* (SR) task (Zheng et al. 2023; Xie et al. 2024). These works first estimate the possible degradation presented by the input image; and then use the degradation-aware features to learn the dynamic convolutional kernels (Gu et al. 2019; Zhou et al. 2022b; Wang et al. 2021a) or synthesizing degraded images (Chen et al. 2024). However, few works try to solve the processing of degraded images been constructed for general computer vision tasks, such as classification, segmentation, generation. Besides, to our best knowledge, existing works haven't considered the significance of quality in receptive fields selection.

Based on the discussions above, in this work, we propose a novel *plug-and-play* network architecture, dubbed *Quality-Adaptive Receptive Fields* (QUARF). Specifically, QUARF mainly comprises two modules: (1) a multi-kernel convolutional block with multiscale continuous receptive fields; and (2) a quality-adaptive router, which predicts the

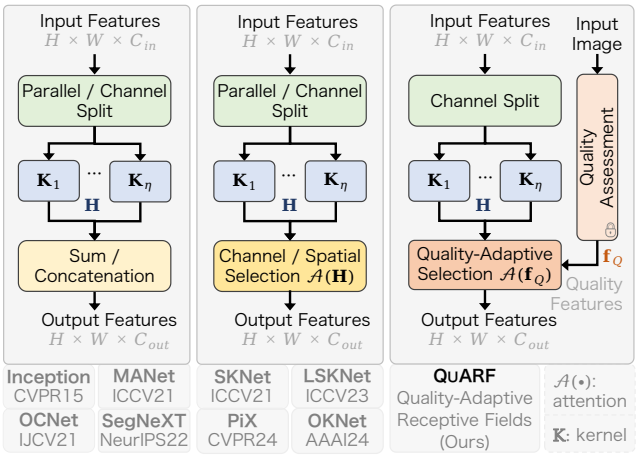

Figure 3: Differences between the proposed QUARF and existing multiscale kernel networks.

optimal inference path, based on the quality features of the input image. To further enhance the efficiency and effectiveness, the input feature map is split into multiple groups, with each group independently learning its quality-adaptive selection weights. We apply QUARF to a variety of DNNs for diverse computer vision tasks, i.e., semantic segmentation, image-to-image translation, style transfer, and restoration. Thorough experimental results show that QUARF significantly and robustly improves the segmentation accuracy and the generation quality for degraded images, and outperforms data augmentation. Our code have been released online.

2 Related Works

Multiscale & Large Kernel Networks have been explored for achieving multiscale receptive fields (Szegedy et al. 2015a; Yu and Koltun 2015; Chollet 2017; Zhang et al. 2022; Guo et al. 2022) for the classification or segmentation tasks. Besides, recent works propose to use large kernels (Liu et al. 2022b; Li et al. 2023; Liu et al. 2022a; Ding et al. 2022), or even global kernels (Cui, Ren, and Knoll 2024), to model long-range dependence. These networks typically use a residual connection or a small kernel for providing local information. The methodologies of existing works are summarized in Fig. 3. The input features might be parallelly fed into the multi-branch convolutional block; or channel-wise split for flexibility (Zhang et al. 2022; Kumar et al. 2024). The multi-branch outputs are merged through addition (or average) (Szegedy et al. 2015b), concatenation (Yuan et al. 2021), or attentively weighted selection (Li et al. 2023). Different from these works, we propose to use multiscale kernels with continuous receptive fields at one layer. Besides, we propose to select the optimal receptive fields based on the quality of input image.

Blind Image Quality Assessment (BIQA) determines an image's quality without referencing a high-quality version. Traditional BIQA techniques extract hand-crafted features from the spatial (Mittal, Soundararajan, and Bovik 2012; Mittal, Moorthy, and Bovik 2012), frequency (Moorthy and Bovik 2011), or wavelet coefficients (Gao et al. 2013), us-

