Rethinking Image Evaluation

Donghun Ryou
dhryou@snu.ac.kr

2024.06.19

ComputerVisionLab
Seoul National University

Introduction

How can we do Image Quality Assessment (IQA)?

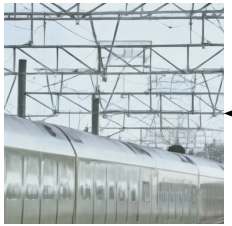
LQ

Restoration / Enhancement

Output

GT

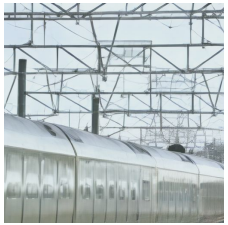
- Full Reference IQA (FR-IQA)



Output

GT

- No Reference IQA (NR-IQA)



Output

FR-IQA metrics - PSNR

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \quad MSE = \frac{\sum_{M,N}[I_1(m,n) - I_2(m,n)]^2}{M * N}$$

- Pros
  - Simple, Computationally inexpensive
- Cons
  - Doesn't Reflect Human Perception

FR-IQA metrics : SSIM (Structural Similarity Index Measure)

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^{\alpha} \cdot [c(\mathbf{x},\mathbf{y})]^{\beta} \cdot [s(\mathbf{x},\mathbf{y})]^{\gamma}$$

- l : luminance, c : contrast, s : structure

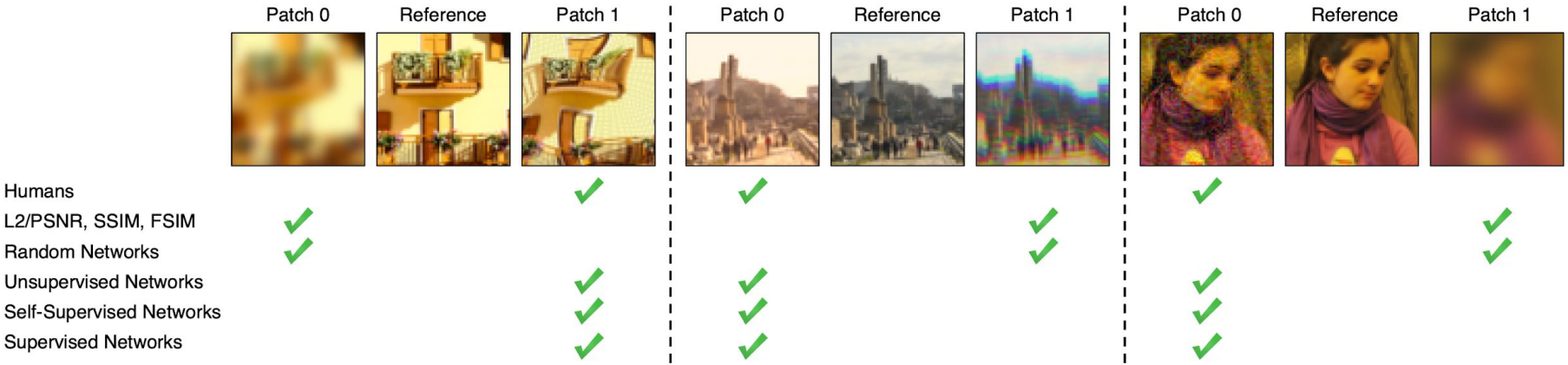
$$l(\mathbf{x},\mathbf{y}) = \frac{2\,\mu_x\,\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \quad c(\mathbf{x},\mathbf{y}) = \frac{2\,\sigma_x\,\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \quad s(\mathbf{x},\mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\,\sigma_y + C_3}$$

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2\right)^{1/2}, \quad \sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

- Pros
  - Better Perceptual Correlation than PSNR
- Cons
  - Still Not a Perfect Perceptual Model

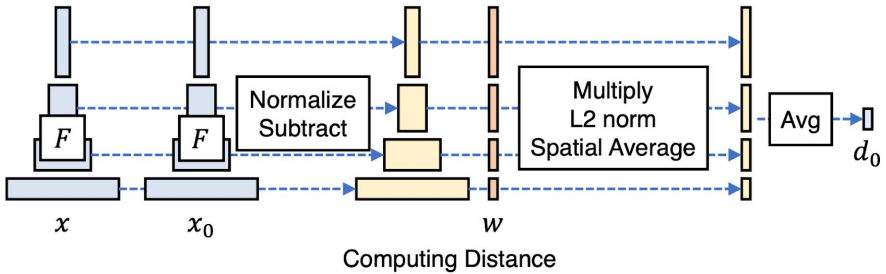
FR-IQA metrics : LPIPS

- The Unreasonable Effectiveness of Deep Features as a Perceptual Metric (2018 CVPR)



FR-IQA metrics : LPIPS

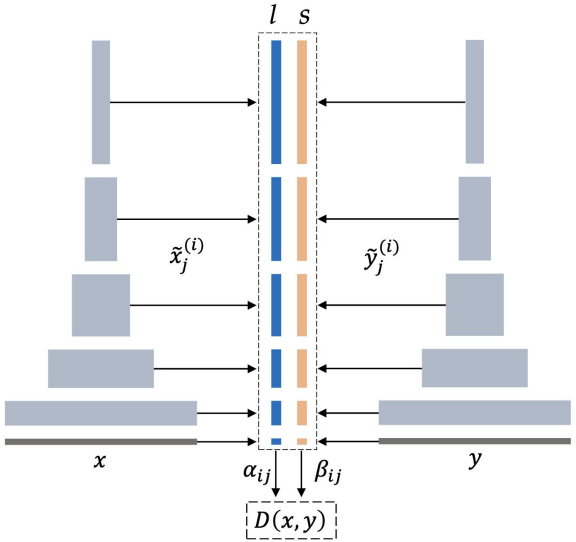
- The Unreasonable Effectiveness of Deep Features as a Perceptual Metric (2018 CVPR)



$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||_2^2$$

FR-IQA metrics : DISTS

- Image Quality Assessment: Unifying Structure and Texture Similarity (2020 TPAMI)

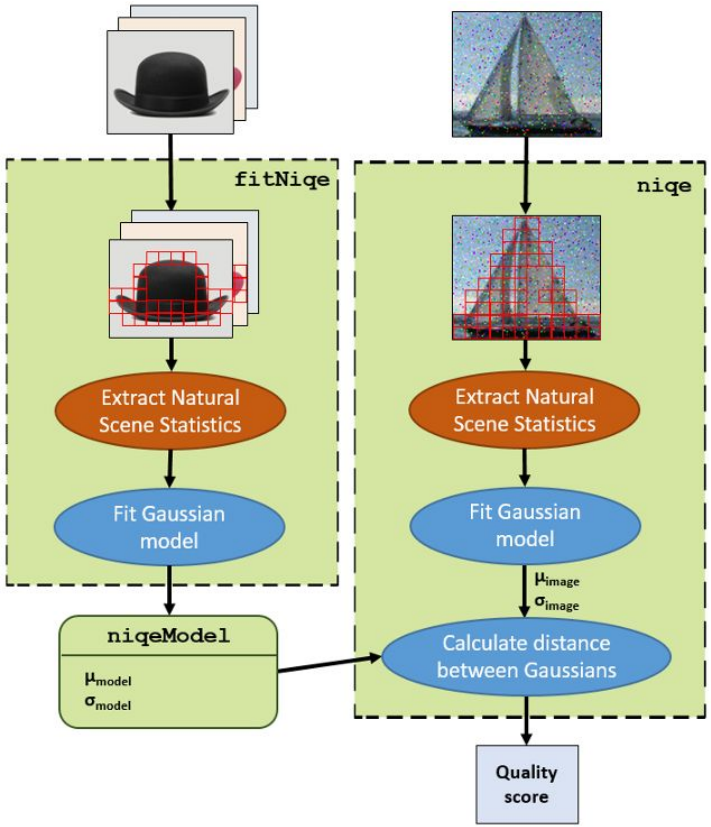


$$l(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}) = \frac{2\mu_{\tilde{x}_j}^{(i)}\mu_{\tilde{y}_j}^{(i)} + c_1}{\left(\mu_{\tilde{x}_j}^{(i)}\right)^2 + \left(\mu_{\tilde{y}_j}^{(i)}\right)^2 + c_1},$$

