清華大學
Tsinghua University

UCL

Attentions Help CNNs See Better: Attention-based Hybrid Image Quality Assessment Network

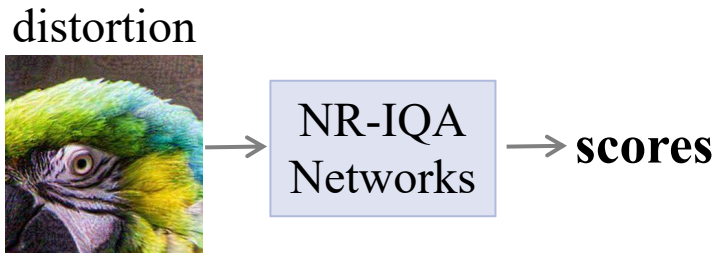
Yuan Gong[1]*, Shanshan Lao[1]*,
Shuwei Shi[1], Sidi Yang[1], Tianhe Wu[1], Jiahao Wang[1], Weihao Xia[2], Yujiu Yang[1§]

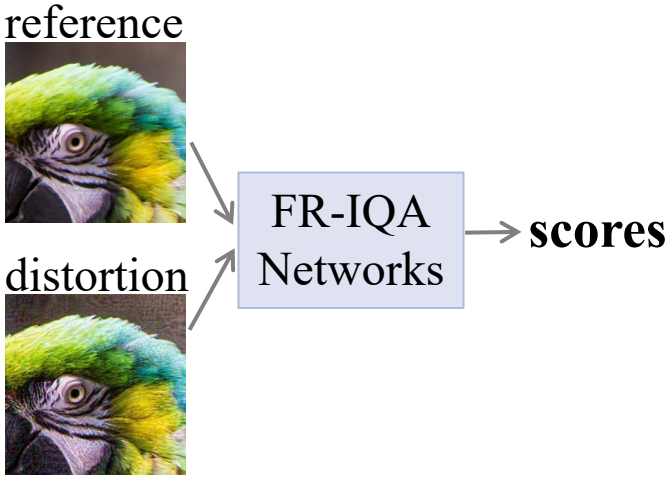
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University; [2]University College London

NTIRE2022 Perceptual Image Quality Assessment Challenge: Track 1 Full-Reference

Image Quality Assessment

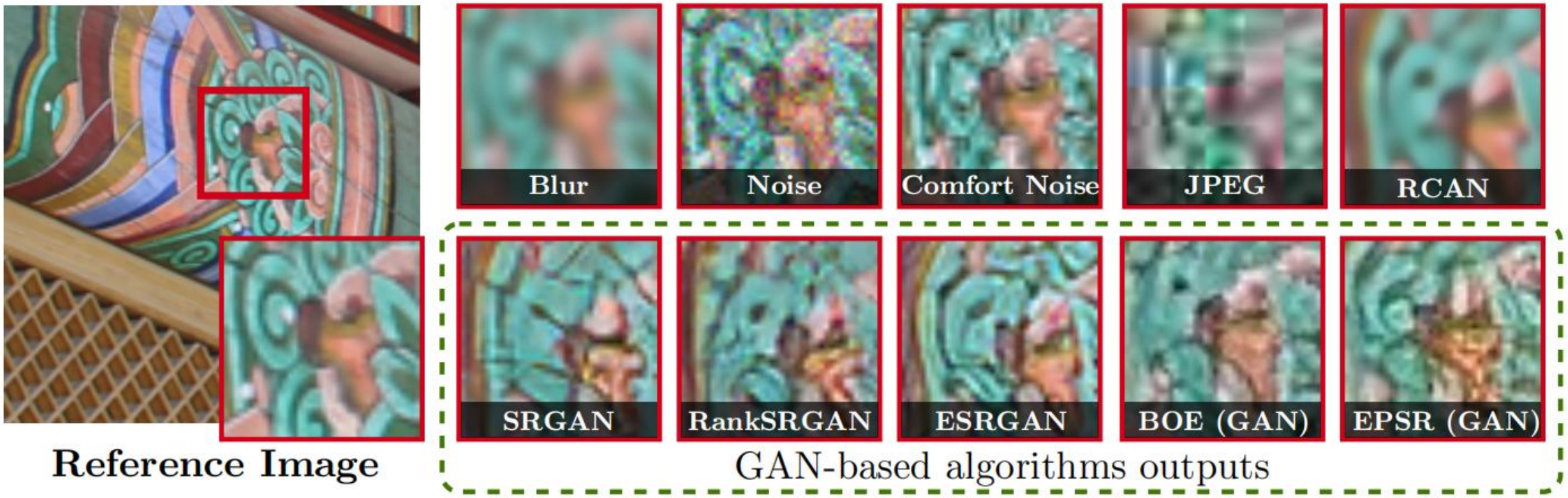
- Image quality assessment (IQA) algorithm aims to **quantify** the human perception of image quality.
- As the "**evaluation mechanism**", IQA plays a critical role in most image processing tasks, such as image super-resolution, denoising, compression and enhancement.
- IQA methods can be divided into **full reference** methods and **no-reference** methods.
- FR-IQA methods take the distortion image and the corresponding reference image as inputs to measure their perceptual similarity.