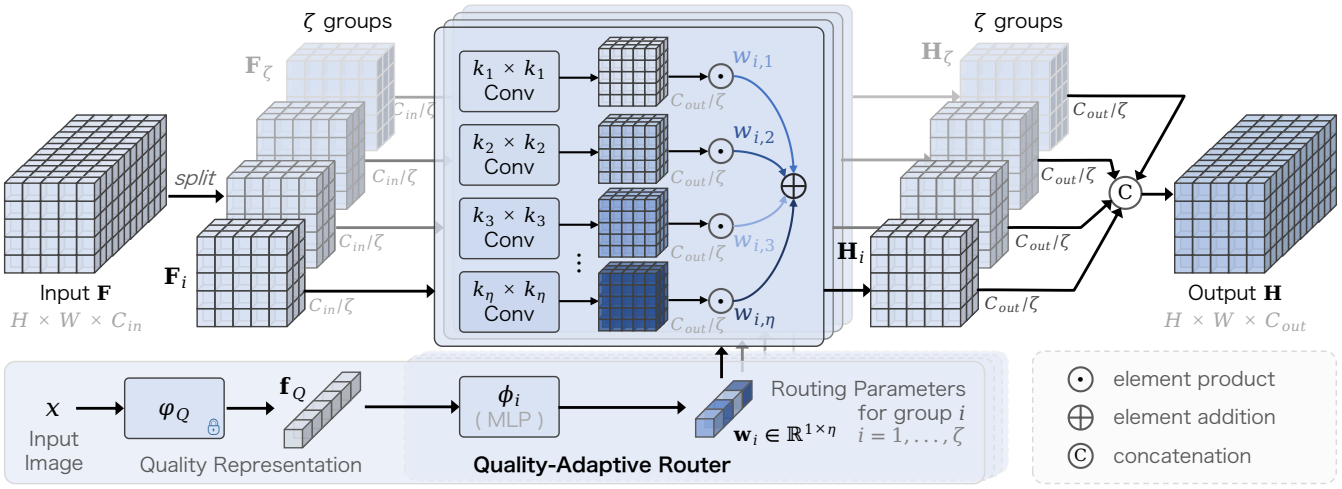

Figure 4: Pipeline of *Quality-Adaptive Receptive Fields* (QUARF). **(1) Channel Split.** The input tensor is channel-wisely split into ζ groups. **(2) Multiscale Receptive Fields.** Afterward, each group is separately fed into η branches of convolutional layers, with kernel sizes $k_j \times k_j, j = 1, ..., \eta$. **(3) Quality-Adaptive Routing.** For each group, we learn the optimal routing parameters, based on the perceptual quality of the input image x, to integrate the multi-kernel outputs. In the implementation, we extract the quality representation $\mathbf{f}_Q$ by using an *blind image quality assessment* (BIQA) method ψ_Q, and maps $\mathbf{f}_Q$ to the routing parameters through *multilayer perceptrons* (MLP). **(4) Merging.** Finally, the outputs of all the groups are channel-wisely concatenated.

ing natural scene statistics (NSS) as a foundation. In recent years, BIQA methods based on DNNs (Su et al. 2020; Qin et al. 2023; Ke et al. 2021) or vision-language contrastive learning (Zhang et al. 2023; Shi, Gao, and Qin 2024), have achieved remarkable advancements. In this work, we use existing BIQA methods to extract quality features, for adaptively selecting the optimal receptive fields.

3 Method

3.1 Overview

Let $\mathbf{F} \in \mathbb{R}^{H \times W \times C_{in}}$ be the input feature map, and $\mathbf{H} \in \mathbb{R}^{H \times W \times C_{out}}$ be the output. H, W, C_{in}, and C_{out} are the height, width, the number of input channels, and the number of output channels, sequentially. The pipeline of our multi-kernel network architecture, i.e. *Quality-Adaptive Receptive Fields* (QUARF), is illustrated in Fig. 4. Similar to the pipeline of previous multiscale or large kernel networks (Fig. 3), QUARF mainly contains the following four stages:

(1) Channel Split The input tensor $\mathbf{F}$ is channel-split into ζ groups: $\mathbf{F}_i \in \mathbb{R}^{H \times W \times C_{in}/\zeta}, i = 1, ..., \zeta$.

(2) Multiscale Receptive Fields (MRF) Each group $\mathbf{F}_i$ is fed into a multi-branch convolutional block, with η kernels: $\mathbf{K}_j \in \mathbb{R}^{k_j \times k_j \times C_{in} \times C_{out}/\eta}, j = 1, ..., \eta$, where k_j is the kernel width. Each kernel produces an output tensor:

$$\mathbf{Z}_{i,j} \in \mathbb{R}^{H \times W \times C_{out}/\eta} = \mathbf{K}_{i,j} \otimes \mathbf{F}_i, \tag{1}$$

where $\otimes$ denotes the convolutional operation.

(3) Quality-Adaptive Routing For each group, the significance of each receptive field (i.e. kernel), is adaptively predicted based on the perceptual quality of the input image x. The output of the i-th group, denoted by $\mathbf{H}_i$, is a weighted combination of $\{\mathbf{Z}_{i,j}\}_{j=1}^{\eta}$.

(4) Merging Finally, the outputs of all the groups are channel-wisely concatenated as the output of QUARF.

In the following part, we'll detail the implementation of multiscale receptive fields and quality-adaptive routing.

3.2 Multiscale Receptive Fields (MRF)

According to the spatial coincidence assumption (Marr and Hildreth 1980), it's better to design multiscale continuous receptive fields. By "continuous" it means the kernels should be "*well separated in the frequency domain and covering an adequate range of the frequency spectrum*" (Marr 2010). Following such instructions, we define the size k_j of $\mathbf{K}_j$ by:

$$k_1 = 3, \quad \text{and} \quad k_j = 2k_{j-1} - 1, \forall j = 2, ..., \eta. \tag{2}$$

We set $\eta = 4$ in default unless otherwise specified.

Computational Complexity To reduce the computational complexity, we use dilated convolution (Yu and Koltun 2015), instead of the standard 2D Covolution, to achieve a receptive field of $k_j \times k_j$ by using 3×3 kernels with a dilation of j. The amount of learnable kernel parameters here is computed by:

$$\mathcal{C}_{\text{MRF}} = \sum_{i=1}^{\zeta} \sum_{j=1}^{\eta} \frac{c_{in}}{\zeta} \cdot 3 \cdot 3 \cdot \frac{c_{out}}{\zeta} = \frac{9 \cdot c_{in} \cdot c_{out}}{\zeta}. \tag{3}$$

Obviously, the amount of parameters is inversely proportional to the number of groups ζ.

