

Cross-IQA: Unsupervised Learning for Image Quality Assessment

Zhen Zhang, Xiaolong Jiang, Chenyi Zhao, Shuang Qiao, and Tian Zhang

Abstract—Automatic perception of image quality is a challenging problem that impacts billions of Internet and social media users daily. To advance research in this field, we propose a no-reference image quality assessment (NR-IQA) method termed Cross-IQA based on vision transformer(ViT) model. The proposed Cross-IQA method can learn image quality features from unlabeled image data. We construct the pretext task of synthesized image reconstruction to unsupervised extract the image quality information based ViT block. The pretrained encoder of Cross-IQA is used to fine-tune a linear regression model for score prediction. Experimental results show that Cross-IQA can achieve state-of-the-art performance in assessing the low-frequency degradation information (e.g., color change, blurring, etc.) of images compared with the classical full-reference IQA and NR-IQA under the same datasets.

Index Terms—Image quality assessment, unsupervised learning, vision transformer.

I. INTRODUCTION

With the advent of the mobile Internet era, a great number of digital images are shared daily on social media platforms such as Instagram, Snapchat, Flickr, and others. Image quality score, an essential metric of digital images [1], can help operators screen and provide high-quality digital images to their subscribers. Therefore, developing an image quality assessment (IQA) method that is highly consistent with the subjective perception of human vision becomes extremely important [2].

Since the IQA technique can predict the quality scores of digital images, it is also widely used in a series of digital image processing tasks such as image restoration [3], image super-resolution reconstruction [4], and so on. Based on dependencies with reference images, IQA can be categorized into full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA), and no-reference IQA (NR-IQA). So far, FR-IQA is the most widely used IQA method (e.g., peak signal to noise ratio (PSNR), structural similarity index measure (SSIM), etc.) [5], which can obtain the quality score by calculating the similarities and differences of the distorted image and the reference image. Unlike FR-IQA requiring the complete reference image, RR-IQA evaluates the quality of the distorted image using only partial information of the reference image

Manuscript received XXXX, XX, 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 11905028, 12105040 and in part by the Scientific Research Project of Education Department of Jilin Province under Grant JJKH20231294KJ. (Corresponding author: Tian Zhang).

Zhen Zhang, Xiaolong Jiang, Chenyi Zhao, Shuang Qiao and Tian Zhang are with the School of Physics, Northeast Normal University, Changchun 130024, China (e-mail:zhangt100@nenu.edu.cn).

(e.g., image entropy, gray histogram, and some transform-domain parameters), which can enhance the flexibility of IQA methods [6]. However, ideal reference images are usually not available in many real photographic environments. Therefore, the more challenging NR-IQA becomes a research hotspot [7].

Initially, traditional NR-IQA methods are mainly based on the assumption that the original image has a specific statistical distribution and the distortions alter this underlying distribution [8]. Therefore, feature extraction methods designed for some special features are used to evaluate the quality of images. Limited by the artificial features and feature extraction methods, traditional NR-IQA methods still have room for improvement. As deep learning (DL) technology has obtained remarkable advancements in various visual tasks such as image classification [9], semantic segmentation [10], and image recognition [11], researchers have started to explore the DL-based NR-IQA methods [12]–[14]. Making full use of the excellent capability of feature extraction, DL-based NR-IQA achieves the significantly improved accuracy compared with those traditional manual feature extraction-based NR-IQA methods. Although the DL-based NR-IQA method has achieved remarkable success, the mainstream NR-IQA methods are based on supervised learning, which require large-scale image datasets with quality labels. Actually, providing quality scores to those datasets is indeed labor-intensive and costly. Therefore, the lack of large-scale IQA datasets is still a problem for the application of supervised learning-based NR-IQA.

To solve these problems, we propose a vision transformer (ViT)-based unsupervised image quality assessment method in this letter. Although the annotation process for the large-scale dataset is difficult, digital images with different degradation levels can be easily synthesized. Based on this, we design a pretext task of reconstructing images with varying levels of quality by training two parameter-shared encoders and two parameter-shared decoders. The well-trained encoder can effectively extract the image quality information in an unsupervised environment.

II. METHOD

A. Cross-IQA

Cross-IQA is a novel method designed for extracting image quality information in an unsupervised environment by image reconstruction and information exchange. Unlike those supervised learning-based NR-IQA using the Imagenet database, the proposed Cross-IQA model is deployed on a synthetic degraded dataset based on the Waterloo database.