Challenges

- **GAN-based** methods often fabricate seemingly realistic yet fake **details** and **textures**.
- Human Visual Systems often ignore part of the subtle differences of textures.



[1] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In ECCV, 2020.

Motivation

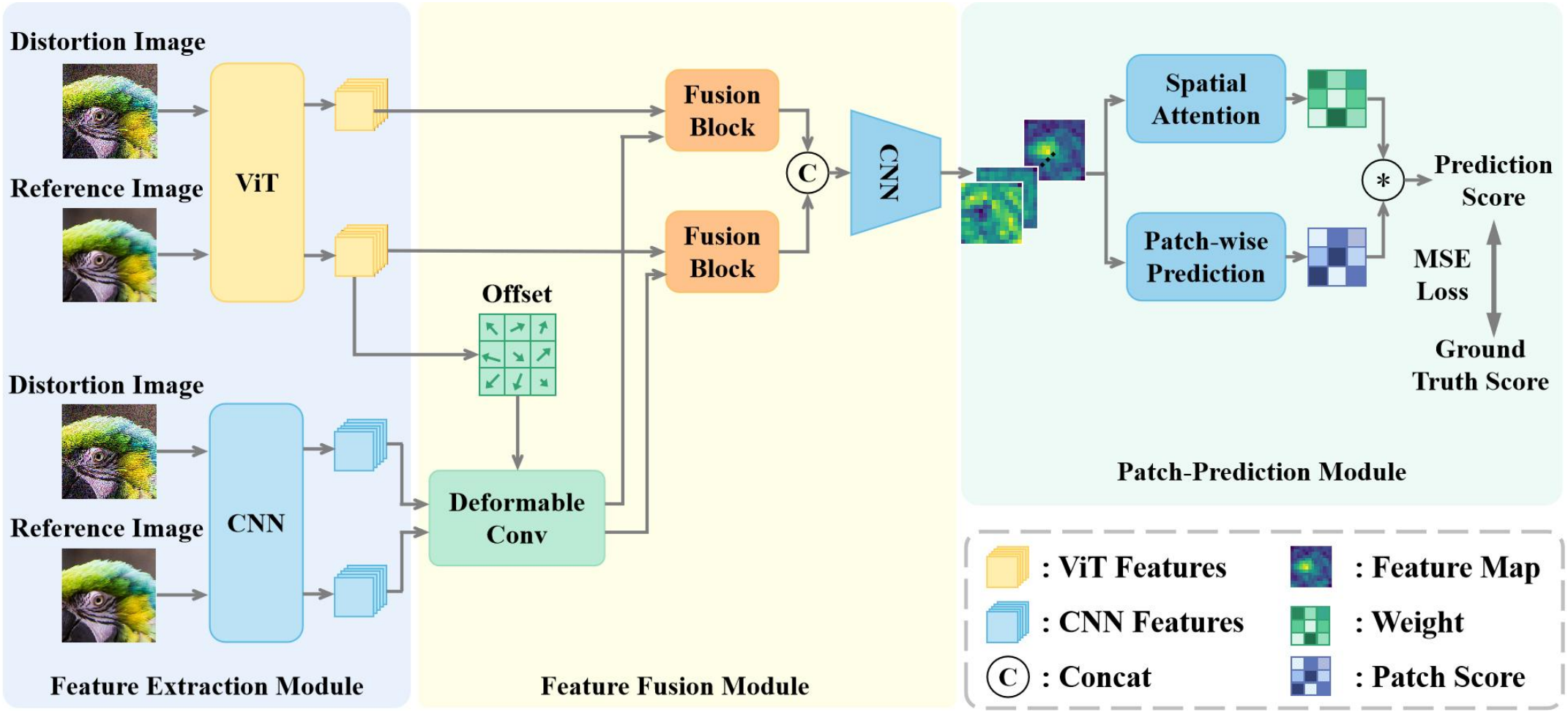
- **Pixel-wise comparison**
  - sensitive to texture misalignment
  - underestimation for GAN-generated images
- **Patch-based prediction**
  - input and calculate each patch separately
  - ignore the context information