3.3 Quality-Adaptive Routing (QAR)

As previously discussed on Fig. 2, images with diverse quality (or degree of degradation), may require different sizes of receptive fields, for extracting the same presentations. Thus, we propose to learn the weighting parameters for every receptive fields, based on the quality of the input image x. To

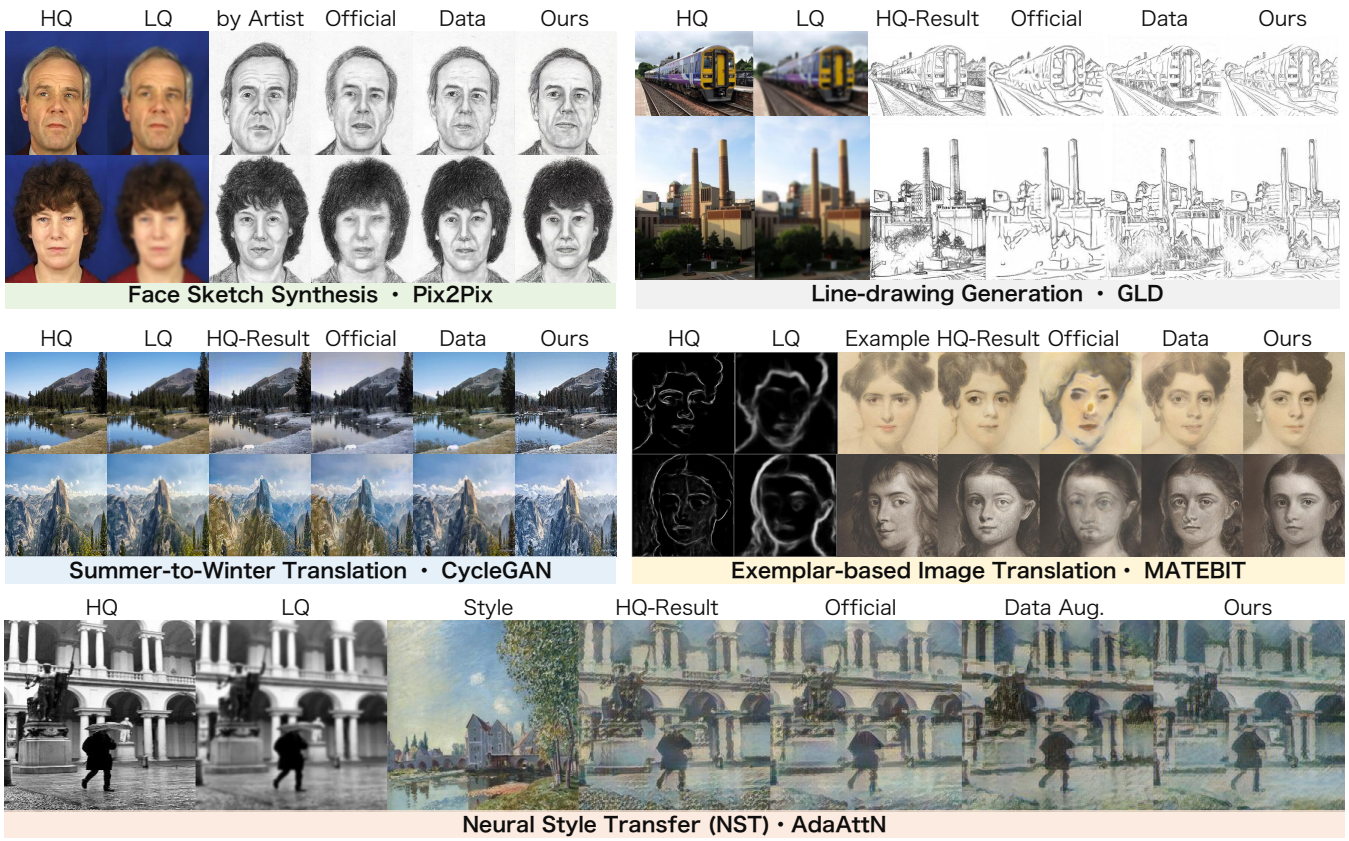

Figure 5: Impact of QUARF on various image-to-image translation (I2IT) models, in diverse I2IT tasks. In each task, we train the model with high-quality samples (*Official*), data augmentation (*Data Aug.*), and QUARF, respectively; and then apply the learned model to both high-quality (HQ) content image and its randomly degraded low-quality (LQ) version.

obtain effective quality representations, we choose an existing IQA method ψ_Q in the implementation. Let $\mathbf{f}_Q$ denote the quality feature vector, we map $\mathbf{f}_Q$ to the routing parameters for each group i, through a shallow MLP ϕ_i with a Softmax activation function. The process is formulated by:

$$\mathbf{w}_i = \phi_i(\mathbf{f}_Q) = \mathrm{Softmax}(\mathrm{MLP}(\mathbf{f}_Q)), \tag{4}$$

where $\mathbf{w}_i = [w_{i,1}, ..., w_{i,\eta}]$; $w_{i,j}$ determines the significance of the j-th receptive field for group i. The integrated output of group i becomes:

$$\mathbf{H}_i = \sum_{j=1}^{\eta} w_{i,j} \odot \mathbf{Z}_j, \tag{5}$$

where $\odot$ denotes the element product.

Quality Representations In this work, we focus on designing the quality-adaptive kernel selection mechanism. Thus, we adopt an existing *blind image quality assessment* (BIQA) models to extract the quality feature vector $\mathbf{f}_Q$ from the input image x. In default, we choose CONTRIQUE (Madhusudana et al. 2022) to extract deep features from the input image; because CONTRIQUE can efficiently and precisely predict the perceptual quality of an image. Note that, it's optional to select other BIQA methods instead. We'll also evaluate the performance of QUARF while using a different BIQA method in the experiments.