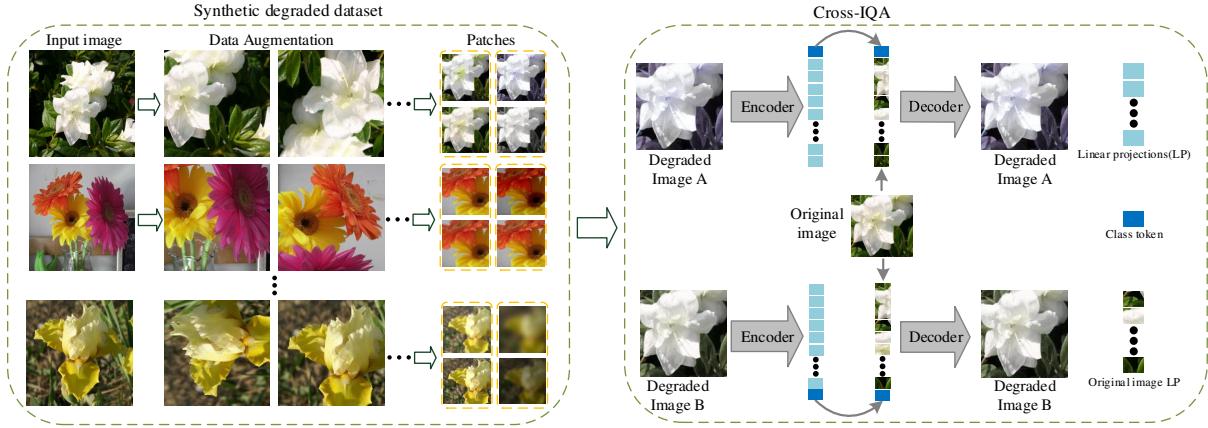


Fig. 1. Schematic diagram of the proposed Cross-IQA.

The schematic diagram of the proposed Cross-IQA is shown in Fig. 1. Specifically, the first two degraded images with different quality levels are put through two parameter-shared encoders for feature extraction. The obtained class token and linear projection of the original image are thereafter fed into the decoder for reconstructing the input degraded images. In the end, the well trained encoder can be used as a backbone for extracting the quality information to realize the subsequent no-reference image quality assessment.

1) *Cross-IQA encoder*: The standard ViT block [15] is used to construct the encoder of Cross-IQA. Firstly, positional embeddings and class token are first added to the input linear projections of the synthetic degraded images. Then, the processed linear projections are fed into a series of Transformer blocks for feature extraction. The reason why we choose ViT as the encoder and decoder can be summarized as:(i)ViT can efficiently extract features to class token, which is more convenient for feature exchange; (ii)The ViT-based encoder and decoder do not require skip connections, which is more convenient for the subsequent linear regression model.

2) *Cross class token*: After obtaining the class token by the encoder, it is thereafter connected to the linear projections of the original image to reconstruct the degraded images. This operation realizes the exchange of image quality information during the process of image reconstruction. It is an essential step for the proposed Cross-IQA method to unsupervised extract the expected image quality information.

3) *Cross-IQA decoder*: The ViT block used for decoder is the same as the encoder but the number is different. The constructed decoder is used to realize the image reconstruction task. The input of the decoder consists of the linear projection of the original image and exchanged class token. Note that, the decoder only performs the image reconstruction task in the pre-training phase of Cross-IQA.

4) *Reconstruction target*: For the reconstruction task, we use a loss function of mean square error (MSE) to quantify the discrepancy between the reconstructed and original images.

$$MSE = \frac{\sum_{i=1}^n (y - f(x_i))^2}{n}, \quad (1)$$

where MSE represents the average of the squared distances between the predicted value of $f(x)$ and the ground truth of y .

At the same time, to better realize the exchange and extraction of image quality information in the proposed Cross-IQA framework, the mean absolute error (MAE) is also employed as a metric to calculate the difference between the input two degraded images and two reconstructed images.

$$MAE = \frac{\sum_{i=1}^n (y - f(x_i))}{n}, \quad (2)$$

where MAE denotes the average distance between the predicted value of $f(x)$ and the the ground truth of y .

