Conformer and Blind Noisy Students for Improved Image Quality Assessment

Marcos V. Conde, Maxime Burchi, Radu Timofte
Computer Vision Lab, Institute of Computer Science, University of Würzburg, Germany
{marcos.conde-osorio,maxime.burchi,radu.timofte}@uni-wuerzburg.de

Abstract

Generative models for image restoration, enhancement, and generation have significantly improved the quality of the generated images. Surprisingly, these models produce more pleasant images to the human eye than other methods, yet, they may get a lower perceptual quality score using traditional perceptual quality metrics such as PSNR or SSIM. Therefore, it is necessary to develop a quantitative metric to reflect the performance of new algorithms, which should be well-aligned with the person's mean opinion score (MOS).

Learning-based approaches for perceptual image quality assessment (IQA) usually require both the distorted and reference image for measuring the perceptual quality accurately. However, commonly only the distorted or generated image is available. In this work, we explore the performance of transformer-based full-reference IQA models. We also propose a method for IQA based on semi-supervised knowledge distillation from full-reference teacher models into blind student models using noisy pseudo-labeled data.

Our approaches achieved competitive results on the NTIRE 2022 Perceptual Image Quality Assessment Challenge: our full-reference model was ranked 4th, and our blind noisy student was ranked 3rd among 70 participants, each in their respective track. https://github.com/burchim/IQA-Conformer-BNS.

1. Introduction

Image quality assessment (IQA) aims at using computational models to measure the perceptual quality of images, which are degraded during acquisition, generation, compression or post-processing operations [47, 65]. Since one of the goals of the image processing is to improve the quality of the content to an acceptable level for the human viewers, IQA, as a "evaluation technique", plays a critical role in most image processing tasks such as image super-resolution, denoising, compression and enhancement [3, 4, 19, 21, 64]. Although it is easy for human beings to distinguish perceptually better images, it has been proved to be difficult for algorithms [19, 42–44].

Recently, Generative Models [15, 17, 37] have shown promising results for image enhancement and generation, producing realistic results to the human eye. For instance, perceptual image processing algorithms based on Generative Adversarial Networks (GANs) [8, 17, 32, 57] have produced images with more realistic textures.

However, these generated images show completely different characteristics and artifacts from traditional distortions (i.e. Gaussian Noise, Blur), for this reason, it has been noticed that the contradiction between the quantitative evaluation results and the real perceptual quality is increasing [4, 5, 19]. Therefore, these methods have posed a great challenge for IQA methods to evaluate their visual quality. New IQA methods need to be proposed accordingly to evaluate new image processing algorithms, as this will also affect the development of such methods [4, 5, 18, 19, 21].

In this context, in order to generate acceptable images we have to accurately measure their perceptual quality, which can be performed via subjective and objective quality assessment [9, 16, 29, 58]. The subjective quality assessment is the most accurate method to measure the perceived quality, which is usually represented by mean opinion scores (MOS) from collected human subjective ratings. However, it is time-consuming and expensive.

Deep Convolutional Neural Networks (CNNs) can extract complex features from the images, and thus, they can provide a powerful IQA metric if there is enough data to train them. Moreover, these represent differentiable functions, allowing to plug them into adversarial training frameworks and optimize for quality directly [6, 13, 19, 44, 70].

In general, we find two different IQA approaches: (i) *Full-Reference (FR)* [1, 10, 21, 43, 47] where an image without distortions is available besides the distorted image. (ii) *No-Reference (NR)* (also known as Blind) [6, 6, 27, 38, 39] where only the distorted or generated image is available. Typically, Full-Reference approaches achieve better performance, however, Blind IQA (BIQA) represents the most realistic scenario and these approaches are more useful because of their feasibility. In Section 2 we present the *state-of-the-art* of each case.

Reference | (a) PSNR 29.22 SSIM 0.37 MOS 1222.33 | (b) PSNR 34.65 SSIM 0.82 MOS 1318.83 | (c) PSNR 29.43 SSIM 0.62 MOS 1390.17 | (d) PSNR 31.66 SSIM 0.90 MOS 1608.99

Figure 1. Training samples from the PIPAL [19, 21]. As we can see, ranking the images by their perceptual quality depends on the metric, and there great discrepancies [4, 5, 18]. IQA models must learn to predict quantitative outputs as much correlated as possible with the MOS human ratings. We appreciate a huge perceptual quality difference between (b) and (d), however, neither PSNR nor SSIM reflect this.