$$s(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}) = \frac{2\sigma_{\tilde{x}_j\tilde{y}_j}^{(i)} + c_2}{\left(\sigma_{\tilde{x}_j}^{(i)}\right)^2 + \left(\sigma_{\tilde{y}_j}^{(i)}\right)^2 + c_2},$$

$$D(x, y; \alpha, \beta) = 1 - \sum_{i=0}^{m}\sum_{j=1}^{n_i}\left(\alpha_{ij} l(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}) + \beta_{ij} s(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)})\right)$$

NR-IQA metrics : NIQE (2013)(without deep learning model)



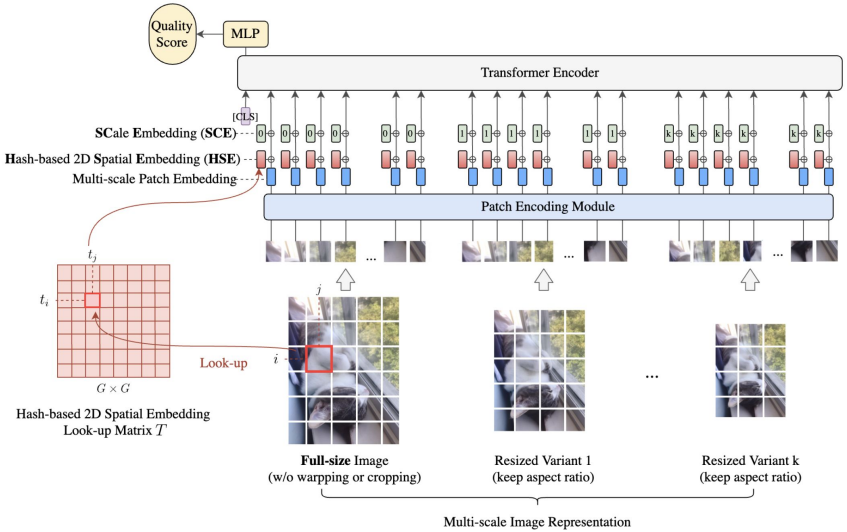
Dataset for NR-IQA metrics

- Image database TID2013: Peculiarities, results and perspectives
- ….
- KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment (2020 TPAMI)
- PIPAL: a Large-Scale Image Quality Assessment Dataset for Perceptual Image Restoration (2020 ECCV)

- Dataset Collection Process
  - Source: Data gathered from diverse sources.
  - Rating Method: Subjective image quality ratings obtained via a crowdsourcing platform.
  - Evaluation Criteria: Assessed based on factors including noise, JPEG artifacts, aliasing, lens blur, motion blur, over-sharpening, incorrect exposure, color fringing, and over-saturation.
  - Rating Scale: Utilized a 5-point Absolute Category Rating (ACR) scale: 1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent
  - Final Metric: Mean Opinion Score (MOS) calculated from individual ratings.

NR-IQA metrics : MUSIQ / MANIQA

- Musiq: Multi-scale image quality transformer (2021 ICCV)
- Maniqa: Multi-dimension attention network for no-reference image quality assessment (2022 CVPR)



- MUSIQ

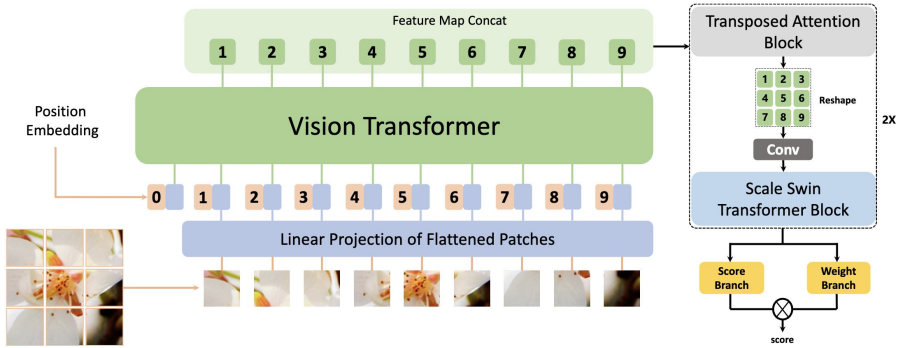
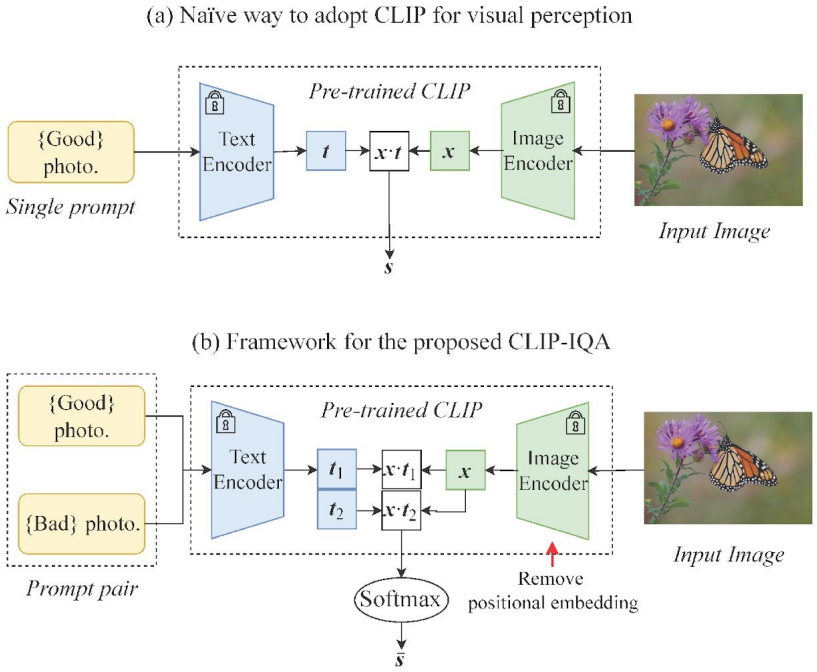
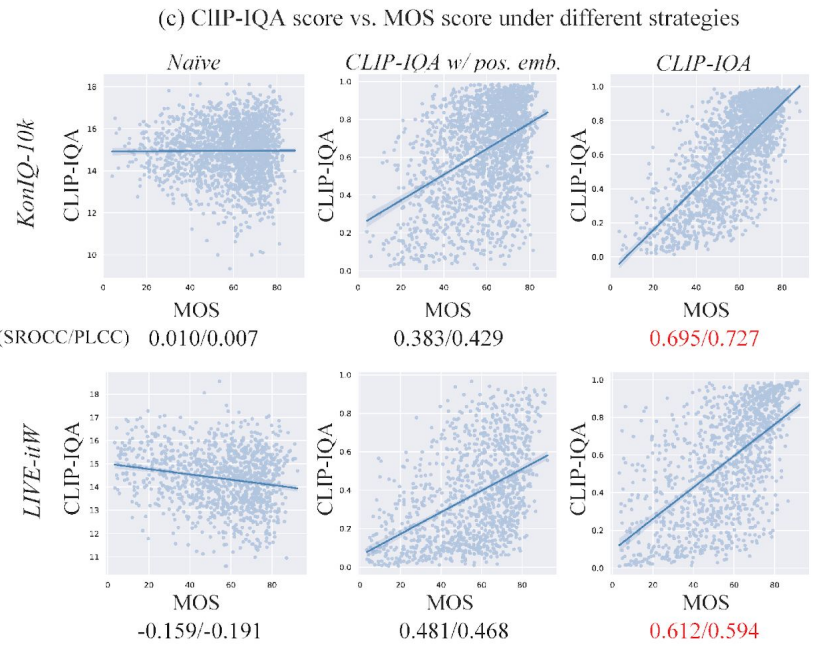


