BALANCING EASY AND HARD DISTORTIONS: A MULTI-RATE KNOWLEDGE DISTILLATION STRATEGY FOR BLIND IMAGE QUALITY ASSESSMENT

Desen Yuan[⋆†]

University of Electronic Science and Technology of China
desenyuan@gmail.com

ABSTRACT

In the evolving fields of computer vision and image processing, Image Quality Assessment (IQA) has become essential due to the prevalence of digital images in today's applications. Our comprehensive study underscores that current IQA models demonstrate varied learning aptitudes towards different image distortions. Notably, these models, employing a uniform learning rate, often yield suboptimal results for certain challenging distortions, affecting the overall evaluation precision. Addressing this challenge, we present an innovative online knowledge distillation strategy named Multi-Rate Knowledge Distillation (MRKD). Our approach fosters the student model to assimilate diverse features from the teacher model, leveraging self-distillation regularization to enhance its generalization capability while enabling the student model to circumvent the pitfalls of local optima. This approach leverages two models with varying learning rates, wherein a high learning rate teacher model mentors a student model with a lower rate. Extensive testing on the TID2013, KADID-10k, and LIVEC datasets has validated the efficacy of our MRKD approach, demonstrating its potential in enhancing performance for challenging distortion types.

Index Terms— Blind image quality assessment, Knowledge Distillation, Learning Rate, Hard Distortions

1. INTRODUCTION

With the rise in digital media, Image Quality Assessment (IQA) [1, 2, 3, 4, 5, 6, 7, 8, 9] has become pivotal. As digital images play an indispensable role in numerous applications, ensuring their quality is paramount for reliable interpretations and applications. IQA has emerged as a central research topic in the fields of computer vision and image processing.

Our observations, illustrated in Figure 1, indicate that there exists a pronounced variance in performance for different distortion types when models are trained at distinct learning rates, highlighting an underexplored area in IQA research. Through a series of experiments, we delved deeply into this phenomenon. A key discovery was that distinct distortion types manifest differential learning outcomes under different learning rates.

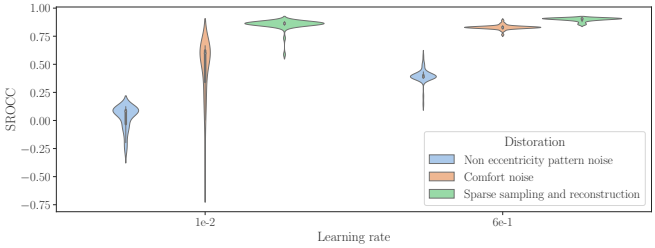

Fig. 1: Performances' violin plot of models trained with large learning rates 1e-2 and models trained with small learning rates 6e-1 at certain distortion types on TID2013. w.r.t each epoch.

Non eccentricity pattern noise
MOS=4.1025

Additive Gaussian noise
MOS=4.1578

Fig. 2: Pictures of similar MOS in different distortion categories on TID dataset.

While traditional IQA models often employ a uniform learning rate, our approach challenges this norm, hypothesizing that different learning rates can be tailored for optimal performance across various distortion types. This raised a question: Why do most existing IQA models adopt a uniform learning rate, overlooking the potential and necessity for targeted learning? In fact, when models process various distortion types under a unified learning rate, the learning outcomes for some challenging distortion types are suboptimal, leading to limited evaluation accuracy.