The **NTIRE 2022 Perceptual Image Quality Assessment Challenge** [20] seeks for novel solutions for Full-Reference and No-Reference IQA. In comparison with previous IQA benchmarks [42, 44], the training and testing datasets in this challenge include the outputs of GAN-based algorithms and the corresponding subjective scores, which provide more diversity and challenging scenarios. In this work we provide the following key contributions:

- In Section 3 we introduce our conformer-based 4th place solution for Full-Reference IQA, as an alternative to transformer-based approaches like IQT [10] (winner of last year challenge).
- In Section 4 we present our 3rd place solution for No-Reference IQA: Exploration of semi-supervised noisy student learning to distill knowledge from FR models into blind noisy student models.
- Comparison with the NTIRE 2021 IQA Challenge [21] methods and extensive ablation studies.

2. Related Work

Image Quality Assessment. CNNs have shown their effectiveness in a wide range of computer vision and image processing tasks, such as super-resolution, denoising and deblurring [41, 54, 64]. Generative models [15, 17, 37], and in particular, GAN-based [17, 32] approaches produce typically more pleasant results to human eyes than the CNNs that do not use adversarial loss. The goal of the developing IQA methods is to accurately predict the perceived quality (by human viewers) of the generated images. However, traditional IQA methods struggle to evaluate these new approaches, and there are contradictions between the perceptual quantitative results and the qualitative results. We can classify IQA methods depending on:

- Input data: (i) Full-Reference (FR) [1, 10, 21, 43, 47] where a reference image without distortions is available besides the distorted image. (ii) No-Reference (NR) [6, 6, 27, 38, 39] where only the distorted or generated image is available.
- Training: (i) Traditional methods do not require training. (ii) Learnable methods (typically CNN-based).

Full-Reference IQA The FR methods focus more on visual similarity or dissimilarity between two images (typically the original or reference image, and the generated one). The most representative IQA FR metrics are the PSNR, which is related to the MSE between both images, and the SSIM proposed by Wang *et al.* [60]. These traditional methods have the advantage of convenience for optimization; however, they poorly predict humans perceived visual quality, especially for evaluating fine textures and details in the images [33]. Since that, various FR metrics have been developed to take into account various aspects of human quality perception, e.g., information-theoretic criterion [46] or structural similarity [61, 68].

Note that the ultimate goal of image enhancement networks is to generate visually pleasant images for humans and have a high MOS, which is not always strictly correlated to these traditional metrics. Recently, learned CNN-based IQA methods have been actively studied and provide the most promising *state-of-the-art* results [6, 14, 21, 27, 44, 70]. Zhang *et al.* proposed a learned perceptual image patch similarity (LPIPS) metric [70], which shows that trained deep features that are optimized by the l_2 distance between distorted and reference images are effective for IQA compared to the conventional IQA methods.

Among the most competitive approaches in the NTIRE 2021 IQA Challenge [21] we can find: ASNA [45] proposed a CNN equipped with spatial and channel-wise attention mechanisms, and Siamese-like network architecture. IQMA [23] proposed a bilateral-branch multi-scale image

quality estimation network, using Feature Pyramid Network (FPN)-like architecture to extract multi-scale features and predict the the quality score of the image at multiple scales.

Cheon *et al.* introduced an image quality transformer IQT [10] that successfully applies a transformer architecture to a perceptual full-reference IQA task. This method combines a CNN backbone as a feature extractor, with a Transformer [56] encoder-decoder to compare a reference and distorted images, and predict the quality score.

Blind IQA The No-Reference (NR) or Blind methods [28, 34] are useful because of its feasibility, they can be plugged-in in adversarial training frameworks and be used for optimizing perceptual quality directly. However, the absence of a reference image makes it challenging to predict image quality accurately compared to the FR methods.

Bosse *et al.* [6] studies the performance of deep neural networks for no-reference and full-reference image quality assessment. Mittal *et al.* [39] explores blind IQA in the spatial domain. Zhang *et al.* [71] proposes a model and a training approach to deal with realistic and synthetic distortions and improve the generalization capabilities.

The NTIRE 2022 IQA Challenge introduced this year a track for Blind image quality assessment (BIQA).