Figure 2. The architecture of the proposed approach - Multi-dimension Attention Network for no-reference Image Quality Assessment(MANIQA). A distorted image is cropped into 8×8 sized patches. Then the patches are inputted into the Vision Transformer (ViT) for extracting the features. Transposed attention block and scale swin transformer block, which are described in detail in Sec. 3.2 and Sec. 3.3, are used to strengthen the global and local interaction. A dual branch structure is proposed for predicting the weight and score of each patch in Sec. 3.4.

- MANIQA

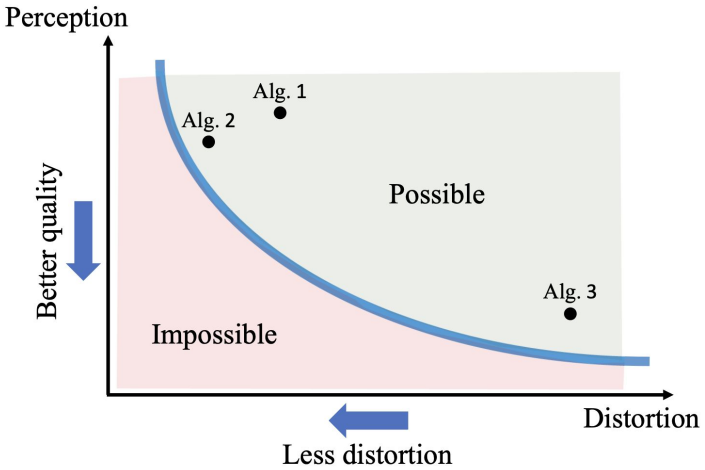
NR-IQA metrics : CLIP-IQA (2023 AAAI)





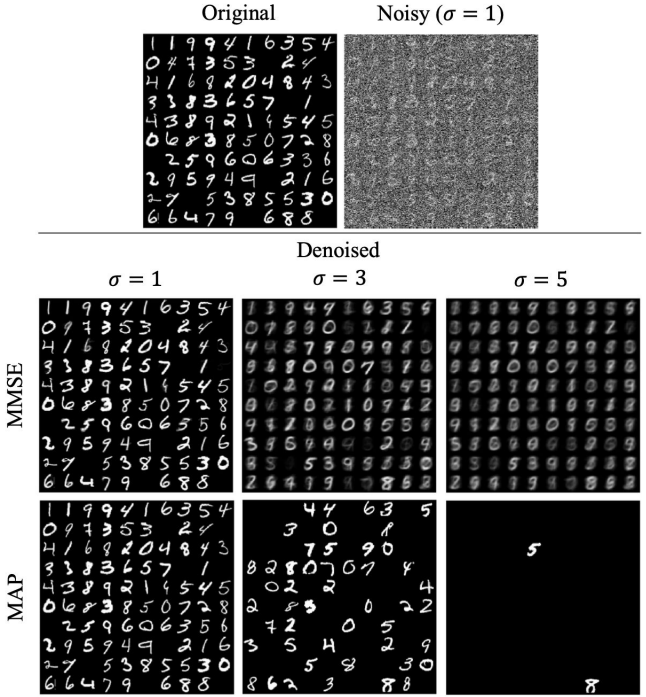
Why it is hard to get good psnr and perceptual quality at the same time?

- Perception-distortion tradeoff (2018 CVPR)



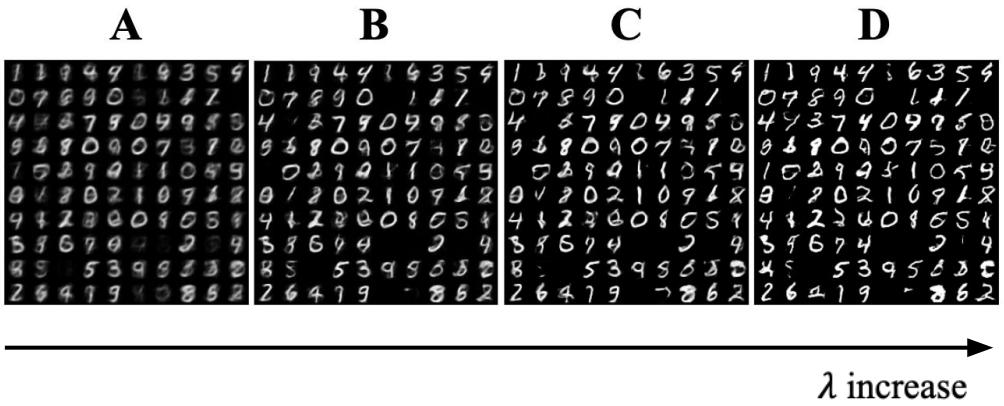
Why it is hard to get good psnr and perceptual quality at the same time?

- Perception-distortion tradeoff (2018 CVPR)



- Image Denoising utilizing a GAN

$$l_{\text{gen}} = l_{\text{MSE}} + \lambda\, l_{\text{adv}}$$



Rethinking Image Evaluation

- Propose New IQA metric

Toward Generalized Image Quality Assessment:
Relaxing the Perfect Reference Quality Assumption

2025 CVPR

Du Chen[1,3,*], Tianhe Wu[2,3,*], Kede Ma[2,†] and Lei Zhang[1,3,†]
[1]The Hong Kong Polytechnic University [2]City University of Hong Kong [3]OPPO Research Institute
csdud.chen@connet.polyu.hk, {tianhewu, kede.ma}@cityu.edu.hk, cslzhang@comp.polyu.edu.hk

- How to use IQA predictors for super-resolution training

Augmenting Perceptual Super-Resolution via Image Quality Predictors

2025 CVPR

Fengjia Zhang* Samrudhdhi B. Rangrej* Tristan Aumentado-Armstrong*
Afsaneh Fazly Alex Levinshtein
AI Center – Toronto, Samsung Electronics
{f.zhang2, s.rangrej, tristan.a, a.fazly, alex.lev}@samsung.com

Toward Generalized Image Quality Assessment: Relaxing the Perfect Reference Quality Assumption (2025 CVPR)