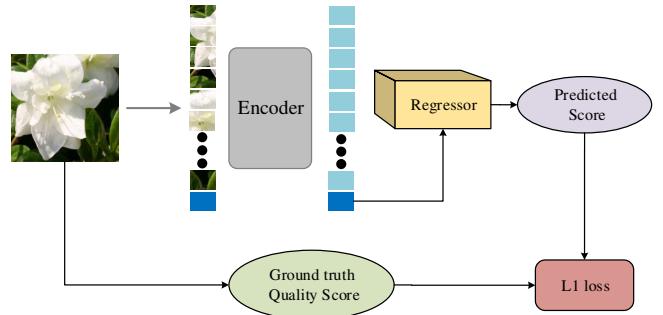


Fig. 2. Process of Cross-IQA Regression.

B. Cross-IQA Regression

After the pre-training phase, the well trained encoder of Cross-IQA is frozen. Then, the standard IQA database is thereafter used to train a regressor to map the obtained features by the encoder to the final perceptual image quality scores and the loss function is set as L1 loss, as shown in Fig. 2. Note that, the obtained class token by the encoder is the regressor head. Two public databases, including LIVE and TID2013 are used to fine-tune the linear regression model for score prediction [16], [17]. LIVE contains 779 distorted images synthesized from 29 reference images while TID2013 comprises 3000 distorted images synthesized from 25 reference images

TABLE I
PLCC COMPARISON ON INDIVIDUAL DISTORTION TYPES.

Database	Type	FR-IQA				NR-IQA				Cross-IQA
		PSNR	SSIM	BRISQUE	CORNIA	GMLOG	DIQA	RANK	DFIA	
TID2013	#01 AGN	0.934	0.867	0.630	0.550	0.748	0.915	0.667	0.718	0.641
	#02 ANC	0.867	0.773	0.424	0.209	0.591	0.755	0.620	0.728	0.597
	#03 SCN	0.916	0.852	0.727	0.717	0.769	0.878	0.821	0.839	0.712
	#04 MN	0.836	0.777	0.321	0.360	0.491	0.734	0.365	0.581	0.597
	#05 HFN	0.913	0.863	0.775	0.797	0.875	0.939	0.760	0.830	0.575
	#06 IN	0.900	0.750	0.669	0.585	0.693	0.843	0.736	0.914	0.653
	#07 QN	0.875	0.866	0.592	0.727	0.833	0.858	0.783	0.824	0.861
	#08 GB	0.910	0.967	0.845	0.840	0.878	0.920	0.809	0.854	0.995
	#09 DEN	0.953	0.925	0.553	0.721	0.721	0.788	0.767	0.821	0.972
	#10 JPEG	0.922	0.920	0.742	0.806	0.823	0.892	0.866	0.902	0.632
	#11 JP2K	0.886	0.947	0.799	0.800	0.872	0.912	0.878	0.924	0.957
	#12 JGTE	0.806	0.845	0.301	0.595	0.400	0.861	0.704	0.767	0.914
	#13 J2TE	0.891	0.883	0.672	0.654	0.731	0.812	0.810	0.859	0.754
	#14 NEPN	0.679	0.782	0.175	0.157	0.190	0.659	0.512	0.455	0.31
	#15 Block	0.330	0.572	0.184	0.016	0.318	0.407	0.622	0.683	0.574
	#16 MS	0.757	0.775	0.155	0.177	0.119	0.299	0.268	0.331	0.712
	#17 CTC	0.447	0.378	0.125	0.262	0.224	0.687	0.613	0.761	0.891
	#18 CCS	0.634	0.414	0.032	0.170	-0.121	-0.151	0.662	0.685	0.777
	#19 MGN	0.883	0.780	0.560	0.407	0.701	0.904	0.619	0.700	0.52
	#20 CN	0.841	0.857	0.282	0.541	0.202	0.655	0.644	0.725	0.747
	#21 LCNI	0.916	0.806	0.680	0.696	0.664	0.930	0.800	0.829	0.945
	#22 ICQD	0.820	0.854	0.804	0.649	0.886	0.936	0.779	0.837	0.652
	#23 CHA	0.880	0.878	0.715	0.689	0.648	0.756	0.629	0.780	0.962
	#24 SSR	0.911	0.946	0.800	0.874	0.915	0.909	0.859	0.915	0.968
	Low Frequency	0.831	0.815	0.521	0.611	0.565	0.723	0.721	0.784	0.863
	Other	0.808	0.787	0.527	0.445	0.626	0.797	0.65	0.729	0.583