Building upon these observations, we introduce a novel knowledge distillation [10, 11] method for IQA named Multi-Rate Knowledge Distillation (MRKD). In our proposed methodology, we initialize two distinct models. The teacher model, trained with a higher learning rate, is designed to ex-

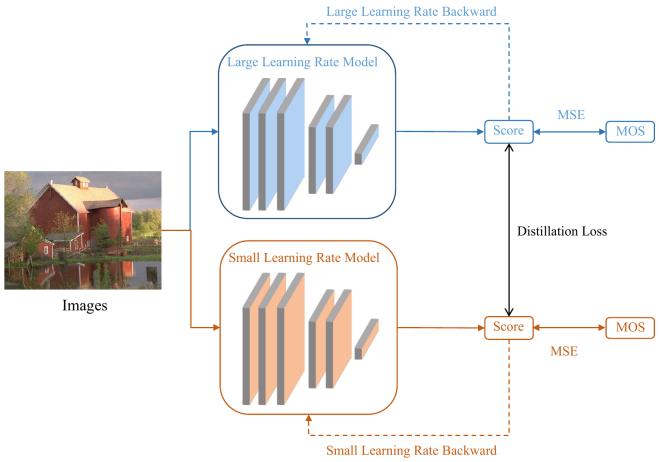

Fig. 3: Framework of our Multi-Rate Knowledge Distillation (MRKD) method

plore a more extensive range of parameter spaces, capturing a wide spectrum of features. In contrast, the student model, optimized with a more conservative learning rate, focuses on achieving consistent and stable performance. By doing so, we aim to blend the robustness of a model trained with a smaller learning rate with the expansive feature understanding of a model trained with a larger one.

Our method offers two significant advantages. First, it enables the student model to learn diverse features [12] from the teacher model by using high learning rates that can accentuate the tiny features that are difficult for models trained with low learning rates to acquire. Second, it employs self-distillation regularization terms to assist the model in escaping local minima and enhancing generalization performance [13, 14].

As shown in Fig. 2, Non-eccentric mode noise presents more subtle distortion characteristics than additive Gaussian noise, even when both manifest similar MOS scores. Given that these models provide complementary insights, our proposed method combines the strengths of both, producing a more comprehensive IQA model through a distillation process.

To validate the effectiveness of our proposed strategy, we conducted extensive experiments on the widely recognized TID2013 [15], KADID-10k [16] and LIVEC [17] datasets. Experimental results reveal that, when compared to the DBCNN [18], HyperIQA [19] and ResNet50 [20] models, our MRKD not only achieves superior performance on challenging distortion types but also significantly elevates the overall evaluation accuracy.

2. METHOD

The general framework of Multi-Rate Knowledge Distillation (MRKD) method is shown in Figure 3. Given a natural image I, a distortion category d and a distortion level s, we can use a function $f_d(I, s)$ to generate a distorted image x:

$$x = f_d(I, s). \tag{1}$$

Repeat the process by giving different pictures, we get a training dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$, where n is the number of training samples and (x_i, y_i) are labeled data pairs. x_i is a distorted picture, and y_i is the corresponding picture mean opinion score (MOS), which can be obtained by human ratings or objective metrics.

The goal of training is to learn a model $P_w(y|x)$. Generally speaking, d has a large number of distortion categories, and the processing function f makes the distorted picture produce different image features. One might naturally question potential biases in the accuracy of a single network across different distortion categories. We use Pearman's Rank Order Correlation Coefficient (SROCC) as a measure of distortion category accuracy:

$$\text{SROCC}_w^d = \text{SROCC}(\{g_w(f_d(I, s_i)), y_{(d,s_i)}\}_{i=1}^{n_d}), \tag{2}$$

where g_w is the network with parameter w, s_i is the degree of distortion of the d-type distortion category, and n_d is the number of d-type distortion categories. We can observe the result shown in Fig. 1, by minimizing the negative log-likelihood loss function, which is as follows:

$$\mathcal{L}_{nll} = \frac{1}{n}\sum_{i=1}^{n} -\log \mathcal{P}_w\left(y_i \mid x_i\right), \tag{3}$$