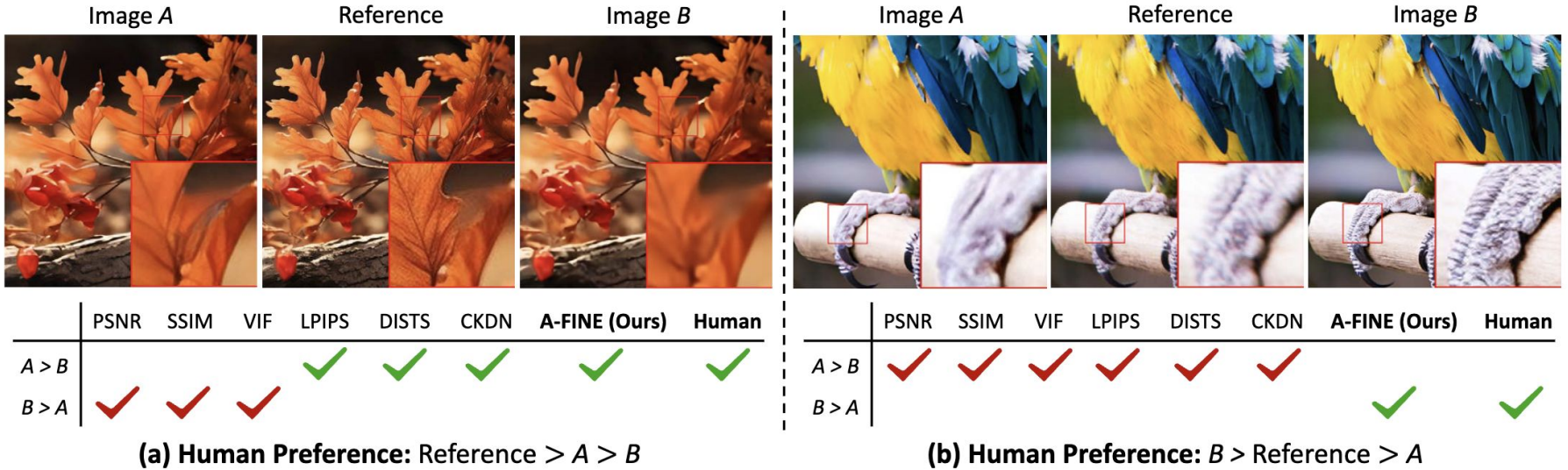
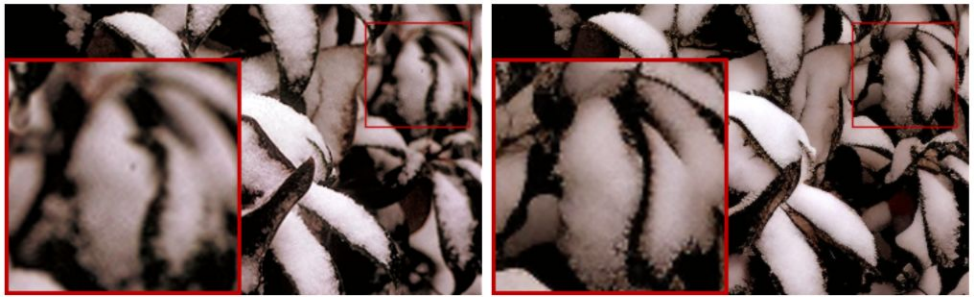


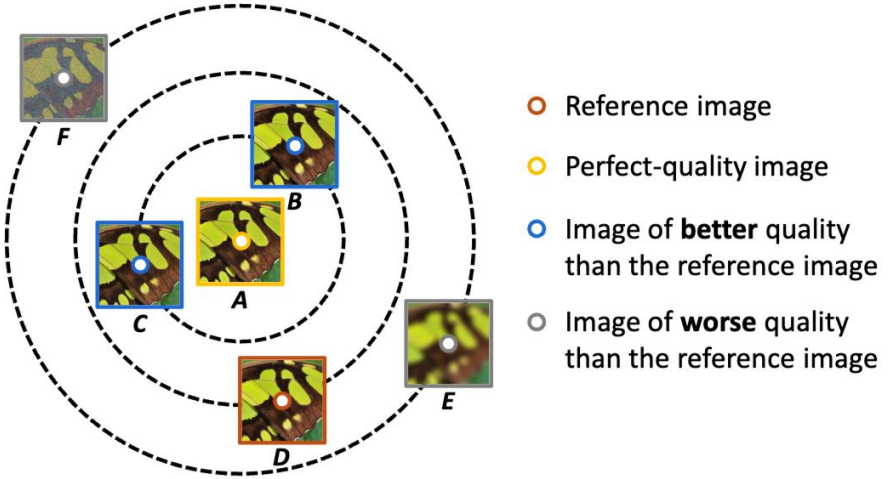
Fig. 1. *With the reference image in the middle, which image, A or B, has better perceived visual quality*? The proposed A-FINE generalizes and outperforms standard FR-IQA models under both perfect and imperfect reference conditions. Zoom in for better visibility.

Motivation

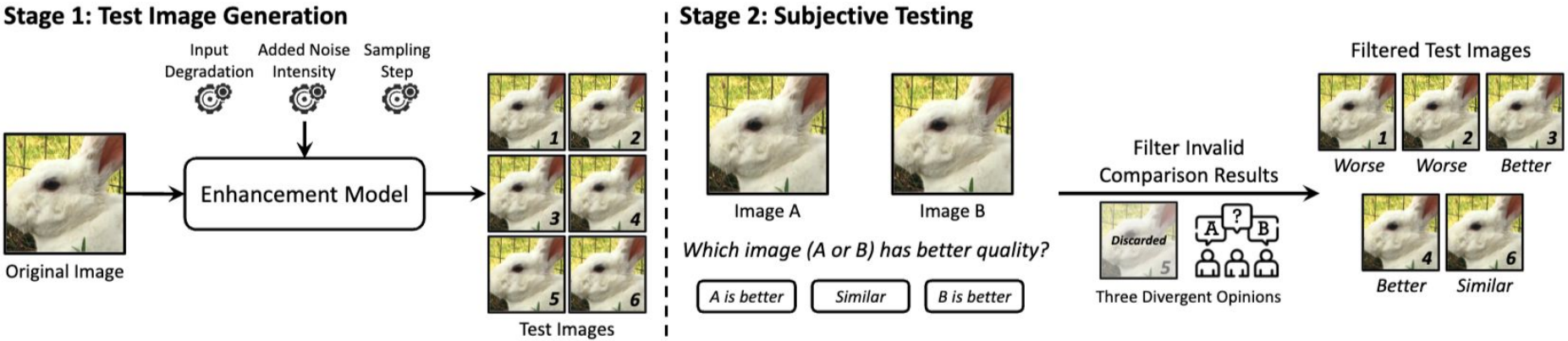
- FR-IQA assumes that the reference image is of perfect quality, but it is not



(a) Reference image (b) Enhanced image



Construction of DiffIQA dataset



- Dataset details
  - Generated by diffusion based super resolution (PASD), ~30K images (512x512), 6 test images per one original image

- Subject Testing details
  - Show two images (reference, test) in random spatial order
  - Choose one from three options : left image is of worse / similar / better
  - 240 subjects, each subject assigned 2240, Each image pair was rated by a minimum of three annotator
  - 232, 285 (43.20%) labeled as worse, 85, 671 (15.94%) as similar, and 219, 668 (40.86%) as better compared to the reference