Hybrid Network

Methods

- We propose an **Attention-based Hybrid Image Quality Assessment Network** to deal with the challenges and get better performance on the **GAN-based IQA task.**
  - Employ the Vision Transformer to model the relationship and capture long-range dependencies among patches.
  - Shallow CNN features are introduced to add detailed spatial information.
  - Use deformable convolution guided by semantic information from ViT.
  - Use an adaptive weighted scoring mechanism to give a comprehensive assessment.

Architecture

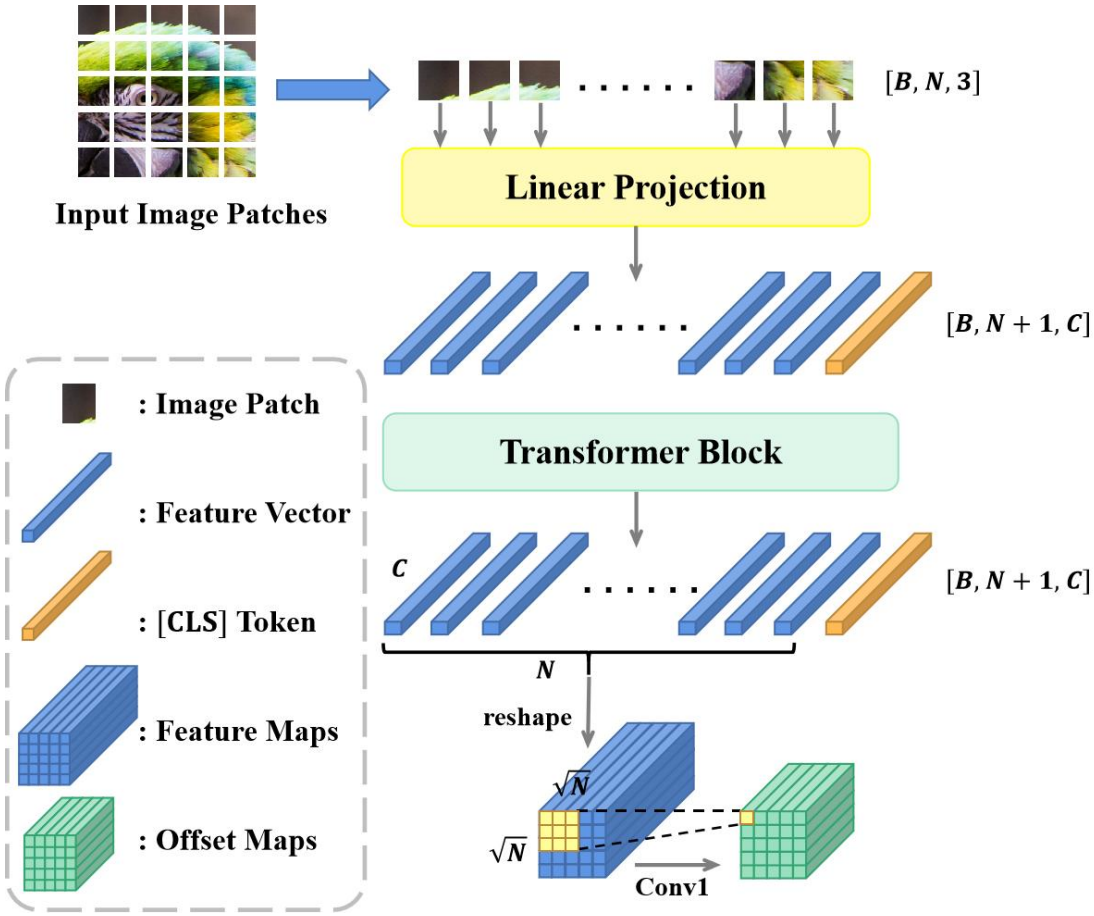


Feature Extraction and Fusion

- **two-branch feature extraction module**
  - **ViT**：global and semantic representations
  - **CNN**：detailed and local information