3.4 Plug-and-Play Applications

As presented previously, the proposed QUARF is simple and intuitive. One can easily apply it to a wide range of networks, by replacing a convolutional layer or block by QUARF, or by additionally plugging QUARF to an appropriate position. As will present in the experimental section, QUARF stably gains performance improvement on various networks, in diverse tasks, and under different settings.

4 Experiments

We apply QUARF to a variety of DNNs, and conduct experiments in the image translation, restoration, and semantic segmentation tasks, respectively. In each task, we select an advanced DNN-based method as the baseline model, and then apply QUARF to the network. Afterward, we compare the following three settings: (1) train the original model exactly following the official settings or using the officially released model (`Official`); (2) train the original model with data augmentation (`+Data Aug.`); and (3) train the model with QUARF on the augmented dataset (`+QUARF`). All the

(a) **I2IT & NST**. Lower FID/SIFID values indicate better style realism; higher CPBD values indicate better clarity.

	face sketch sythesis **Pix2Pix**			*summer-to-winter* **CycleGAN**			*photo-to-sketch* **GLD**			*sketch-to-painting* **MATEBIT**			*neural style transfer* **AdaAttN**		
	FID↓	SIFID↓	CPBD↑	FID↓	SIFID↓	CPBD↑	FID↓	SIFID↓	CPBD↑	FID↓	SIFID↓	CPBD↑	FID↓	SIFID↓	CPBD↑
Original	40.54	2.398	0.613	167.00	3.044	0.251	**124.45**	**0.823**	**0.806**	96.26	1.700	0.272	**84.98**	**0.280**	**0.699**
+Data Aug.	26.68	2.372	0.639	**80.41**	0.566	0.670	163.30	1.763	0.776	35.14	0.660	0.712	89.39	0.356	0.658
+QUARF	**26.25**	**2.157**	**0.649**	83.97	**0.467**	**0.717**	154.95	1.245	0.796	**29.87**	**0.437**	**0.720**	86.39	0.336	0.665

(b) **Image Restorition**. AST: deraining on AGAN-Data . Degradations in `Data Aug.` are included in *Official*.

	Real-ESRGAN (×4 SR) (Wang et al. 2021b)										**AST** (Zhou et al. 2024)	
	BSD100		**Manga109**		**Urban100**		**DF1K**				**AGAN-Data**	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	FID↓	LPIPS↓	PSNR↑	SSIM↑
Official	25.21	0.660	26.01	0.837	21.20	0.657	24.03	0.704	23.49	0.2743	32.32	0.935
+QUARF	**26.69**	**0.702**	**27.11**	**0.841**	**22.35**	**0.692**	**26.67**	**0.779**	**18.09**	**0.1966**	**32.39**	**0.946**

(c) **Face Parsing**: SegNeXt (Guo et al. 2022) on the CelebAMask-HQ dataset (Lee et al. 2020).

	mIoU ↑ (Mask IoU)							**BIoU** ↑ (Boundary IoU)						
SegNeXt	HQ	LQ-1	LQ-2	LQ-3	LQ-4	LQ-5	avg	HQ	LQ-1	LQ-2	LQ-3	LQ-4	LQ-5	avg
Official	**74.51**	**71.21**	57.80	45.61	21.87	9.76	41.25	**64.15**	59.85	43.35	27.32	8.20	3.89	28.52
+ Data Aug.	71.79	70.00	63.99	59.98	54.96	49.94	59.77	62.41	59.78	51.57	47.87	42.40	36.87	47.70
+ QUARF	72.59	70.16	**64.06**	**60.10**	**55.35**	**50.48**	**60.03**	62.76	**60.34**	**52.46**	**48.82**	**43.54**	**37.97**	**48.62**

(d) **Semantic Segmentation**: average performance aross the DIS, COIFT, HRSOD, and ThinObject datasets.

	mIoU ↑ (Mask IoU)							**BIoU** ↑ (Boundary IoU)						
HQ-SAM	HQ	LQ-1	LQ-2	LQ-3	LQ-4	LQ-5	avg	HQ	LQ-1	LQ-2	LQ-3	LQ-4	LQ-5	avg
Official	**88.67**	79.17	73.85	70.11	64.70	54.86	71.89	77.74	62.37	52.88	47.31	42.37	29.74	52.07
+ Data Aug.	88.38	**80.84**	**76.75**	70.41	**70.51**	**59.29**	**74.36**	78.13	64.32	56.28	**53.25**	48.16	35.70	55.97
+ QUARF	87.70	77.89	76.21	**72.28**	69.15	57.69	73.49	**79.79**	**65.90**	**59.03**	51.64	**49.50**	**36.21**	**57.01**

Table 1: Impacts of QUARF on generative and discriminative vision tasks. HQ: high-quality pristine image; LQ-l: low-quality image with the l-level degradation.

other experimental settings are exactly the same; and all the learned models are evaluated on the same testing set.