Construction of DiffIQA dataset

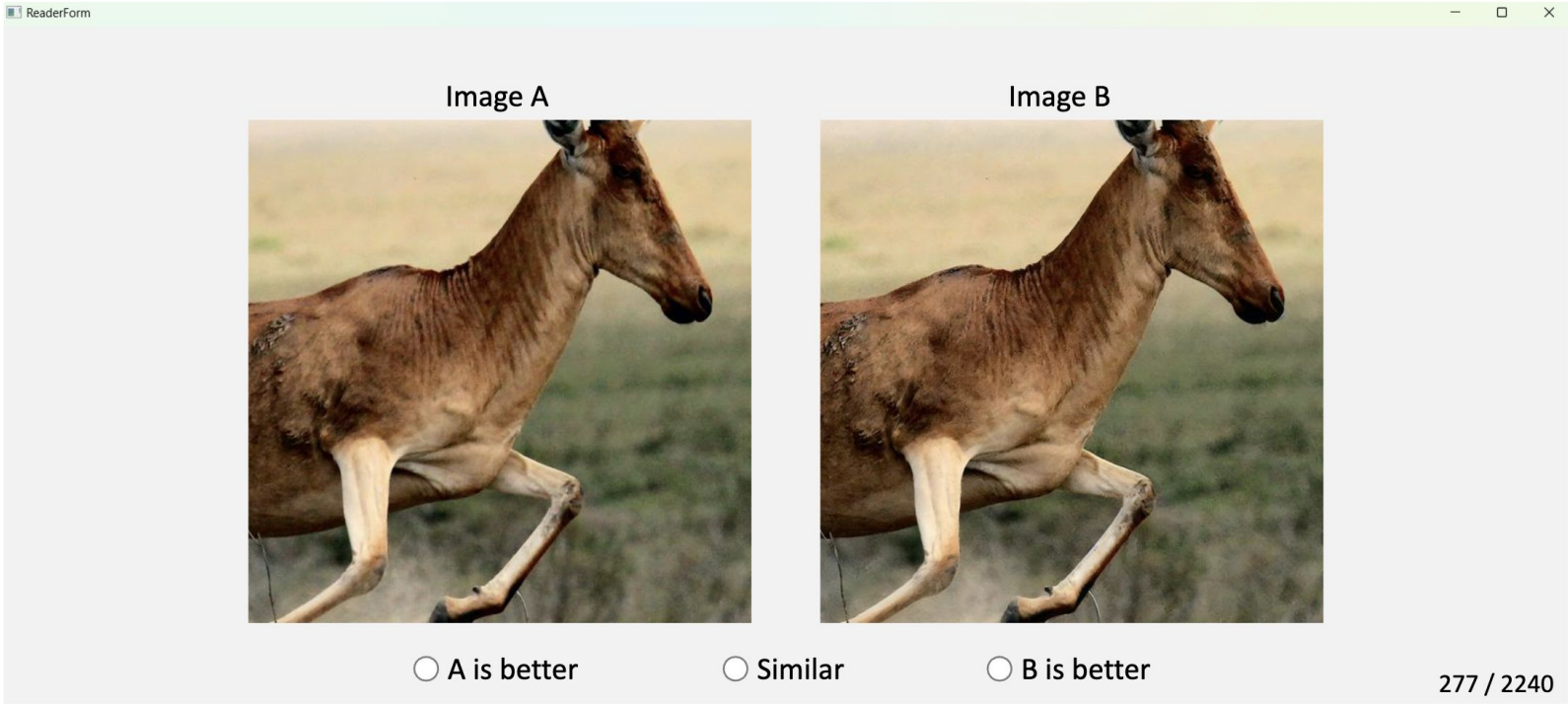


Fig. 8. The GUI used for constructing DiffIQA.

Proposed FR-IQA Model: A-FINE

- Evaluate the perceptual quality of **y** relative to **x**

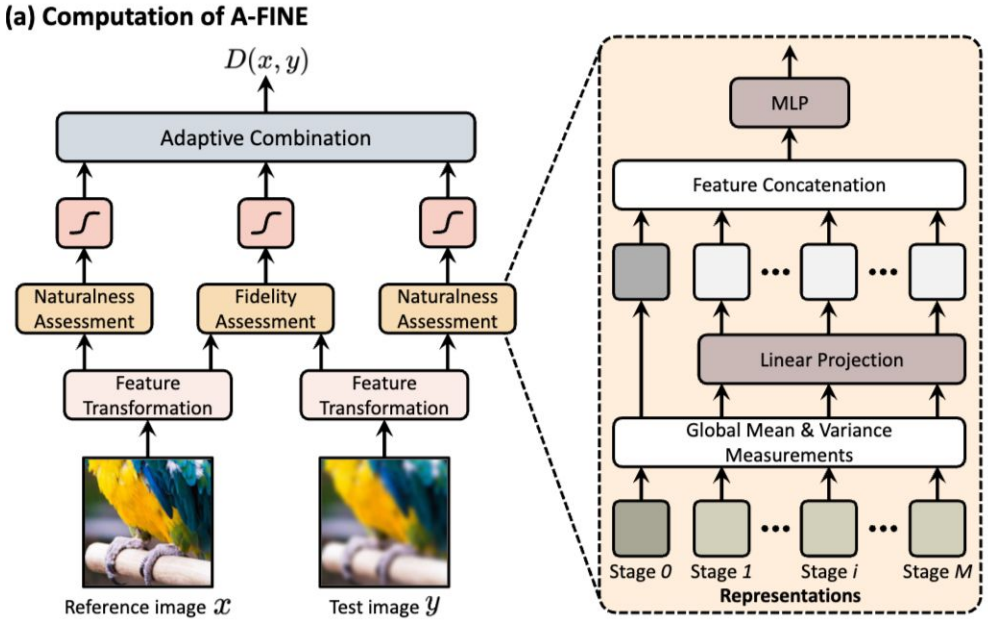
$$D(x,y) = F(x,y) + \lambda(x,y)N(y)$$

  - $F(x,y)$: Fidelity
  - $N(y)$: Naturalness
  - Smaller value indicates better quality

$$\lambda(x,y) = \exp\left(k(N(x) - N(y)\right)$$

  - $k \geq 0$, is learnable scale parameter
  - If y's perceptual quality is better than x, $D(x,y)$ depends more $N(y)$
  - If y's perceptual quality is worse than x, $D(x,y)$ depends more $F(x,y)$

Proposed FR-IQA Model: A-FINE



- For fidelity, DISTS like approach :

$$F(x, y) = 1 - \sum_{i=0}^{M} \sum_{j=1}^{N_i} F\left(x_j^{(i)}, y_j^{(i)}\right)$$

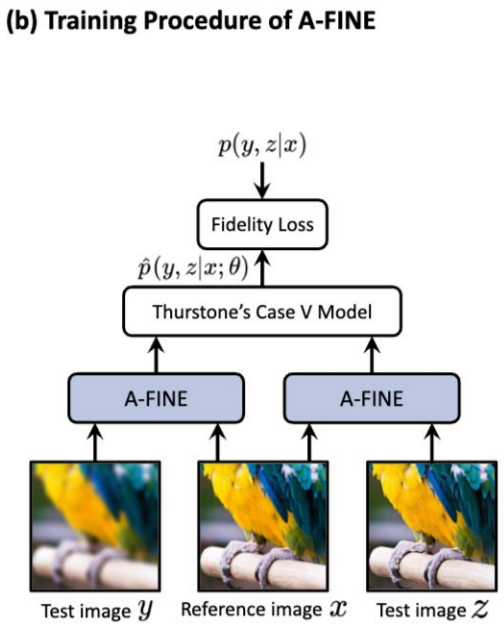
$$F\left(x_j^{(i)}, y_j^{(i)}\right) = \alpha_{ij} L\left(x_j^{(i)}, y_j^{(i)}\right) + \beta_{ij} S\left(x_j^{(i)}, y_j^{(i)}\right)$$

$$L(x_j^{(i)}, y_j^{(i)}) = \frac{2\mu_{x_j}^{(i)}\mu_{y_j}^{(i)} + c_1}{\left(\mu_{x_j}^{(i)}\right)^2 + \left(\mu_{y_j}^{(i)}\right)^2 + c_1}$$

$$S(x_j^{(i)}, y_j^{(i)}) = \frac{2\sigma_{x_j y_j}^{(i)} + c_2}{\left(\sigma_{x_j}^{(i)}\right)^2 + \left(\sigma_{y_j}^{(i)}\right)^2 + c_2}$$