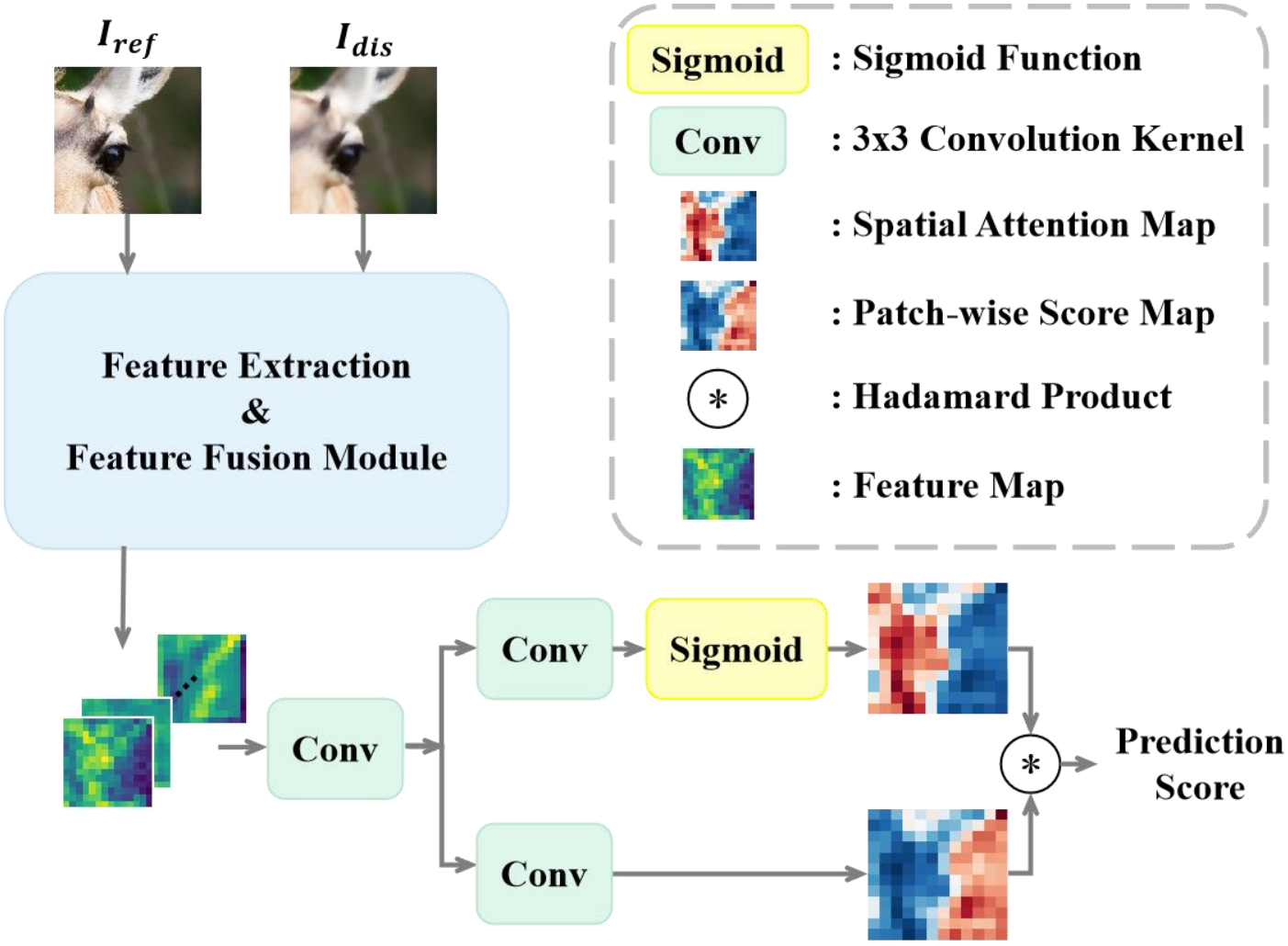
bring noise

- **find the salient region**
  - use feature maps from ViT as guidance
  - preform deformable convolution on CNN feature



Patch-wise Prediction Module

- **two-branch patch-wise prediction module**
  - **prediction:** calculates a score for each pixel in the feature map
  - **spatial attention:** calculates an attention map for each corresponding score
  - **final score:** weighted summation of scores.



Experiment

Datasets

Table 1. IQA datasets for performance evaluation and model training.

Database	# Ref	# Dist	Dist. Type	# Dist. Type	Rating	Rating Type	Env.
LIVE	29	779	traditional	5	25k	MOS	lab
CSIQ	30	866	traditional	6	5k	MOS	lab
TID2013	25	3,000	traditional	25	524k	MOS	lab
KADID-10k	81	10.1k	traditional	25	30.4k	MOS	crowdsourcing
PIPAL	250	29k	trad.+alg.outputs	40	1.13m	MOS	crowdsourcing

PIPAL: includes the results of **GAN-based** IR algorithms

Model Settings

- **data split**: train (60%), test (20%), validate (20%)
- **input image**: random crop 224×224
- **backbone**:

  CNN: ResNet50

  Transformer: ViT-B/16 (traditional); ViT-B/8 (PIPAL)

Performance on traditional IQA datasets

Table 2. Performance comparisons on LIVE, CSIQ, and TID2013 Databases. Performance scores of other methods are as reported in the corresponding original papers and. The best scores are **bolded** and missing scores are shown as "–" dash.