where the probability distribution of the mapping function is denoted as $P_w(y|x)$. We can obtain two distributions of the model predictions, denoted as $P_{w1}(y_i|x_i)$ and $P_{w2}(y_i|x_i)$. As depicted in Fig. 1, when minimizing the negative log-likelihood loss function, significant patterns emerge that elucidate our approach's effectiveness, models $P_1^w(y_i|x_i)$ trained with large learning rates and models $P_2^w(y_i|x_i)$ trained with small learning rates have significant performance differences on certain distortion types, even though they are both trained on the same data $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$. To be precise, there exists distortion d on each epoch with a high probability such that

$$\text{SROCC}_{w1}^d > \text{SROCC}_{w2}^d. \tag{4}$$

Based on above observations, we propose our Multi-Rate Knowledge Distillation (MRKD) to regularize the output predictions of sub-models. At each iteration of the training step, our method attempts to regularize the model predictions by minimizing the mean squared error (MSE):

$$D(P_{w1}(y_i|x_i), P_{w2}(y_i|x_i)) = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^2)^2, \tag{5}$$

where y_i^1 and y_i^2 are the outputs of the two models respectively. Since it is difficult to ensure the convergence of the network when training with a large learning rate, we adopt

Table 1: Model performance improvements with MRKD across different models such as DBCNN and ResNet50 on TID2013 and KADID-10K datasets.

Method	TID2013		KADID-10k	
	SRCC	PLCC	SRCC	PLCC
DBCNN	0.8043	0.8303	0.8148	0.8077
w/ URKD	0.8074	0.8331	0.8152	0.8085
w/ MRKD	**0.8397**	**0.8633**	**0.8258**	**0.8231**
	+0.0354	**+0.0332**	**+0.0110**	**+0.0154**
ResNet50	0.7278	0.7725	0.7117	0.7139
w/ URKD	0.7284	0.7726	0.7188	0.7218
w/ MRKD	**0.7700**	**0.8063**	**0.7447**	**0.7530**
	+0.0422	**+0.0338**	**+0.0330**	**+0.0391**
HyperIQA	0.7764	0.8081	0.7818	0.7595
w/ URKD	0.7608	0.8062	0.7862	0.7626
w/ MRKD	**0.7974**	**0.8296**	**0.8011**	**0.7875**
	+0.0210	**+0.0215**	**+0.0193**	**+0.0280**

Table 2: Model performance improvements with MRKD of DBCNN on the two hardest distortions (eg. Non eccentricity pattern noise and Mean shift) and two easy distortions (eg. Additive Gaussian and JPEG transmission errors) of the TID2013 dataset.

Hard	Non eccentricity		Mean shift	
	SRCC	PLCC	SRCC	PLCC
DBCNN	0.0969	0.0328	0.1523	0.1274
w/ URKD	0.0869	0.0385	0.1546	0.1332
w/ MKRD	**0.4277**	**0.4014**	**0.2285**	**0.1550**
	+0.3308	**+0.3686**	**+0.0762**	**+0.0276**
Easy	Additive Gaussian		JPEG transmission	
	SRCC	PLCC	SRCC	PLCC
DBCNN	0.8892	0.8806	0.9423	0.9385
w/ URKD	0.8900	0.8817	0.9423	0.9387
w/ MKRD	**0.9069**	**0.8934**	**0.9700**	**0.9582**
	+0.0177	**+0.0128**	**+0.0277**	**+0.0197**

a learning rate decay strategy to simultaneously decay the learning rates of the two networks. Specifically, cosine annealing rule [21] is used as a learning rate scheduling method that sets the learning rate for each parameter group according to the following formula:

$$\begin{cases} \eta_{w1}^{t} = \eta_{w1}^{min} + \frac{1}{2}(\eta_{w1}^{max} - \eta_{w1}^{min})(1 + \cos(\frac{T}{T_{w1}^{max}}\pi)) \\ \eta_{w2}^{t} = \eta_{w1}^{min} + \frac{1}{2}(\eta_{w2}^{max} - \eta_{w1}^{min})(1 + \cos(\frac{T}{T_{w2}^{max}}\pi)) \end{cases}, \tag{6}$$

where η^{max} is the initial learning rate, η^{min} is the minimum learning rate and T is the number of iterations since the last restart.