4.1 Image Translation

Settings We conduct experiments on both *image-to-image translation* (I2IT) and *neural style transfer* (NST) tasks, and evaluate our approach against four baselines: (1) *Face sketch synthesis* (FSS) with Pix2Pix (Isola et al. 2017) on the CUFS dataset (Wang and Tang 2008), (2) *Summer-to-winter translation* with CycleGAN on the *summer2winter* dataset (Zhu et al. 2017), (3) *Exemplar-based image translation* with MATEBIT (Jiang et al. 2023) on the *Metface* (Karras et al. 2020) dataset, (4) *Image-to-sketch translation* with GLD (Chan, Durand, and Isola 2022) on the COCO dataset and the Anime Colorization dataset (Chan, Durand, and Isola 2022), and (5) *Neural Style Transfer* (NST) with AdaAttN (Liu et al. 2021), using the MSCOCO and WikiArt datasets (Lin et al. 2014; Phillips and Mackintosh 2011).

Criteria We use the *Fréchet Inception Score* (FID) (Seitzer 2020), *Single Image Fréchet Inception Distance* (SIFID), and *Cumulative Probability of Blur Detection* (CPBD) (Narvekar and Karam 2011) for evaluating the quality of generated images. Lower FID or SIFID values indicate better style realism; while lower CPBD values indicate more serious blurring effects.

Results Table 1 (a) shows that QUARF consistently delivers significant improvements across various baselines, achieving the lowest FID and competitive results on SIFID and CPBD. Besides, Fig. 5 shows that the images generated by our method present the best visual quality, in terms of clarity and style, across all these tasks.

4.2 Image Restoration (IR)

Settings For *Single-Image Super Resolution* (SISR), we use Real-ESRGAN (Wang et al. 2021b) as the baseline network, and conduct experiments on the DF2K (Lim et al. 2017) and OST300 (Wang et al. 2018) datasets. The dataset is split 7:3 into training and testing sets, with the test set referred to as DF1K. For *Image Restoration*, we utilize AST (Zhou et al. 2024) as the base network and perform deraining experiments under the official settings. In each task, we compare (1) the official baseline model and (2) the model variant with QUARF learned from the same training set.

Results Following the official settings, we use PSNR and SSIM on Urban100, BSD100, Manga109 and AGAN-Data, and FID and LPIPS on DF1K. As shown in Table 1(b), QUARF stably and significantly boost the SR performance across all the datasets, according to all the indices. Besides,

Figure 6: Illustration of raindrop removal results on the AGAN-Data dataset. Our results exhibit superior visual quality.

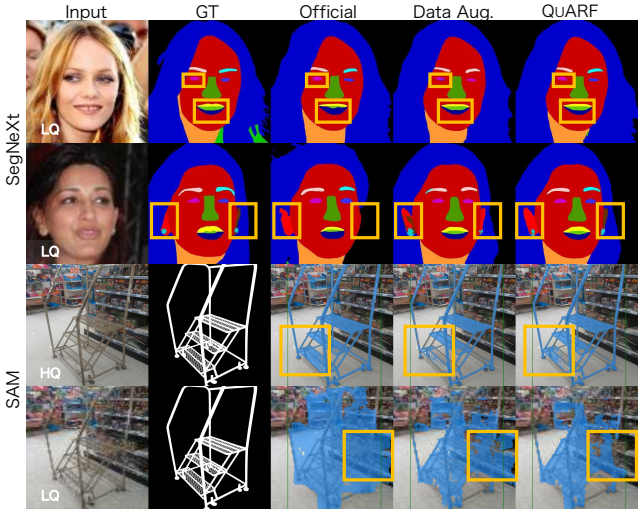

Figure 7: Illustration of semantic segmentation results.

QuARF outperforms the official AST model in terms of both PSNR and SSIM. Consistently, Fig. 6 shows that the images generated by QuARF present less artifacts.

4.3 Face Parsing

Settings We used SegNeXt (Guo et al. 2022) as the base network for face parsing experiments on CelebAMask-HQ (Lee et al. 2020), to validate QuARF's effectiveness, under the same standard settings.

Results We report the mIoU and BIoU on the CelebAMask-HQ test set, across different levels of hybrid degradation levels (following CodeFormer (Zhou et al. 2022a)). Table 1(c) shows that QuARF significantly boost the face parsing accuracy for degraded images, in terms of both mIoU and BIoU. Besides, QuARF lessens the performance decrease on high-quality images caused by data augmentation. Fig. 7 shows that QuARF leads to preciser parsing results, compared to the `Official` and `Data Aug.` models, especially in terms of regional boundaries. Such observations are consistent with the superiority of QuARF in BIoU.

Position	Pix2Pix FID↓	Pix2Pix SIFID↓	Pix2Pix CPBD↑	MATEBIT FID↓	MATEBIT SIFID↓	MATEBIT CPBD↑
Encoder	**26.25**	2.157	0.649	**29.87**	**0.437**	0.720
Decoder	26.89	2.330	**0.653**	32.14	0.448	**0.726**
both	60.16	**2.076**	0.608	32.63	0.651	0.698

Table 2: Impact of the plug positions of QuARF.

	Pix2Pix FID↓	Pix2Pix SIFID↓	Pix2Pix CPBD↑	CycleGAN FID↓	CycleGAN SIFID↓	CycleGAN CPBD↑
QUARF $_{\text{CNN}}$	27.08	2.310	**0.655**	133.64	0.776	0.665
QUARF $_{\text{BRSQ}}$	26.47	2.342	0.651	95.19	0.517	0.683
QUARF $_{\text{CTRQ}}$	**26.25**	**2.157**	0.649	**83.97**	**0.467**	**0.717**
	GLD FID↓	**GLD** SIFID↓	**GLD** CPBD↑	**MATEBIT** FID↓	**MATEBIT** SIFID↓	**MATEBIT** CPBD↑
QUARF $_{\text{CNN}}$	**144.1**	1.662	0.778	34.40	0.711	0.708
QUARF $_{\text{BRSQ}}$	158.2	1.390	0.741	33.45	0.610	0.690
QUARF $_{\text{CTRQ}}$	154.9	**1.245**	**0.796**	**29.87**	**0.437**	**0.720**

Table 3: Impact of quality representations on QuARF, in various I2IT tasks. `CTRQ` is our default choice.