- Using same CLIP-VIT backbone for fidelity and naturalness
- Additional MLP for naturalness

Proposed FR-IQA Model: A-FINE



- Ground-truth ranking label

$$p(y, z|x) = \begin{cases} 1 & \text{if } Q(y|x) > Q(z|x) \\ 0.5 & \text{if } Q(y|x) = Q(z|x) \\ 0 & \text{otherwise,} \end{cases}$$

- Assume that the perceptual quality of a test image follows a Gaussian distribution

$$\hat{p}(y, z|x; \theta) = \Phi\left(\frac{D(x, y; \theta) - D(x, z; \theta)}{\sqrt{2}}\right)$$

ϕ : standard Gaussian cumulative distribution function

- Loss function

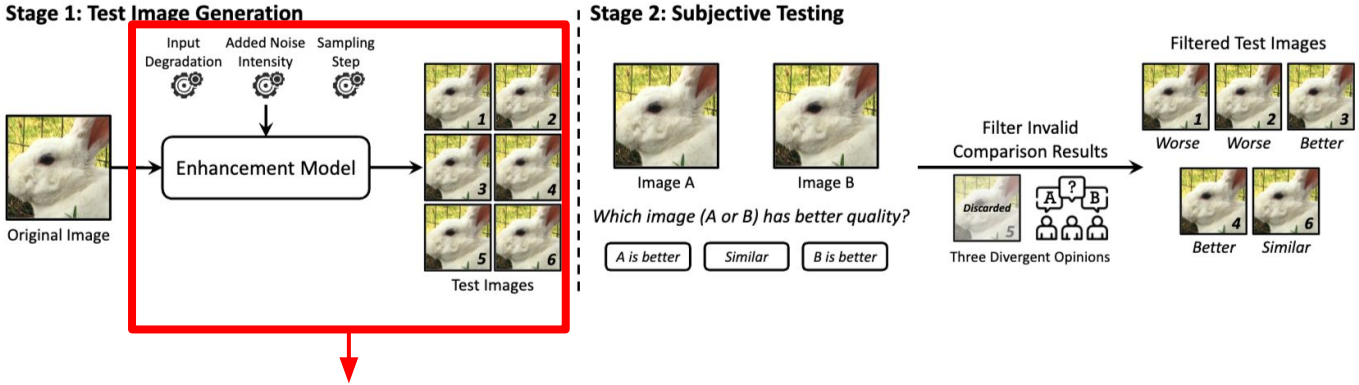
$$\ell(y, z|x; \theta) = 1 - \sqrt{p(y, z|x)\hat{p}(y, z|x; \theta)} - \sqrt{(1 - p(y, z|x))(1 - \hat{p}(y, z|x; \theta))}.$$

Experiments

Scenario	Method	Training Dataset	TID2013	KADID	PIPAL	Average	DiffIQA Ref < Test	DiffIQA Ref > Test	DiffIQA Average	All Average
Standard	PSNR	N.A.	75.8	74.8	70.7	72.2	18.2	92.1	45.6	58.9
	SSIM [45]	N.A.	68.9	74.0	72.1	72.4	20.1	93.0	47.1	60.0
	MS-SSIM [44]	N.A.	83.4	81.8	72.5	75.9	20.1	93.0	47.1	61.5
	FSIM [55]	N.A.	86.0	83.4	76.2	79.0	20.2	93.1	47.2	63.1
	VSI [56]	N.A.	87.3	84.8	76.2	79.5	19.7	93.1	46.9	63.2
	LPIPS [58]	BAPPS	78.7	77.0	74.3	75.4	23.7	94.7	50.0	62.7
	LPIPS-FT	Combined	72.5	78.2	71.7	73.6	35.4	91.6	55.6	64.6
	DISTS [5]	KADID	78.4	81.4	75.3	77.2	21.4	94.8	48.6	62.9
	DISTS-FT	Combined	78.4	81.9	72.1	75.3	38.2	89.5	56.7	66.0
	AHIQ [14]	PIPAL	74.6	76.4	79.3	78.1	34.1	88.1	54.1	66.1
	AHIQ-FT	Combined	81.0	79.7	74.9	76.7	78.4	73.8	76.7	76.7
	TOPIQ [1]	KADID	**90.4**	**94.3**	80.5	**85.1**	22.1	**95.1**	49.1	67.1
	TOPIQ-FT	Combined	78.9	85.0	79.0	80.6	78.6	74.2	77.0	78.8
Generalized	VIF [29]	N.A.	78.5	75.2	72.4	73.7	20.0	92.8	46.9	60.3
	PCQI [37]	N.A.	66.6	65.4	56.7	59.9	17.3	90.3	44.3	52.1
	SFSN [63]	N.A.	75.6	70.5	69.8	70.5	19.5	89.6	45.4	58.0
	CKDN [62]	PIPAL	76.9	70.9	79.8	77.1	33.3	82.4	51.4	64.3
	CKDN-FT	Combined	75.0	80.1	68.1	72.0	**79.4**	71.0	76.4	74.2
	A-FINE (Ours)	Combined	88.1	88.3	**81.0**	83.6	78.5	82.3	**79.9**	**81.8**

Experiment – SRIQA bench

- Similar pipeline with DiffIQA



- Using 2 regression based SR methods (SwinIR, RRDB), 8 generation based SR methods (Real-ESRGAN, BSR-GAN, HGGT, SUPIR, SeeSR, StableSR, SinSR, OSEDiff)

Experiment – SRIQA bench

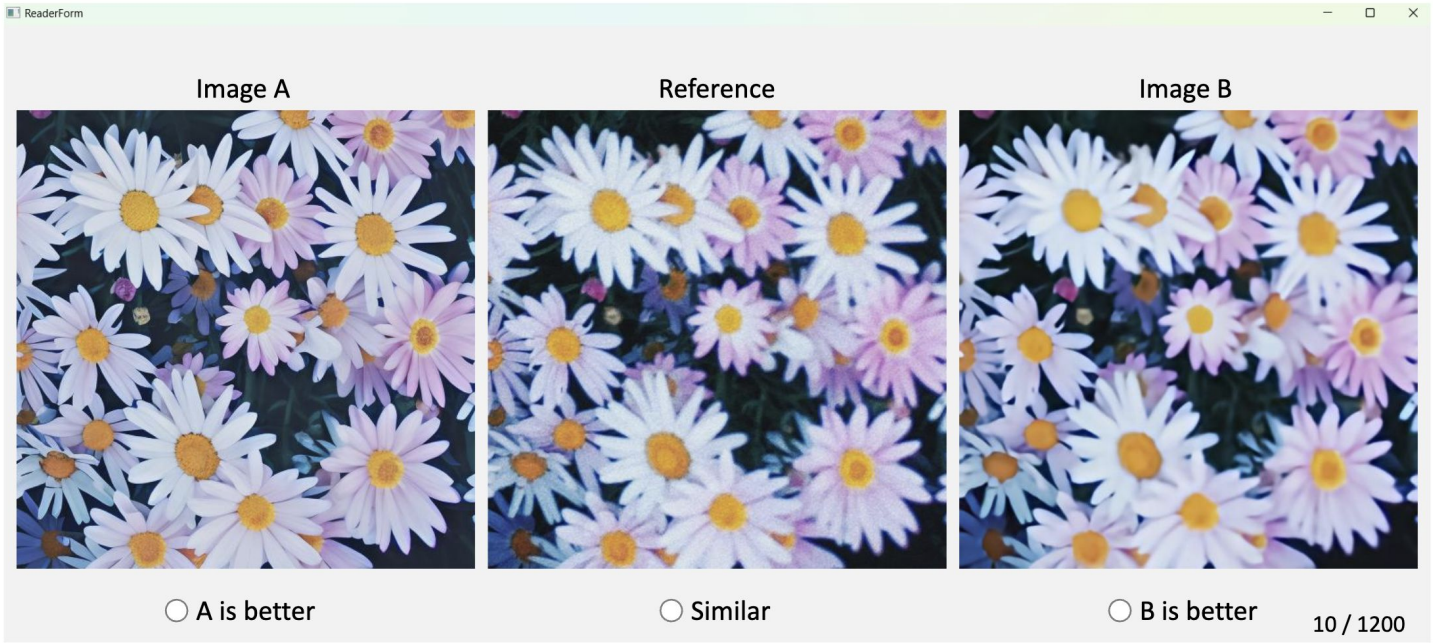


Fig. 9. The GUI used for constructing SRIQA-Bench.