3. EXPERIMENTS

3.1. Experimental Settings

Datasets and Evaluation Metrics. We evaluated various NR-IQA models and MRKD on two commonly used IQA datasets, TID2013, KADID-10k and LIVEC datasets. The

Table 3: Model performance improvements with MRKD of DBCNN on the two hardest distortions (eg. Non-eccentricity patch and Color saturation 1) and two easy distortions (eg. Gaussian blur and Lens blur) of the KADID-10k dataset.

Hard	Non eccentricity		Color saturation	
	SRCC	PLCC	SRCC	PLCC
DBCNN	0.2430	0.2082	0.3333	0.3505
w/ URKD	0.2456	0.2088	0.3400	0.3581
w/ MKRD	**0.3024**	**0.2796**	**0.4208**	**0.4661**
	+0.0594	**+0.0714**	**+0.0875**	**+0.1156**
Easy	Gaussian blur		Lens blur	
	SRCC	PLCC	SRCC	PLCC
DBCNN	0.8710	0.8795	0.7697	0.8069
w/ URKD	0.8726	0.8808	0.7736	0.8093
w/ MKRD	**0.8830**	**0.8905**	**0.8776**	**0.8505**
	+0.0120	**+0.0110**	**+0.1079**	**+0.0436**

Table 4: Model performance improvements with MRKD of ResNet50 on the two hardest distortions (eg. Non eccentricity pattern noise and Mean shift) and two easy distortions (eg. Gaussian blur and JPEG2000 compression) of the TID2013 dataset.

Hard	Non eccentricity		Mean shift	
	SRCC	PLCC	SRCC	PLCC
ResNet50	0.3838	0.4561	0.1008	0.0495
w/ URKD	0.3646	0.4583	0.1100	0.0625
w/ MKRD	**0.4262**	**0.4603**	**0.4000**	**0.3245**
	+0.0424	**+0.0042**	**+0.2992**	**+0.2750**
Easy	Gaussian blur		JPEG2000 comp	
	SRCC	PLCC	SRCC	PLCC
ResNet50	0.8585	0.8588	0.9346	0.9335
w/ URKD	0.8608	0.8595	0.9285	0.9320
w/ MKRD	**0.8946**	**0.8817**	**0.9485**	**0.9450**
	+0.0361	**+0.0229**	**+0.0139**	**+0.0115**

performance of the NR-IQA models was assessed using two standard evaluation metrics: Spearman Rank Order Correlation Coefficient (SRCC) [22] and Pearson Linear Correlation Coefficient (PLCC) [23]. The datasets were randomly divided, allocating 80% of the images for training and 20% for testing. We selected the results from the final epoch after the model had converged during training as the definitive outcomes. Additionally, we conducted multiple experiments on the models and reported the average results for consistency.

Models and Implementation Details. To evaluate the proposed multi-rate knowledge distillation approach, we applied the MRKD method on three representative NR-IQA models, including DBCNN, HyperIQA, and ResNet50, thereby assessing the effectiveness of the MRKD. For the model training configuration, the batch size was set to 16, with a cosine decay schedule, and optimization was carried out using SGD. The original learning rate (small lr) for DBCNN was set at 1e-2, while the large learning rate teacher model was designated at 7e-1. The learning rate for ResNet50 and HyperIQA was set at 5e-2 and the large lr was set at 2e-1 and 1e-1. The distillation loss coefficient between the large and small learning rate models, relative to their MSE loss, is 1.

Table 5: Performances of ORI, URKD, and MRKD models on the LIVEC dataset.