4.4 Semantic Segmentation

Settings For semantic segmentation, we evaluate the impact of QuARF on SAM models (Kirillov et al. 2023b), by using HQ-SAM (Ke et al. 2024) as the baseline. We conducted experiments on the HQSeg-44K dataset, following the official settings (Ke et al. 2024). For HQ-SAM, we chose the image encoder of the ViT-B model as the visual backbone. We test the model performance on ThinObject-5K (Liew et al. 2021)(test set), DIS (Qin et al. 2022) (validation set), HR-SOD (Zeng et al. 2019) and COIFT (Liew et al. 2021). The original clear images as well as their degraded versions (5 different levels of hybrid distortion following CodeFormer (Zhou et al. 2022a)), are used for training and testing accordingly.

Results Fig. 7 shows that both `Official` and `Data Aug.` present prediction errors in low-quality images; while our module reduces the error regions. Similarly, Table 1(d) shows that QuARF achieves optimal BIOU values, across diverse levels of degradation. Both the quantitative and qualitative results imply that QuARF typically leads to continuous segmentation regions with precise boundaries.

4.5 Ablation Study

Impacts of Plug Position We first analyze the performance of QuARF when plugged in different layers of a network. Specifically, we insert QuARF after the down-sampling operation in the front encoder layer (`Encoder`), or after the up-sampling operation in the decoder layer (`Decoder`), or `both`. As shown in Table 2, using QuARF in the encoder leads to the best balanced performance according to all the indices. This is consistent with our motivation, of using multiscale kernels for extracting effective information, since the encoder is near to the input.

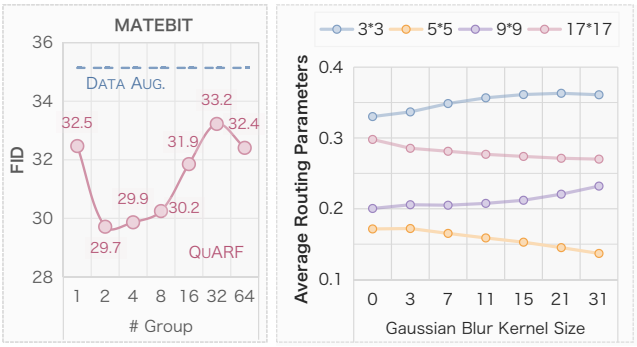

Figure 8: Left: impacts of the number of groups (ζ) on MATEBIT; Right: average kernel selection (routing) parameters vs. the Gaussian blurring kernel size.

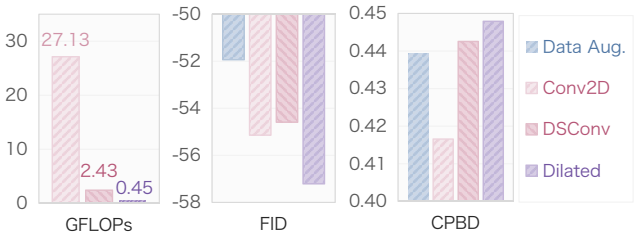

Figure 9: Analysis of the implementation of QUARF. To highlight the differences between these variants, we here show the changes of FID and CPBD, compared with the original baseline model (*middle* & *right*).

Impacts of Quality Representations. In this part, we aim to analyze the impact of quality features. To this end, we extract quality feature vectors by using three BIQA models, i.e. a 5-layer CNN, BRISQUE (Mittal, Moorthy, and Bovik 2012) and CONTRIQUE (Madhusudana et al. 2022), respectively. The corresponding variants of QUARF are denoted by the subscripts, `CNN`, `BRSQ` and `CTRQ`, sequentially. As shown in Table 3, all the variants of QUARF achieve superior performance in most cases. Such consistent performance improvement demonstrates that we can use different quality features in QUARF. It's potential to further boost the performance by using better quality representations.

Impacts of Groups We here analyze the impact of the number of groups (ζ) in QUARF (Fig. 4). As shown in the left plot of Fig 8, QUARF stably leads to optimal performance when the input feature tensor is chanel-wisely divided into 2 to 8 groups. Recall that the amount of parameters is inversely proportional to the number of groups (Eq. 3). We choose $\zeta = 4$ in default to achieve optimal balance between performance stability and efficiency.

Learned Kernel Routing Parameters To have an insight into QUARF, we compute the average kernel routing parameters of each receptive field, for images degraded with different sizes of Gaussian blurring. As visualized in Fig. 8, as image quality decreases, the weights of small receptive fields, i.e. $(3, 9)$, increase; while the weights of large receptive fields, i.e. $(5, 17)$, decrease. As previously shown in Fig. 2, features extracted by small receptive fields focus more on local characteristics, while large receptive fields focus more on extracting global features. As image quality declines, the model needs to pay more attention to local features that might be lost. This is consistent with the trend reflected in the curve.