Evaluation IQA methods should present the following two desired characteristics: (i) high Pearson linear correlation coefficient (PLCC) between the scores produced by the proposed method and the ground-truth MOS, which indicates the linear relationship between them, (ii) high Spearman rank order correlation coefficient (SRCC), which shows the monotonicity of relationship between the proposed method and the ground-truth MOS. Both metrics separately, and the sum of both as a "Main Score", serve as evaluation metric to compare the performance of IQA methods [19, 21, 47]. The Kendall Rank-order Correlation Coefficient (KRCC) is also used to estimate the monotonicity and consistency of the quality prediction [18].

Datasets TID2013 [42], LIVE [47] and PIPAL [18, 19] provide images with their corresponding reference images and MOS to train models in a supervised manner. We compare these datasets in Table 1. The NTIRE 2022 IQA Challenge [20] uses the PIPAL [19] dataset, which takes a step forward in benchmarking perceptual IQA by incorporating the perceptual quality of images obtained by perceptual-oriented algorithms (i.e. GANs), missing in previous datasets. The PIPAL [19] as our training set, contains 200 reference images, 23k distorted images and their respective human judgements. To ensure that the models can generalize properly, the challenge has an extended dataset of PIPAL for validation and testing. This dataset contains 3300 distorted images (1650 for training and testing respectively) for 50 reference images, and all of them are the outputs of perceptual-oriented algorithms. It collects 753k human judgements to assign subjective scores for the extended images, ensuring the objectivity of the testing data. Participants do not have access to the ground-truth for validation or test, results are submitted using a public website.

Database	# Ref.	# Dist.	Dist. Type	# Dist. Type	# Rating
LIVE [47]	29	779	trad.	5	25k
TID2013 [42]	25	3k	trad.	25	524k
PIPAL [19]	250	29k	trad.+alg.	40	1.13m

Table 1. Comparison of IQA datasets for performance evaluation. The NTIRE Challenge Dataset, PIPAL [19] presents the highest and more various number of distorsions and human ratings.

3. IQA Conformer Network

We propose an alternative architecture to IQT [10] by replacing the Transformer encoder-decoder [56] by a Conformer architecture [7, 22], which uses convolution and attention operations to model local and global dependencies.

We use a Inception-ResNet-v2 [52] network pre-trained on ImageNet to extract feature maps from the reference and distorted image. The network weights are kept frozen and a Conformer [22] encoder-decoder is trained to regress MOS using the MSE loss. As done by Cheon *et al.* [10], we concatenate the feature maps from the following blocks: `mixed5b, block35_2, block35_4, block35_6, block35_8 and block35_10`. We do this for the reference and distorted images generating f_{ref} and f_{dist}, respectively. In order to obtain difference information between reference and distorted images, a difference feature map, $f_{diff} = f_{ref} - f_{dist}$ is also used.

Concatenated feature maps are then projected using a point-wise convolution but not flattened to preserve spatial information. We used a single Conformer block [22] for both encoder and decoder. The model hyper-parameters are set as follows: $L = 1$, $D = 128$, $H = 4$, $D_{feat} = 512$, and $D_{head} = 128$. The input image size of the backbone model is set to (192 × 192 × 3) which generates feature maps of size 21 x 21. **IQA Conformer** has 2,831,841 total parameters, we illustrate the pipeline in Figure 2. Note that we only use the first feature maps from the CNN, not the whole network, therefore the number of parameters is substantially smaller. In Table 4 we compare our method with the *state-of-the-art* on the NTIRE 2021 and 2022 IQA Challenges [20, 21]. For a fair comparison, all the models were trained using the same PIPAL training dataset [19, 21]. We use RADN [48] and ASNA [2] public available pre-trained weights and code. Our proposed solution allows to reach better PLCC and SRCC at inference than IQT [10] under the same setup (see Table 4). Note that to the best of our

knowledge, there is no public code or models for reproducing IQT [10] results, therefore we report results of our best implementation following the original paper. We also compare our results with other top performing teams at the NTIRE 2022 IQA FR Challenge [20] in Table 2, where our IQA Conformer was ranked 4th. We show qualitative samples and analysis in Figures 3 and 8.

Team	Main Score ↑	PLCC	SRCC
THU1919Group	1.651	0.828	0.822
Netease OPDAI	1.642	0.827	0.815
KS	1.640	0.823	0.817
Ours	1.541	0.775	0.766
Yahaha!	1.538	0.772	0.765
debut_kele	1.501	0.763	0.737
Pico Zen	1.450	0.738	0.713
Team Horizon	1.403	0.703	0.701

Table 2. Performance comparison of the top teams on the testing dataset of the NTIRE 2022 Full-Reference IQA Challenge.