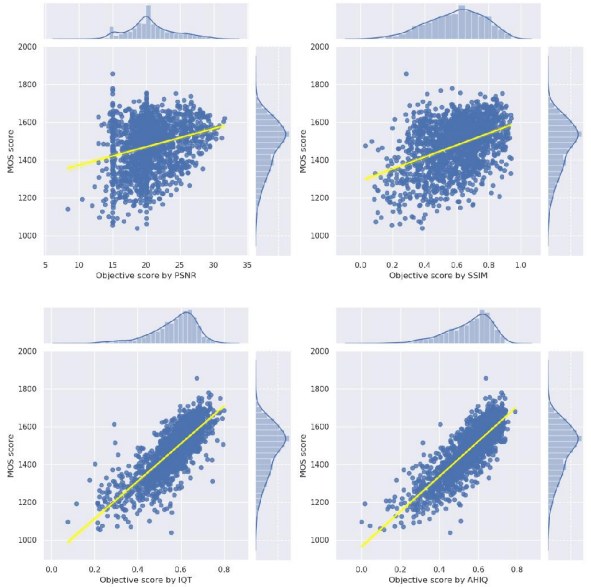
Method	LIVE		CSIQ		TID2013	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
PSNR	0.865	0.873	0.819	0.810	0.677	0.687
SSIM	0.937	0.948	0.852	0.865	0.777	0.727
MS-SSIM	0.940	0.951	0.889	0.906	0.830	0.786
FSIMc	0.961	0.965	0.919	0.931	0.877	0.851
VSI	0.948	0.952	0.928	0.942	0.900	0.897
MAD	0.968	0.967	0.950	0.947	0.827	0.781
VIF	0.960	0.964	0.913	0.911	0.771	0.677
NLPD	0.932	0.937	0.923	0.932	0.839	0.800
GMSD	0.957	0.960	0.945	0.950	0.855	0.804
SCQI	0.937	0.948	0.927	0.943	0.907	0.905
DOG-SSIMc	0.966	0.963	0.943	0.954	0.934	0.926
DeepQA	0.982	0.981	0.965	0.961	0.947	0.939
DualCNN	-	-	-	-	0.924	0.926
WaDIQaM-FR	0.98	0.97	-	-	0.946	0.94
PieAPP	0.986	0.977	0.975	0.973	0.946	0.945
JND-SalCAR	0.987	**0.984**	0.977	**0.976**	0.956	0.949
AHIQ (ours)	**0.989**	**0.984**	**0.978**	0.975	**0.968**	**0.962**

Performance on PIPAL

Table 4. Performance comparison of different IQA methods on PIPAL dataset. AHIQ-C is the ensemble version we used for the NTIRE 2022 Perceptual IQA Challenge.

Method	Validation		Test	
	PLCC	SROCC	PLCC	SROCC
PSNR	0.269	0.234	0.277	0.249
NQM	0.364	0.302	0.395	0.364
UQI	0.505	0.461	0.450	0.420
SSIM	0.377	0.319	0.391	0.361
MS-SSIM	0.119	0.338	0.163	0.369
RFSIM	0.285	0.254	0.328	0.304
GSM	0.450	0.379	0.465	0.409
SRSIM	0.626	0.529	0.636	0.573
FSIM	0.553	0.452	0.571	0.504
VSI	0.493	0.411	0.517	0.458
NIQE	0.129	0.012	0.132	0.034
MA	0.097	0.099	0.147	0.140
PI	0.134	0.064	0.145	0.104
Brisque	0.052	0.008	0.069	0.071
LPIPS-Alex	0.606	0.569	0.571	0.566
LPIPS-VGG	0.611	0.551	0.633	0.595
DISTS	0.634	0.608	0.687	0.655
IQT	0.840	0.820	0.799	0.790
AHIQ (ours)	0.845	0.835	0.823	0.813
AHIQ-C (ours)	**0.865**	**0.852**	**0.828**	**0.822**

Scatter plots of the objective scores vs. the MOS scores:



Higher correlation means better performance of the IQA method.

Table 5. Performance comparison for cross-database evaluations.

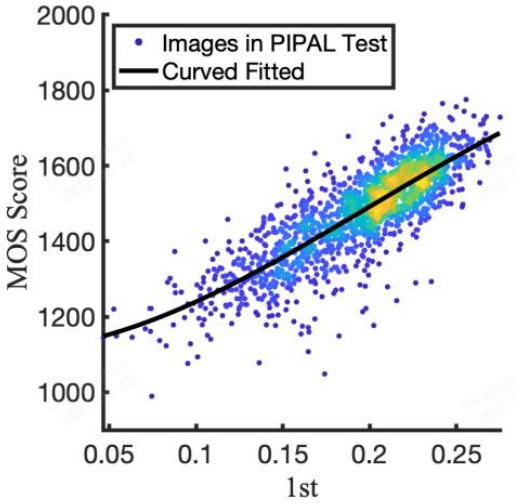
Method	LIVE	CSIQ	TID2013
	PLCC/SROCC	PLCC/SROCC	PLCC/SROCC
PSNR	0.865/0.873	0.786/0.809	0.677/0.687
WaDIQaM	0.837/0.883	-/-	0.741/0.698
RADN	0.878/0.905	-/-	0.796/0.747
AHIQ (ours)	**0.911/0.920**	**0.861/0.865**	**0.804/0.763**