	MATEBIT			AdaAttN		
	FID↓	SIFID↓	CPBD↑	FID↓	SIFID↓	CPBD↑
Sum	33.77	0.650	0.692	93.88	0.362	0.674
Concatenation	31.75	0.592	0.706	89.64	0.341	0.670
Channel Para.	33.29	0.673	**0.692**	91.20	0.384	**0.692**
Channel Att.	34.56	0.650	0.693	88.93	0.412	0.667
Spatial Att.	33.57	0.626	0.705	88.77	0.357	0.656
QUARF	**29.87**	**0.437**	**0.720**	**86.39**	**0.336**	0.665
QUARF$_{(\mathbf{F},\mathbf{f}_Q)}$	32.22	0.617	0.712	87.08	0.349	0.682
QUARF$_{\text{hard}}$	33.13	0.622	0.715	88.12	0.384	0.660

Table 4: Impact of kernel selection mechanisms (Fig. 3).

Implementation of Convolutions We further investigate the implementation method of QUARF by using different types of convolutions, i.e. *standard 2D convolution* (`Conv2D`), *depth-separable convolution* (`DSConv`), and *dilated convolution* (`Dilated`). As shown in Fig 9, the dilated convolution lead to the best image quality according to both FID and CPBD, with the best efficiency in terms of *giga floating point of operations per second* (GFLOPs).

Comparison with Existing Feature Fusion Mechanisms In this part, we compare our quality-adaptive selection methods, with existing ones (Fig. 3), i.e. Sum, Concatenation, Channel-wise learnable parameters (`Channel Para.`), Channel-wise Attention (`Channel Att.`), and Spatial-wise Attention (`Spatial Att.`). Besides, we modify QUARF by (1) using both the global average of the input tensor and the quality features (QUARF$_{(\mathbf{F},\mathbf{f}_Q)}$), or (2) merely selecting the receptive fields with the greatest routing parameter (`hard`). From above mentioned, using quality features is superior to using current feature maps to adjust. Table 4 shows that most feature fusion mechanisms lead to better performance, compared with data augmentation. Besides, our default QUARF achieves the best performance in general. Such observations again demonstrate our basic motivation, i.e. learning quality-adaptive receptive fields.

5 Conclusions

In this paper, we propose a novel methodology of *Quality-Adaptive Receptive Fields* (QUARF). Qualitative and quantitative experiments have shown that QUARF is capable of achieving quality adaptation and can significantly improve the generation quality across multiple existing models. The experimental results demonstrate the effectiveness of quality-adaption on various tasks. In the future, we will explore more approaches to solve the challenge of processing, analyzing, and generation tasks for degraded images.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants U22A2096, 62441601 and 62072148, in part by the Proof of Concept Foundation of Xidian University Hangzhou Institute of Technology under Grant GNYZ2023YL0301, in part by the Shaanxi Province Core Technology Research and Development Project under Grant 2024QY2-GJHX-11, in part by the Fundamental Research Funds for the Central Universities under Grants QTZX23042 and ZYTS24012, and in part by the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant GK239909299001-009.

References

Chan, C.; Durand, F.; and Isola, P. 2022. Learning to generate line drawings that convey geometry and semantics. In *CVPR*, 7915–7925.

Chen, Z.; Zhang, Z.; Li, H.; Li, M.; Chen, Y.; Li, Q.; Feng, H.; Xu, Z.; and Chen, S. 2024. Deep Linear Array Pushbroom Image Restoration: A Degradation Pipeline and Jitter-Aware Restoration Network. In *AAAI*, volume 38, 1290–1298.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 1251–1258.

Cui, Y.; Ren, W.; and Knoll, A. 2024. Omni-Kernel Network for Image Restoration. In *AAAI*, volume 38, 1426–1434.

Ding, X.; Zhang, X.; Han, J.; and Ding, G. 2022. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *CVPR*, 11963–11975.

Fan, Q.; Tao, X.; Ke, L.; Ye, M.; Zhang, Y.; Wan, P.; Wang, Z.; Tai, Y.-W.; and Tang, C.-K. 2023. Stable Segment Anything Model. *arXiv preprint arXiv:2311.15776*.

Gao, X.; Gao, F.; Tao, D.; and Li, X. 2013. Universal Blind Image Quality Assessment Metrics Via Natural Scene Statistics and Multiple Kernel Learning. *IEEE TNNLS*, 24(12): 2013–2026.

Gu, J.; Lu, H.; Zuo, W.; and Dong, C. 2019. Blind super-resolution with iterative kernel correction. In *CVPR*, 1604–1613.

Guo, M.-H.; Lu, C.-Z.; Hou, Q.; Liu, Z.; Cheng, M.-M.; and Hu, S.-M. 2022. SegNeXt: Rethinking convolutional attention design for semantic segmentation. *NeurIPS*, 35: 1140–1156.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, P.; Jiao, L.; Shang, R.; Wang, S.; Liu, X.; Quan, D.; Yang, K.; and Zhao, D. 2022. MANet: Multi-scale aware-relation network for semantic segmentation in aerial scenes. *IEEE TGRS*, 60: 1–15.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*, 1125–1134.

Jiang, C.; Gao, F.; Ma, B.; Lin, Y.; Wang, N.; and Xu, G. 2023. Masked and Adaptive Transformer for Exemplar Based Image Translation. In *CVPR*, 22418–22427.

Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; and Aila, T. 2020. Training generative adversarial networks with limited data. *NeurIPS*, 33: 12104–12114.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *ICCV*, 5148–5157.