Method	ORI		URKD		MRKD	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
ResNet-50	0.7814	0.8215	0.7834	0.8235	**0.8312**	**0.8552**
DBCNN	0.7812	0.8296	0.7780	0.8325	**0.8199**	**0.8491**
HyperIQA	0.7875	0.8015	0.7813	0.8075	**0.8297**	**0.8671**

Table 6: Results of cross dataset evaluations trained on different fractions of datasets adopting the state-of-the-art model DBCNN.

Training	Testing	Method	SRCC	PLCC
TID2013	KADID-10k	DBCNN	0.5347	0.5678
		w/ MRKD	**0.6013**	**0.6140**
			+0.0666	**+0.0462**
KADID-10k	TID2013	DBCNN	0.6240	0.6423
		w/ MRKD	**0.6433**	**0.6670**
			+0.0193	**+0.0247**

3.2. Enhancing IQA Models with Multi-Rate Knowledge Distillation

In the NR-IQA task, we trained three representative models, namely DBCNN, HyperIQA, and ResNet50, to investigate whether MRKD can enhance the performance of models on two IQA benchmark datasets (TID2013 and KADID-10k), especially towards challenging distortion types.

As shown in Table 1, the MRKD method led to a significant improvement over the baseline model performances. Specifically, for the overall SRCC and PLCC, MRKD enabled DBCNN to achieve a performance gain of +0.0354 in SRCC and +0.0332 in PLCC on the TID2013 dataset, and a gain of +0.011 in SRCC and +0.0154 in PLCC on the KADID-10K dataset. The comparative method URKD (Uniform Rate Knowledge Distillation), which maintains consistent learning rates for both models, indicates that the performance gain introduced by our proposed approach does not merely arise from direct consistent knowledge distillation between two models, but rather necessitates multi-rate knowledge distillation.

Furthermore, to highlight the capability of addressing more challenging distortion types, Tables 2, 3 and 4 display our evaluations on the TID2013 and KADID-10k datasets. We selected two distortion types with the poorest performance and two with relatively better performance to assess the impact of MRKD on model performance. Experimental results demonstrated that MRKD effectively enhances the model's capability towards challenging samples. For instance, in the Non-eccentricity type, DBCNN achieved a performance gain of +0.3308 in SRCC. For the Mean shift type, ResNet50 attained a performance gain of +0.2992 in SRCC. Additionally, for distortion types where the performance was already commendable, MRKD still rendered significant performance improvements.

We extended our experiments to the LIVEC dataset. Unlike TID2013 and KADID-10k, LIVEC doesn't categorize specific distortion types, thus detailed performance metrics for each distortion aren't provided. However, our comprehensive results on the LIVEC dataset are presented in Table 5, underscoring the effectiveness of our proposed method in real-world scenarios.

Table 7: Ablation studies on MRKD models with different learning rates on the DBCNN on TID2013 dataset.

Model (Learning Rate)	Large LR Model		Small LR Model	
	SRCC	PLCC	SRCC	PLCC
DBCNN-URKD (1e-2)	0.8074	0.8331	0.8074	0.8331
w/ MRKD (3e-2)	0.8292	0.8558	0.8119	0.8415
w/ MRKD (5e-2)	0.8159	0.8394	0.8239	0.8519
w/ MRKD (7e-2)	0.7942	0.8196	0.8241	0.8528
w/ MRKD (1e-1)	0.7711	0.8022	0.8265	0.8547
w/ MRKD (3e-1)	0.7070	0.7549	0.8299	0.8589
w/ MRKD (5e-1)	0.7410	0.7904	0.8309	0.8585
w/ MRKD (7e-1)	0.7728	0.8166	**0.8397**	**0.8633**

3.3. Multi-Rate Knowledge Distillation improves generalization ability

To demonstrate the generalization capability of the proposed MRKD method, we conducted a cross-dataset evaluation. We trained the DBCNN on the TID2013 dataset and tested it on the KADID-10K dataset. Table 6 attests to the effectiveness of MRKD in enhancing the generalization performance of NR-IQA models.