Experiment – SRIQA bench

Method	Regression-based	Generation-based	All
PSNR	80.7	41.7	34.7
SSIM [45]	83.0	45.3	37.4
MS-SSIM [44]	83.0	45.6	37.6
FSIM [55]	<u>85.3</u>	49.5	41.0
VSI [56]	81.3	50.1	41.2
LPIPS [58]	82.0	63.9	65.8
LPIPS-FT	84.7	63.8	72.2
DISTS [5]	83.3	66.6	72.4
DISTS-FT	**86.0**	63.9	71.4
AHIQ [14]	83.7	70.0	68.4
AHIQ-FT	71.0	71.5	69.6
TOPIQ [1]	83.7	63.9	67.0
TOPIQ-FT	78.3	<u>73.0</u>	<u>77.7</u>
VIF [29]	<u>85.3</u>	47.1	39.0
PCQI [37]	79.0	39.8	32.2
SFSN [63]	80.3	48.4	39.9
CKDN [62]	45.0	60.1	47.4
CKDN-FT	76.7	64.3	59.1
A-FINE (Ours)	83.3	**78.9**	**82.4**

Experiment – Ablation

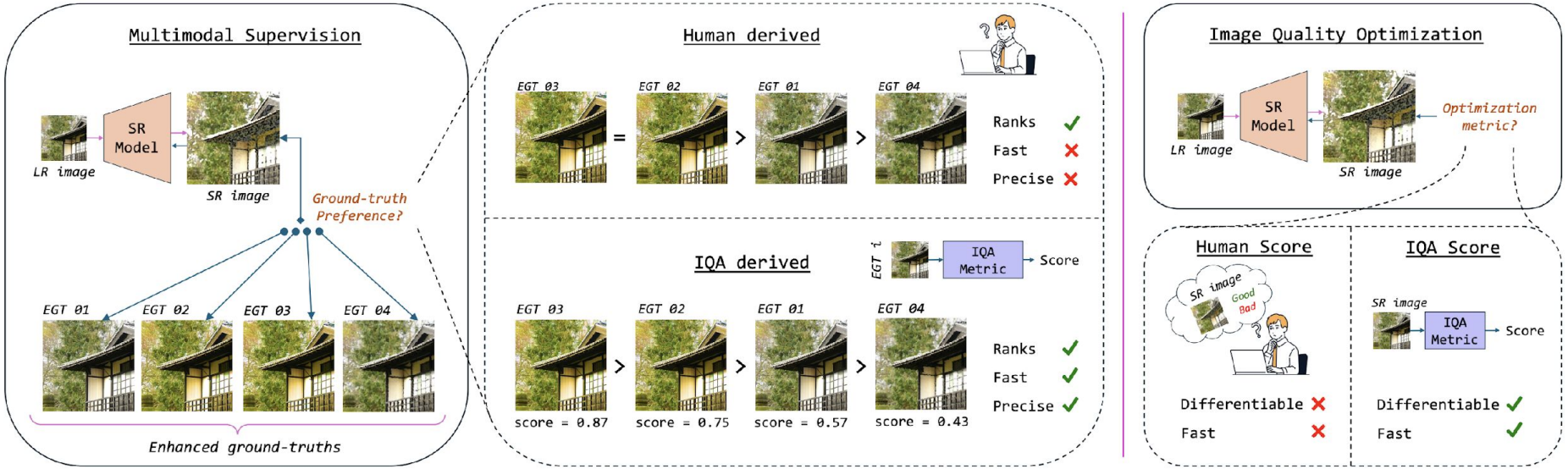
Table 4. Ablation study on backbone networks. The accuracy values in the "Standard" column are averaged across the test sets of TID2013 [26], KADID-10K [20], and PIPAL [10]. The results of TOPIQ-FT are included for reference.

Backbone	Standard	DiffIQA	SRIQA-Bench		
			Reg.	Gen.	All
TOPIQ-FT	80.6	77.0	78.3	73.0	77.7
VGG16	77.6	77.0	79.0	75.0	79.8
ResNet50 (ImageNet)	74.8	69.6	84.7	70.7	77.2
ResNet50 (CLIP)	76.1	71.1	**85.2**	70.3	75.6
ViT-B/32 (ImageNet)	81.0	77.7	81.3	75.5	80.4
ViT-B/32 (CLIP)	**83.6**	**79.9**	83.3	**78.9**	**82.4**

Table 5. Ablation study on training datasets.

Training Dataset	Standard	DiffIQA	SRIQA-Bench		
			Reg.	Gen.	All
Standard	**84.1**	65.6	**86.7**	71.8	78.7
DiffIQA	70.6	79.6	78.3	72.9	76.0
Combined	83.6	**79.9**	83.3	**78.9**	**82.4**

Augmenting Perceptual Super-Resolution via Image Quality Predictors (2025 CVPR)



- Two ways to improve perceptual quality :
  - 1. providing supervision through multiple enhanced ground-truth, 2. Direct optimization for the quality
- NR-IQA metrics can replace human raters

Analysis of NR-IQA metrics

- Dataset : SBS180K,
- Phase1 : 404 pairs form trainset, Compare 42 NR-IQA metrics
- Phase2 : Compare top 7 NR-IQA metrics from Phase1
  - Phase1

Method	PaQ-2-PiQ [97]	NIMA† [74]	MUSIQ◁ [41]	LIQE♡ [101]	ARNIQA-TID⋆ [1]	Q-Align◇ [88]	TOPIQ-NR [11]
Accuracy (%)	76.41	74.91	74.47	74.03	74.03	73.77	73.06