Ke, L.; Ye, M.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; Yu, F.; et al. 2024. Segment anything in high quality. *NeurIPS*, 36.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023a. Segment anything. In *ICCV*, 4015–4026.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023b. Segment anything. In *ICCV*, 4015–4026.

Kumar, A.; Kim, D.; Park, J.; and Behera, L. 2024. Pick-or-Mix: Dynamic Channel Sampling for ConvNets. In *CVPR*, 5873–5882.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 5549–5558.

Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.-M.; Yang, J.; and Li, X. 2023. Large selective kernel network for remote sensing object detection. In *ICCV*, 16794–16805.

Liew, J. H.; Cohen, S.; Price, B.; Mai, L.; and Feng, J. 2021. Deep interactive thin object selection. In *WACV*, 305–314.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *CVPR workshops*, 136–144.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755.

Liu, S.; Chen, T.; Chen, X.; Chen, X.; Xiao, Q.; Wu, B.; Kärkkäinen, T.; Pechenizkiy, M.; Mocanu, D.; and Wang, Z. 2022a. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv preprint arXiv:2207.03620*.

Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; and Ding, E. 2021. AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style Transfer. In *ICCV*, 6649–6658.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022b. A convnet for the 2020s. In *CVPR*, 11976–11986.

Madhusudana, P. C.; Birkbeck, N.; Wang, Y.; Adsumilli, B.; and Bovik, A. C. 2022. Image quality assessment using contrastive learning. *IEEE TIP*, 31: 4149–4161.

Marr, D. 2010. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Marr, D.; and Hildreth, E. 1980. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167): 187–217.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE TIP*, 21(12): 4695–4708.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE SPL*, 20(3): 209–212.

Moorthy, A. K.; and Bovik, A. C. 2011. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE TIP*, 20(12): 3350–3364.

Narvekar, N. D.; and Karam, L. J. 2011. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE TIP*, 20(9): 2678–2683.

Phillips, F.; and Mackintosh, B. 2011. Wiki art gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3): 593–608.

Qin, G.; Hu, R.; Liu, Y.; Zheng, X.; Liu, H.; Li, X.; and Zhang, Y. 2023. Data-efficient image quality assessment with attention-panel decoder. In *AAAI*, volume 37, 2091–2100.

Qin, X.; Dai, H.; Hu, X.; Fan, D.-P.; Shao, L.; and Van Gool, L. 2022. Highly accurate dichotomous image segmentation. In *ECCV*, 38–56.

Rebuffi, S.-A.; Gowal, S.; Calian, D. A.; Stimberg, F.; Wiles, O.; and Mann, T. A. 2021. Data augmentation can improve robustness. *NeurIPS*, 34: 29935–29948.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *CVPR*, 779–788.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Seitzer, M. 2020. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid. Version 0.3.0.

Shi, J.; Gao, P.; and Qin, J. 2024. Transformer-based no-reference image quality assessment via supervised contrastive learning. In *AAAI*, volume 38, 4829–4837.

Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *CVPR*, 3667–3676.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015a. Going deeper with convolutions. In *CVPR*, 1–9.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015b. Going deeper with convolutions. In *CVPR*, 1–9.

Wang, L.; Wang, Y.; Dong, X.; Xu, Q.; Yang, J.; An, W.; and Guo, Y. 2021a. Unsupervised degradation representation learning for blind super-resolution. In *CVPR*, 10581–10590.

Wang, X.; and Tang, X. 2008. Face photo-sketch synthesis and recognition. *IEEE TPAMI*, 31(11): 1955–1967.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021b. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 1905–1914.

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 606–615.

Xie, L.; Zheng, C.; Xue, W.; Jiang, L.; Liu, C.; Wu, S.; and Wong, H. S. 2024. Learning Degradation-unaware Representation with Prior-based Latent Transformations for Blind Face Restoration. In *CVPR*, 9120–9129.

Xu, M.; Yoon, S.; Fuentes, A.; and Park, D. S. 2023. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, 137: 109347.

Yu, F.; and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; and Wang, J. 2021. OCNet: Object context for semantic segmentation. *IJCV*, 129(8): 2375–2398.

Zeng, Y.; Zhang, P.; Zhang, J.; Lin, Z.; and Lu, H. 2019. Towards high-resolution salient object detection. In *ICCV*, 7234–7243.

Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. 2022. Resnest: Split-attention networks. In *CVPR*, 2736–2746.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*, 3836–3847.

Zhang, W.; Zhai, G.; Wei, Y.; Yang, X.; and Ma, K. 2023. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *CVPR*, 14071–14081.

Zheng, N.; Huang, J.; Zhou, M.; Yang, Z.; Zhu, Q.; and Zhao, F. 2023. Learning semantic degradation-aware guidance for recognition-driven unsupervised low-light image enhancement. In *AAAI*, volume 37, 3678–3686.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *AAAI*, volume 34, 13001–13008.

Zhou, S.; Chan, K.; Li, C.; and Loy, C. C. 2022a. Towards robust blind face restoration with codebook lookup transformer. *NeurIPS*, 35: 30599–30611.

Zhou, S.; Chen, D.; Pan, J.; Shi, J.; and Yang, J. 2024. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *CVPR*, 2952–2963.

Zhou, Y.; Lin, C.; Luo, D.; Liu, Y.; Tai, Y.; Wang, C.; and Chen, M. 2022b. Joint learning content and degradation aware feature for blind super-resolution. In *ACM MM*, 2606–2616.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *ICCV*, 2223–2232.