Implementation details The model was trained using only the NTIRE 2022 PIPAL training dataset [19]. Adam optimizer by setting $\beta_1 = 0.9$, $\beta_2 = 0.999$. We set mini-batch size as 16. The learning rate was set to 10^{-4} and the model trained for 30 epochs (43479 gradient steps). Last 10 epoch checkpoints were averaged using SWA [30].

Inference During inference, we use *enhanced prediction* [55] (a.k.a Test-Time Augmentations). The prediction for an input image is enhanced by averaging the predictions on a set of transformed images derived from it. We use 10 crops (2 flips of 4 image corners + center crop) for reference and distorted images.

Ensembles and fusion strategies As shown in Table 4 an ensemble of our model, RADN [48] and ASNA [2], improves notably the performance (+0.4 boost in main score).

3.1. Cross Database Evaluations

IQA methods tend to overfitting, they commonly struggle to generalize to data distributions different from the one they were trained with. To validate the generalization capabilities of our approach, we use our FR IQA Conformer trained on PIPAL [19] and we conduct the cross-dataset evaluation on two other benchmarks: TID2013 [42] and LIVE [47] (using the full datasets). As shown in Table 3, our model generalizes better than NTIRE 2021 [21] top methods: ASNA [45], RADN [49] and IQT [10]. It also achieves competitive results in comparison with other learnt methods like PieAPP [44], WaDIQaM [6] or LPIPS [70],

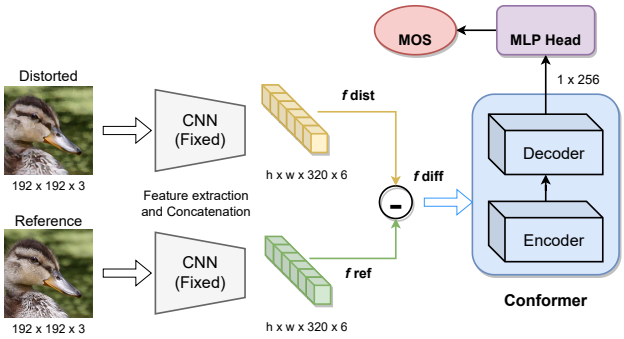


Figure 2. FR IQA Conformer setup inspired by IQT [10].

which are trained on the specific datasets. Figure 6 shows a qualitative analysis of the predictions for LIVE [47].

Method	LIVE [47]		TID2013 [42]	
	SRCC	KRCC	SRCC	KRCC
PSNR [26]	0.873	0.680	0.687	0.496
SSIM [60]	0.948	0.796	0.727	0.545
MS-SSIM [61]	0.951	0.805	0.786	0.605
VIF [47]	0.964	0.828	0.677	0.518
NLPD [31]	0.937	0.778	0.800	0.625
GMSD [63]	0.960	0.827	0.804	0.634
WaDIQaM [6]	0.947	0.791	0.831	0.631
PieAPP [44]	0.919	0.750	0.876	0.683
LPIPS [70]	0.932	0.765	0.670	0.497
DISTS [14]	0.954	0.811	0.830	0.639
SWDN [18]	-	-	0.819	0.634
ASNA [45]	0.92	-	0.73	-
RADN [49]	0.905	-	0.747	-
IQT-C [10]	0.917	0.737	0.804	0.607
Ours	0.921	0.752	0.82	0.630

Table 3. Performance comparison on LIVE [47] and TID2013 [42]. Some results are borrowed from [13, 18]. We separate traditional and learnt methods, and we highlight in blue the NTIRE 2021 [21] methods trained on PIPAL [19].

3.2. Ablation Study

In Table 4 we compare the performance of Transformer [10, 56] and Conformer [22] models, both using the same backbone (Inception-ResNet-v2 [52]) and training setup. We also explore the effect of different backbone architectures for feature extraction like ConvNext [36] (SOTA in image classification) and VGG [50] (common backbone in IQA). We find Inception-ResNet-v2 [52] features to be the best representation.These approaches are very sensitive to the backbone selection, and more specifically, to the feature block selection (as introduced by EGB [24]).