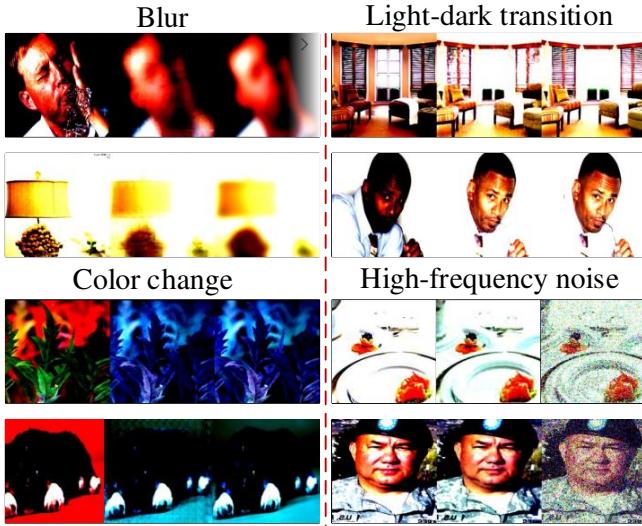


Fig. 3. Example results on Waterloo database validation images. For each triplet, we show the original image (left), the reconstructed image by decoder of Cross-IQA (middle), and the synthetic degraded image (right).

with 24 synthetic distortion types and 5 degradation levels for each reference image.

III. EXPERIMENTS

The experiments are carried out on a workstation with a 8-core AMD Ryzen 7 CPU and an NVIDIA GeForce RTX 4090 GPU. We compared the performance of the proposed Cross-IQA method with the state-of-the-art IQA methods on public databases.

A. Implementation details

Pre-training. The proposed Cross-IQA was pre-trained on the synthetic degraded dataset (i.e., Waterloo database) described below. After balancing the time and resource constraints, we train the encoder of Cross-IQA by 200 epochs. The input image is randomly cropped to a size of 224×224 pixels and undergoes a random folding with the probability of 0.5 and normalized operation. AdamW is used as the optimizer, which is the common choice for ViT training. The number of ViT blocks is set to 12 and 8 respectively for the encoder and decoder. The batchsize is set to 16 and the learning rate is $0.0001 * \text{batchsize}/256$. Fig. 3 illustrates the original image, the reconstructed image by the decoder of Cross-IQA, and the synthetic degraded image, which indicates the proposed Cross-IQA is more sensitive to the low-frequency degradation information (e.g., color change, blurring, light-dark transition, etc.) than the high-frequency information.

Fine-tuning. In our experiment, the distorted images of a target IQA database (e.g., LIVE, TID2013) are divided into two portions: 80% is used for finetuning the Cross-IQA, and the remainder is used for testing. Note that, the database is divided according to the reference image to guarantee image content independence between the training and testing sets. Besides, the experiments are repeated ten times to obtain a fair evaluation, and the mean values of Spearman rank-order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC) are reported as the final results [5]. All training samples are randomly flipped left and right with a probability of 0.5 and cropped to images with the size of 224×224 for data augment.

TABLE II
GENERALIZATION ABILITY OF GB PRETRAINED-BASED CROSS-IQA ON TID2013.

Pre-training	Distortions with blur property	PLCC	SROCC
Training in GB	#07 QN	0.861	0.854
	#08 GB	0.995	0.988
	#09 DEN	0.963	0.916
	#11 JP2K	0.911	0.871
	#12 JGTE	0.914	0.785
	#20 CN	0.747	0.750
	#21 LCNI	0.945	0.891
	#23 CHA	0.967	0.921
	#24 SSR	0.968	0.905

TABLE III
PLCC AND SROCC COMPARISON ON THE LIVE DATABASE.