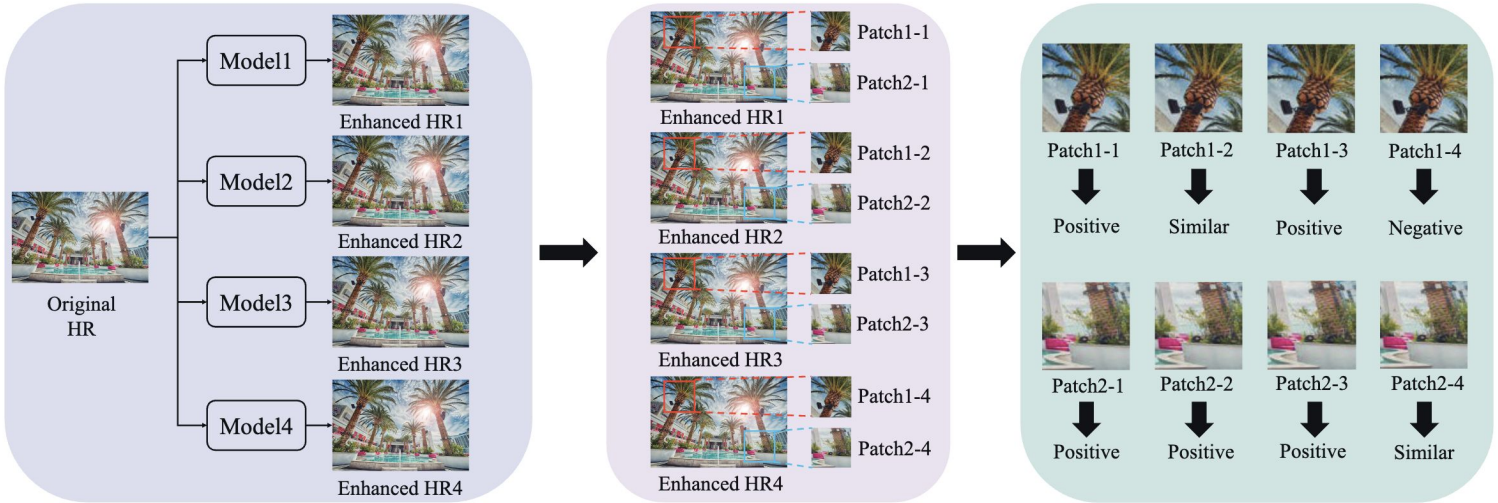
  - Phase2

Method	PaQ-2-PiQ [97]	NIMA† [74]	MUSIQ◁ [41]	LIQE♡ [101]	ARNIQA-TID⋆ [1]	Q-Align◇ [88]	TOPIQ-NR [11]
Train Acc. (%)	80.41	79.32	79.96	77.70	77.74	80.00	78.30
Test Acc. (%)	80.57	81.37	82.73	77.45	77.07	80.68	81.28

How to use NR-IQA metric for training super-resolution?
Background : HGGT (2023 CVPR)

Human Guided Ground-truth Generation for Realistic Image Super-resolution

Du Chen[1]*, Jie Liang[1,2]*, Xindong Zhang[1,2], Ming Liu[1,3], Hui Zeng[2] and Lei Zhang[1,2]†
[1]The Hong Kong Polytechnic University, [2]OPPO Research Institute, [3]Harbin Institute of Technology
{csdud.chen, c-ming.liu}@connet.polyu.hk, {liang27jie, cshzeng}@gmail.com
{csxdzhang, cslzhang}@comp.polyu.edu.hk



Step 1: HR Image Enhancement | Step 2: Patch Selection | Step 3: Annotation

Change Sampling method

- HGGT models are trained using uniformly sampled GT from positive samples.

$$\mathcal{L}(\theta|\widehat{I}, I) = \lambda_{\ell_1}||I - \widehat{I}||_1 + \lambda_P d_{\mathrm{P}}(\widehat{I}, I) + \lambda_A D(\widehat{I}),$$
$$I \sim \mathcal{U}_{I_{\mathrm{LQ}}}[\{I_1, \ldots, I_n\}]$$

d_P : perceptual loss (LPIPS),
D : adversarial discriminator

- Reweighted Sampling

$$I \sim \mathcal{P}[S_I \mid \mathrm{SoftMax}_\tau(Q(S_I))],$$
$$Q(S_I) = \{Q(I_1), \ldots, Q(I_n)\},$$
$$S_I = \{I_1, \ldots, I_n\} \in \{A_I, P_I\}$$

- Softmax-All (SMA)
  - Using IQA weight, apply softmax to all GT samples
- Softmax-Positives (SMP)
  - Using IQA weight, apply softmax to all positive GT samples
- Argmax-online(AMO)
  - Using IQA, choose the best one

Direct Optimization using NR-IQA model

- Problem :

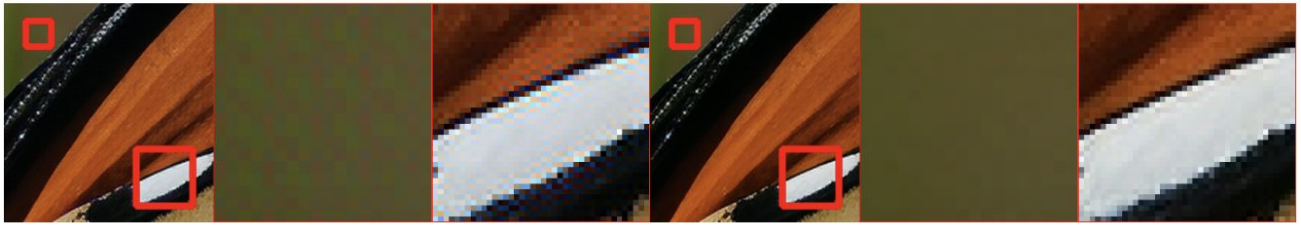


Figure 3. **Structured Optimization Noise.** Optimizing via an NR-IQA metric (MUSIQ [41]) generates structured artifacts (left), similar to an adversarial attack, while utilizing LoRA removes this noise (right; see §4.3 and Supp. §10.2). Zoom in for details.

$$\widetilde{\mathcal{L}}(\phi|\widehat{I}, I) = \mathcal{L}(\phi|\widehat{I}, I) - \lambda_Q Q(\widehat{I}),$$

Experiments

Model		FR Low-Lev. Dist.		FR Mid-Lev. Dist.			NR High-Lev. Perceptual Quality			
		PSNR ↑	SSIM ↑	LPIPS ↓	LPIPS-ST ↓	DISTS ↓	MUSIQ ↑	NIMA ↑	Q-Align ↑	TOPIQ ↑
Gold Standard	✗	–	–	–	–	–	69.64	5.28	3.78	0.69
SwinIR-OrigsOnly	✓	22.72	0.652	0.227	0.174	0.162	59.47	4.87	3.17	0.48
SwinIR-Rand	✓	22.45	0.650	0.180	0.139	0.131	65.27	5.11	3.52	0.59
SwinIR-UPos*	✗	22.30	0.647	0.169	0.129	0.123	66.39	5.16	3.56	0.62
SwinIR-SMA	✓	22.27	0.646	0.171	0.129	0.124	66.73	5.16	3.60	0.63
SwinIR-SMP	✗	22.29	0.647	0.171	0.130	0.124	66.83	5.17	3.62	0.62
SwinIR-AMO	✓	22.08	0.641	0.167	0.124	0.123	68.08	5.21	3.67	0.66
SwinIR-UPos + FT_{HP}	✗	22.17	0.642	0.166	0.123	0.122	68.38	5.23	3.64	0.65
SwinIR-UPos + FT_{IG}	✗	22.03	0.635	0.168	0.122	0.123	69.37	5.24	3.69	0.66
SwinIR-UPos + FT	✗	22.01	0.633	0.169	0.123	0.124	69.70	5.26	3.70	0.67
SwinIR-AMO + FT	✓	21.77	0.624	0.174	0.121	0.128	70.81	5.29	3.75	0.70

- Gold Standard : average of best metric value per quintuplet of test GT.
- Rand : randomly chose GT
- Upos : Uniform sample from positive samples

- FT_{HP} : High perceptual loss (High LPIPS loss)
- FT_{IQ} : Using GAN Loss for finetuning

Limitation

- NR-IQA does not consider artifact

Conclusion

- Need for a comprehensive metric
  - Balancing fidelity and perceptual quality effectively.
  - Evaluating generative artifacts
  - Investigate metrics from text-to-image generation for potential insights.

- Metric-driven model improvement
  - Well-designed metrics could enable model enhancement via backpropagation.
  - Aim for co-development where better metrics lead to better models, and vice-versa.

Thank you!