Method	Validation 2021		Testing 2021		Validation 2022		Testing 2022	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
PSNR [26]	0.292	0.255	0.277	0.249	0.269	0.234	0.277	0.249
SSIM [60]	0.398	0.340	0.394	0.361	0.377	0.319	0.391	0.361
VSI [67]	0.516	0.450	0.517	0.458	0.493	0.411	0.517	0.458
NQM [11]	0.416	0.346	0.395	0.364	0.364	0.302	0.395	0.364
UQI [59]	0.548	0.486	0.450	0.420	0.505	0.461	0.450	0.420
GSM [35]	0.469	0.418	0.465	0.409	0.450	0.379	0.465	0.409
RFSIM [69]	0.304	0.266	0.328	0.304	0.285	0.254	0.328	0.304
SRSIM [66]	0.654	0.566	0.636	0.573	0.626	0.529	0.636	0.573
LPIPS-VGG [70]	0.647	0.591	0.633	0.595	0.611	0.551	0.633	0.595
DISTS [14]	0.686	0.674	0.687	0.655	0.634	0.608	0.687	0.655
EGB [24]	0.775	0.776	0.677	0.700	0.746	0.723	0.700	0.677
ASNA [2]	0.820	0.830	0.750	0.710	0.796	0.765	0.752	0.719
RADN [48]	0.866	0.865	0.771	0.777	0.789	0.777	0.753	0.757
IQT (2021 Winner) [10]	0.876	0.865	0.790	0.799	0.840	0.820	**0.799**	**0.790**
Ours IQA Transformer					0.790	0.765	0.757	0.751
Ours IQA Conformer A					0.804	0.790	**0.775**	**0.766**
Ours IQA Conformer B					0.740	0.740	0.730	0.730
Ours IQA Conformer C					0.790	0.770	0.754	0.740
Ensemble							**0.787**	**0.793**

Table 4. Performance comparison of IQA methods on the PIPAL NTIRE 2021 and 2022 Full-Reference benchmark [19–21]. We highlight in blue the top performing methods on the NTIRE 2021 IQ Challenge [21]. The different IQA Conformer versions correspond to different backbones: (A) Inception-ResNet-v2 [52], (B) ConvNext [36], (C) VGG19 [50]. The ensemble method is: Ours + RADN [48] + ASNA [2]. Ours IQA "Transformer" is our own implementation of IQT [10], since, to the best of our knowledge, there is not public code available.

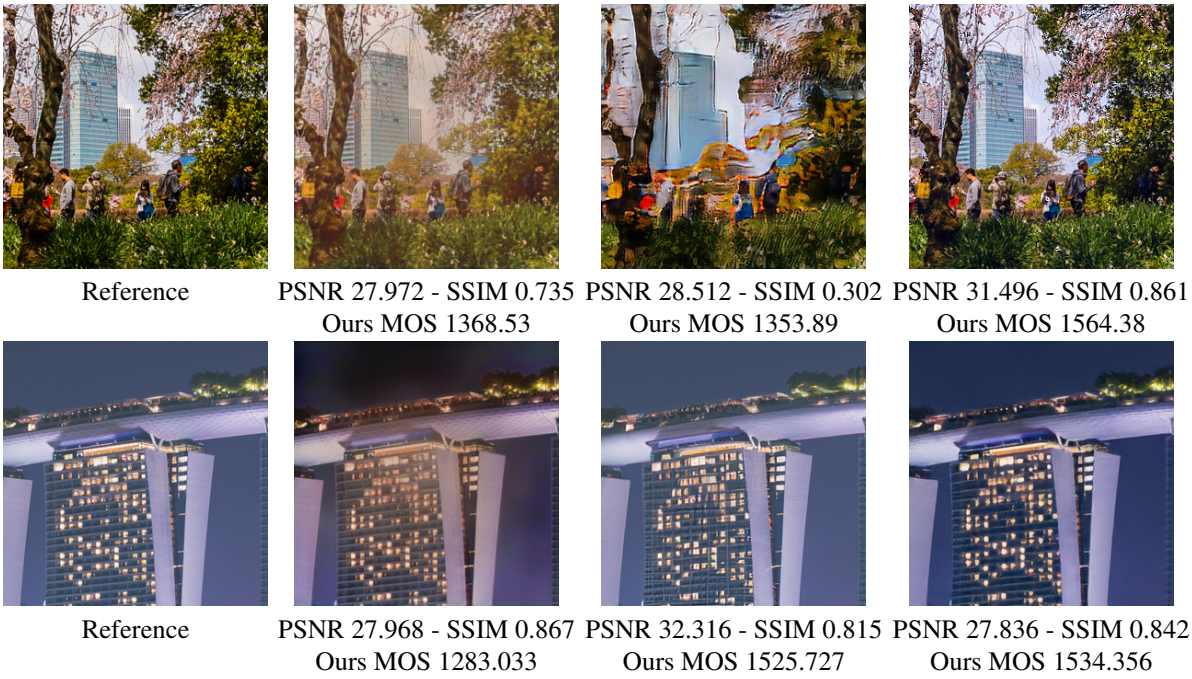


Figure 3. Example images from the test set of the NTIRE 2022 challenge. For each distorted image we provide predicted scores of PSNR, SSIM and MOS from our model. As we can see, ranking the images by their quality depends on the metric, and there great discrepancies [4, 5, 18, 19]. However, our model is the most correlated quantitative metric to the real human MOS ratings.