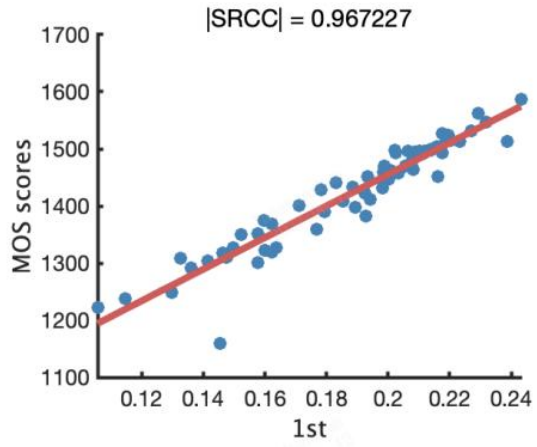
NTIRE 2022 Perceptual IQA Challenge

Table 9. The results of NTIRE 2022 challenge FR-IQA track on the testing dataset. This table only shows part of the participants and best scores are **bolded**.

Method	PLCC	SROCC	Main Score
Ours	**0.828**	**0.822**	**1.651**
2nd	0.827	0.815	1.642
3rd	0.823	0.817	1.64
4th	0.775	0.766	1.541
5th	0.772	0.765	1.538

$$\text{Main Score} = |\text{SRCC}| + |\text{PLCC}|.$$





Ablation study

- **Feature fusion strategy**

Table 6. Comparison of different feature fusion strategies on the NTIRE 2022 IQA Challenge testing datasets. CNN refers to Resnet50 and ViT refers to ViT-B/8 in this experiment.

No.	Feature		Fusion Method	PLCC	SROCC
	CNN	ViT			
1	✓	✓	deform+concat	**0.823**	**0.813**
2	✓	✓	concat	0.810	0.799
3	✓		-	0.792	0.789
4		✓	-	0.799	0.788

- **Backbones**

Table 7. Comparison of different feature extraction backbones on the NTIRE 2022 IQA Challenge testing datasets.

CNN	ViT	PLCC	SROCC	Main Score
Resnet50		**0.823**	**0.813**	**1.636**
Resnet101		0.802	0.788	1.590
Resnet152	ViT-B/8	0.807	0.793	1.600
HRnet		0.806	0.796	1.601
IncepResV2		0.806	0.793	1.599
Resnet50	ViT-B/16	0.811	0.803	1.614

- **Pooling strategy**

Table 8. Comparison of different pooling strategy on the NTIRE 2022 IQA Challenge testing datasets. Note that "Patch" denotes the patch-wise prediction and "Spatial" denotes the spatial pooling.

Pooling Strategy	PLCC	SROCC	Main Score
Patch	**0.823**	**0.813**	**1.636**
Spatial	0.794	0.795	1.589
Patch + Spatial	0.801	0.791	1.593

Conclusion

- Propose a hybrid network called AHIQ for full-reference IQA task. It takes advantage of the long-term relationship modeling ability of ViT and the local texture information from CNN.

- AHIQ not only outperforms the SOTA methods on standard datasets, but also has a strong generalization ability on unseen samples and hard samples, especially GAN-based distortions. The ensembled version ranked **first place** in the FR track of the NTIRE 2022 Perceptual Image Quality Assessment Challenge.

Thanks!

https://github.com/IIGROUP/AHIQ