	PLCC	JP2K	JPEG	GN	GB	FF
FR-IQA	PSNR	0.873	0.876	0.926	0.779	0.87
	SSIM [18]	0.921	0.955	0.982	0.893	0.939
	FSIM [19]	0.91	0.985	0.976	0.978	0.912
NR-IQA	DIVINE [20]	0.922	0.921	0.988	0.923	0.888
	BLIINDS-II [21]	0.935	0.968	0.98	0.938	0.896
	BRISQUE [22]	0.923	0.973	0.985	0.951	0.903
	CORNIA [23]	0.951	0.965	0.987	0.968	0.917
	CNN [24]	0.953	0.981	0.984	0.953	0.933
	SOM [25]	0.952	0.961	0.991	0.974	0.944
Cross-IQA						
	SROCC	JP2K	JPEG	GN	GB	FF
FR-IQA	PSNR	0.87	0.885	0.942	0.763	0.874
	SSIM [18]	0.939	0.946	0.964	0.907	0.941
	FSIM [19]	0.97	0.981	0.967	0.972	0.949
NR-IQA	DIVINE [20]	0.913	0.91	0.984	0.921	0.863
	BLIINDS-II [21]	0.929	0.942	0.969	0.923	0.889
	BRISQUE [22]	0.914	0.965	0.979	0.951	0.887
	CORNIA [23]	0.943	0.955	0.976	0.969	0.906
	CNN [24]	0.952	0.977	0.978	0.962	0.908
	SOM [25]	0.947	0.952	0.984	0.976	0.937
Cross-IQA						
	0.971	0.755	0.769	0.985	0.941	

B. Performance comparison

Evaluation on TID2013. For testing on TID2013, we generated 17 out of a total of 24 distortions (except for #03, #04, #07, #12, #13, #20, #21, and #24 as shown in Table I). For the distortions that we could not generate, we fine-tuned the network according to the other trained distortions. During the testing process for each distortion type, only the corresponding synthetic distortion or similar distortion datasets with 5 distortion levels are used for the Cross-IQA pre-training. The original images used for constructing the synthetic degraded dataset were derived from the Waterloo database. For the TID2013 dataset with low-frequency distortion types (i.e., #07 #08 #09 #10 #11 #12 #13 #16 #17 #18 #20 #21 #23 #24), our Cross-IQA method substantially outperforms the current state-of-the-art NR-IQA method in terms of PLCC, and matches or even exceeds the results of some FR-IQAs for some distortion types. Next, the Cross-IQA is pretrained with only the synthetic distortion of #08 Gaussian blur (GB), the PLCC and SROCC results on the distortion types containing blur property from TID2013 dataset further demonstrate its generalization ability as shown in Table II.

Evaluation on LIVE. To test on the LIVE database, we generated the training datasets with five distortion types

TABLE IV
SROCC AND PLCC FOR ViT USING DIFFERENT PRE-TRAINING METHODS.

Method	TID2013		LIVE	
	PLCC	SROCC	PLCC	SROCC
ViT(Baseline)	0.285	0.258	0.385	0.362
ViT(ImageNet)	0.369	0.385	0.587	0.535
ViT(Cross-IQA)	0.863	0.842	0.976	0.965

including GB, Gaussian noise (GN), JPEG, JP2K and fast-fading (FF). The pre-training method used for LIVE was the same as the above TID2013. Table 3 shows that Cross-IQA has a significant advantage over other methods in evaluating features with low-frequency degradation, such as GB, JP2K, FF [18]–[25]. That means the proposed Cross-IQA without any labeled data for pre-training outperforms the existing NR-IQA and state-of-the-art FR-IQA methods in low-frequency degradation assessment.

The main reason for such excellent results is that ViT and its variants are good at capturing the low-frequency information of visual data including the global shape and structure of a scene or object. On the contrary, high-frequency information such as local edges and textures are usually not skillful [15]. This can be explained intuitively: self-attention, the main operation used in ViTs for exchanging information between non-overlapping patch tokens is a global operation, which is more suitable to capture global information (i.e., low-frequency) than local information (i.e., high-frequency). Therefore, for the low-frequency degradation information, the proposed Cross-IQA outperforms the state-of-the-art NR-IQA algorithm.

Baseline performance analysis. In this section, different pre-training methods are chosen to validate the effectiveness of Cross-IQA including the ViT network initialized with random parameters termed ViT baseline, ViT network pre-trained using ImageNet termed ViT ImageNet, and the proposed Cross-IQA as shown in Table IV. The fine-tuning methods used therein are all the same as in the experiment above. All experiments were performed ten times, and the average SROCC and PLCC were given in Table IV. The proposed unsupervised Cross-IQA achieves quite good results in the evaluation of low-frequency degradation features of images. Since the proposed pretraining method uses unlabeled data to realize IQA, it has good potential for NR-IQA method with limited IQA data.