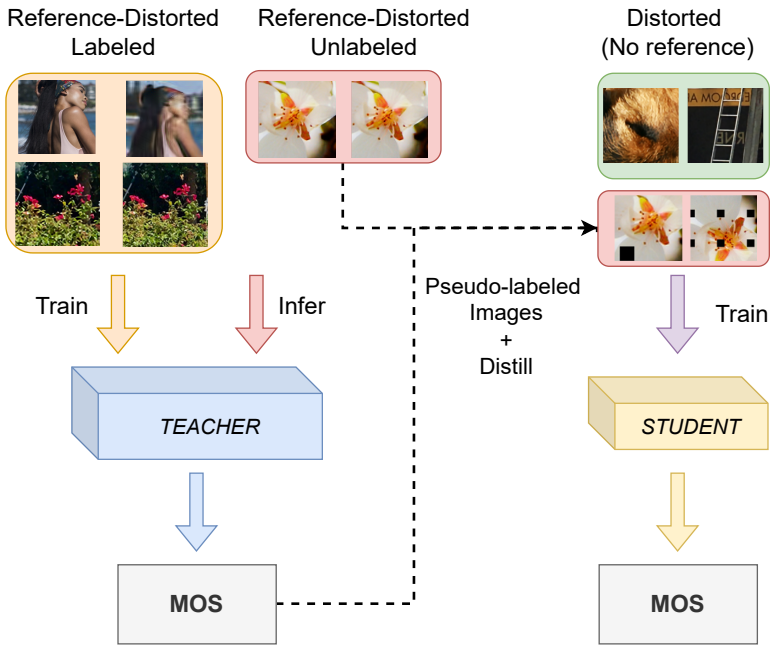


Figure 4. Full-Reference Teacher and Blind Noisy Student. Unlabeled samples are annotated using pseudo-labels inferred using the teacher.

4. Blind Noisy Student

A simple CNN backbone Φ takes as input a distorted image x and aims to minimize the MOS y (see Figure 5) using the following loss function from [2], where $\Phi(x) = \hat{y}$.

$$\mathcal{L} = MSE(y, \hat{y}) + (1 - Pearson(y, \hat{y})) \quad (1)$$

Initial setup We train EfficientNet B0 [53] (pre-trained on ImageNet) to perform this task. Using a 90/10 validation split (i.e. validating on roughly 1000 images locally), we achieved 1.02 on the development phase using a single model. We use as augmentations in all our experiments the following pipeline: random horizontal and vertical flips, random rotations of 90/180/270 degrees, and finally take a random crop of size 224 x 224. We find the main performance limitation to be overfitting due to the small dataset: 23200 images, yet only 200 reference images [19]. We select EffNet B0 [53] as backbone as it is *stat-of-the-art* in Image classification and has only 4 million parameters.

Noisy student a semi-supervised learning approach that extends the idea of self-training and distillation, and has achieved *state-of-the-art* results on Image Classification [62]. We distinguish a Full-Reference **teacher** model, and a blind **student** model trained only with distorted images. This method allows to increase the amount of training distorted images, and to transfer "dark" knowledge [62] from FR models and ensembles, into simple NR models. The process is as follows:

1. Train the FR teacher using the training dataset [19, 21] consisting on 23k reference-distorted pairs.
2. Teacher infers on **unlabeled** reference-distorted pairs, and annotate the images. These MOS annotations are noisy, we refer to them as *pseudo-labels*.
3. Add the pseudo-labeled distorted images to the original training set: approximately 2k new images.
4. Train a student model for NR IQA, which takes as input only the distorted images using the extended dataset (original + pseudo-labels) and extra augmentations to the initial setup: (i) CutOut [12] as further regularization to ensure the model learns useful features without looking to the entire image. (ii) Small perturbations on the Saturation, Brightness and Contrast. We show some examples in Figure 7.