3.4. Ablation Study

To examine the proposed MRKD method's impact, we conducted ablation studies on the adoption of models with varying learning rates and the large learning rate teacher model. The results, presented in Table 7, reveal that models with a smaller learning rate outperform those with a larger one. This supports our hypothesis that smaller learning rate models learn more robustly, while larger ones guide knowledge transfer for challenging distortions and assist in avoiding local optima. Additionally, Table 7 indicates that 7e-1 is the optimal parameter for the DBCNN model on the TID2013.

4. CONCLUSION

Our research highlights that standard IQA models often struggle with certain challenging distortions when using a consistent learning rate. Addressing this gap, we introduced the Multi-Rate Knowledge Distillation (MRKD) strategy, employing dual models with varying learning rates. This method promotes a thorough understanding of diverse distortion features, and tests on TID and KADID datasets have showcased its superiority over conventional models. Our work provides a modest contribution to the community by suggesting a nuanced model training strategy. We anticipate that future refinements of MRKD could further enhance IQA systems.

5. REFERENCES

[1] Guangtao Zhai and Xiongkuo Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, no. 11, pp. 1–52, 2020.

[2] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2017.

[3] Jupo Ma, Jinjian Wu, Leida Li, Weisheng Dong, Xuemei Xie, Guangming Shi, and Weisi Lin, "Blind image quality assessment with active inference," *IEEE Transactions on Image Processing*, vol. 30, pp. 3650–3663, 2021.

[4] Feng Shao, Weisi Lin, Shanbo Gu, Gangyi Jiang, and Thambipillai Srikanthan, "Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1940–1953, 2013.

[5] Xinfeng Zhang, Weisi Lin, and Qingming Huang, "Fine-grained image quality assessment: A revisit and further thinking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2746–2759, 2021.

[6] Long Xu, Jia Li, Weisi Lin, Yongbing Zhang, Lin Ma, Yuming Fang, and Yihua Yan, "Multi-task rank learning for image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 9, pp. 1833–1843, 2016.

[7] Heliang Zheng, Huan Yang, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo, "Learning conditional knowledge distillation for degraded-reference image quality assessment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10242–10251.

[8] Guanghao Yin, Wei Wang, Zehuan Yuan, Chuchu Han, Wei Ji, Shouqian Sun, and Changhu Wang, "Content-variant reference image quality assessment via knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 3134–3142.

[9] Haiyu Zhang, Shaolin Su, Yu Zhu, Jinqiu Sun, and Yanning Zhang, "Boosting no-reference super-resolution image quality assessment with knowledge distillation and extension," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.

[11] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin, "Regularizing class-wise predictions via self-knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13876–13885.

[12] Zeyuan Allen-Zhu and Yuanzhi Li, "Towards understanding ensemble, knowledge distillation and self-distillation in deep learning," in *The Eleventh International Conference on Learning Representations*, 2022.

[13] Samy Jelassi and Yuanzhi Li, "Towards understanding how momentum improves generalization in deep learning," in *Proceedings of the 39th International Conference on Machine Learning*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, Eds. 17–23 Jul 2022, vol. 162 of *Proceedings of Machine Learning Research*, pp. 9965–10040, PMLR.

[14] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari, "The large learning rate phase of deep learning: the catapult mechanism," *arXiv preprint arXiv:2003.02218*, 2020.

[15] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al., "Image database tid2013: Peculiarities, results and perspectives," *Signal processing: Image communication*, vol. 30, pp. 57–77, 2015.

[16] Hanhe Lin, Vlad Hosu, and Dietmar Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–3.

[17] Deepti Ghadiyaram and Alan C Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2015.

[18] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.

[19] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667–3676.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] Ilya Loshchilov and Frank Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[22] Jerrold H Zar, "Spearman rank correlation," *Encyclopedia of biostatistics*, vol. 7, 2005.

[23] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*, pp. 1–4. Springer, 2009.