IV. CONCLUSIONS

In this letter, we proposed a novel no-reference image quality assessment method termed Cross-IQA based on the vision transformer model, which can extract the image quality information from unlabeled image data. Extensive experimental results demonstrate that the proposed Cross-IQA outperforms the classical full-reference IQA and NR-IQA in terms of low-frequency degradation information of images, thus demonstrating its application potential.

REFERENCES

- [1] G. Zhai and X. Min, "Perceptual image quality assessment: A survey," *Sci.China Inf. Sci.*, vol. 63, pp. 1-52, 2020.
- [2] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol .15,no. 11, pp. 3440-3451, Nov. 2006.
- [3] H. Zhao, O. Gallo, I. Frosio and J. Kautz, "Loss Functions for Image Restoration With Neural Networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47-57, Mar. 2017.
- [4] C. Dong, C. C. Loy, K. He and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295-307, Feb. 2016.
- [5] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, Apr. 2004.
- [6] Yutao Liu, Guangtao Zhai, Ke Gu, Xianming Liu, Debin Zhao, and Wen Gao, "Reduced-reference image quality assessment in free-energy principle and sparse representation," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 379-391, Feb. 2018, doi: 10.1109/TMM.2017.2729020.
- [7] S. Bosse, D. Maniry, K. R. Miller, T. Wiegand and W. Samek, "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206-219, Jan. 2018, doi: 10.1109/TIP.2017.2760518.
- [8] G. Cheron, I. Laptev and C. Schmid, "P-CNN: Pose-Based CNN Features for Action Recognition," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 3218-3226, doi: 10.1109/ICCV.2015.368.
- [9] F. Wang et al., "Residual Attention Network for Image Classification," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 6450-6458, doi: 10.1109/CVPR.2017.683.
- [10] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 3431-3440, doi: 10.1109/CVPR.2015.7298965.
- [11] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
- [12] J. Kim and S. Lee, "Fully Deep Blind Image Quality Predictor," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206-220, Feb. 2017.
- [13] H. Zhu, L. Li, J. Wu, W. Dong and G. Shi, "MetaIQA: Deep Meta-Learning for No-Reference Image Quality Assessment," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),Seattle, WA, USA, 2020, pp.14131-14140, doi: 10.1109/CVPR42600.2020.01415.
- [14] Q. Yan, D. Gong and Y. Zhang, "Two-Stream Convolutional Networks for Blind Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2200-2211, May 2019, doi: 10.1109/TIP.2018.2883741.
- [15] W. Wang et al., "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 548-558.
- [16] H. Sheikh, "Live image quality assessment database release 2," 2005. [Online]. Available: <http://live.ece.utexas.edu/research/quality>
- [17] N. Ponomarenko et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, Vol. 30,2015,Pages 57-77,ISSN 0923-5965.
- [18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, Apr. 2004.
- [19] L. Zhang, L. Zhang, X. Mou, and D. Zhang. "Fsim: a feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378-2386, Aug. 2011, doi: 10.1109/TIP.2011.2109730.
- [20] A. K. Moorthy and A. C. Bovik. "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350-3364, Dec. 2011, doi: 10.1109/TIP.2011.2147325.
- [21] M. A. Saad, A. C. Bovik, and C. Charrier. "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339-3352, Aug. 2012, doi: 10.1109/TIP.2012.2191563.
- [22] A. Mittal, A. K. Moorthy, and A. C. Bovik. "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695-4708, Dec. 2012, doi: 10.1109/TIP.2012.2214050.
- [23] P. Ye, J. Kumar, L. Kang, and D. Doermann. "Unsupervised feature learning framework for no-reference image quality assessment," 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 1098-1105, doi: 10.1109/CVPR.2012.6247789.
- [24] L. Kang, P. Ye, Y. Li, and D. Doermann. "Convolutional neural networks for no-reference image quality assessment," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1733-1740, doi: 10.1109/CVPR.2014.224.
- [25] P. Zhang, W. Zhou, L. Wu, and H. Li. "Som: Semantic obviousness metric for image quality assessment," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 2394-2402, doi: 10.1109/CVPR.2015.7298853.