We illustrate this process in Figure 4. Using this approach we can distill knowledge from the FR models into the NR models. We show the benefits of this approach and augmentations in Table 5. We obtain the unlabeled samples from two different ways: (i) using the unlabeled data provided at the challenge and PIPAL [19, 21], (ii) augmenting the reference images using traditional methods and GAN-based methods like SRGAN [32, 57] to upscale the images and resize back to the original resolution. We do not use other datasets for training.

Therefore, our single model EffNet B0 [53] has only 4 million parameters, in comparison with other well-known architectures used for this task such as VGG [51] (15 million parameters) or ResNet 50 [25] (24 million parameters). In our experiments, deep models tend to overfitting quickly and did not perform great. In Table 6 we show our performance in comparison with other top teams.

Method	# Extra	Augs.	Score	PLCC	SRCC
EffNet B0 [53]	No	No	0.84	0.42	0.42
EffNet B0 [53]	No	Yes	1.02	0.51	0.51
EffNet B0 [53]	1.6k	Yes	1.42	0.73	0.70
VGG 19 [51]	1.6k	Yes	1.25	0.63	0.61
ResNet50 [25]	1.6k	Yes	1.37	0.70	0.67
EffNet B0 [53]	2k	Yes	1.48	0.75	0.73
EffNet B0 [53] + TTA	2k	Yes	1.49	0.76	0.73

Table 5. Ablation study of our NR models. We indicate the number of extra pseudo-labeled samples added to the original training dataset, the use of "extra" augmentations, and the scores for each model in the NTIRE 2022 IQA Challenge test set. TTA [55] indicates test-time-augmentations (i.e. average of 4 random crops).

Team	Main Score ↑	PLCC	SRCC
THU_IIGROUP	1.444	0.740	0.704
DTIQA	1.437	0.737	0.700
Ours	1.422	0.725	0.697
KS	1.407	0.726	0.681
NetEase OPDAI	1.390	0.720	0.671
Minsu Kwon	1.183	0.607	0.576
NTU607QCO-IQA	1.112	0.585	0.527

Table 6. Performance comparison of the top teams on the testing dataset of the NTIRE 2022 No-Reference IQA Challenge Main score is calculated as the sum of PLCC and SRCC.

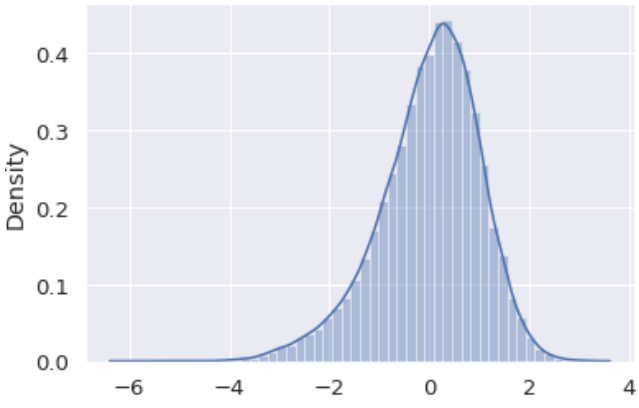


Figure 5. PIPAL [19] training MOS standardized distribution with $\mu = 1448.96$ and $\sigma = 121.53$.

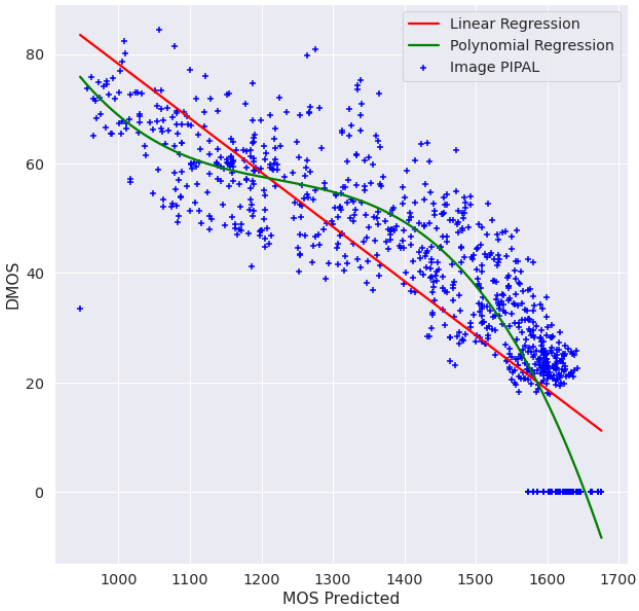


Figure 6. Ground-truth DMOS LIVE [47] against the predicted MOS. Our predictions have $|SRCC| = 0.92$, which indicates they are very correlated with the real qualitative ratings.

Implementation details We train each NR model to convergence, approximately 20 epochs. We use Adam optimizer with default parameters and learning rate 0.0001. We set batch size to 32. The learning rate is reduced by factor 0.5 on plateaus. The loss function is presented in Equation 1. We use a Tesla P100 GPU to run all our experiments. This model allows to predict the quality of the test set (1650 images) in just 8s (143ms/step) on a single GPU.

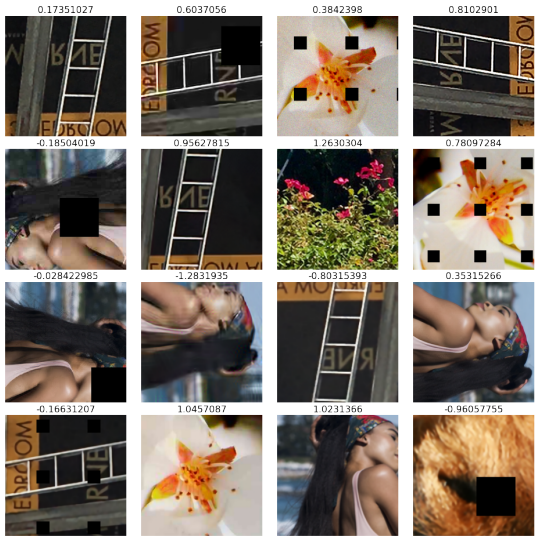
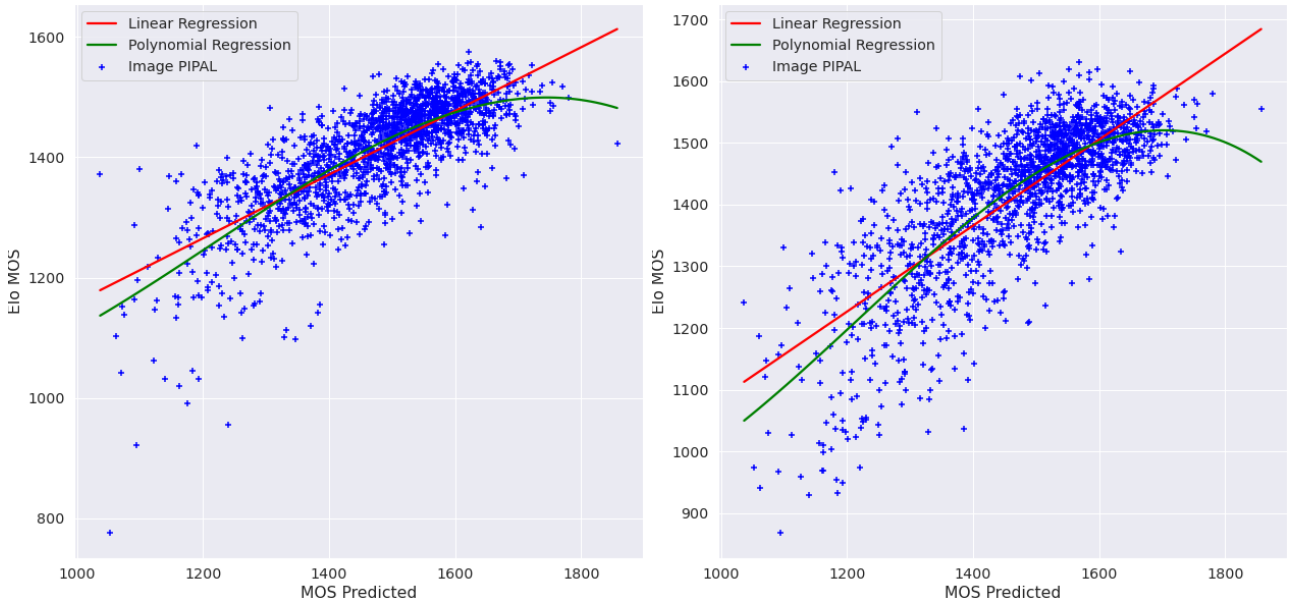


Figure 7. Example of "Extra" augmentations [12] for the noisy student training. We show standardized MOS above each image.

Method	Validation 2022			Testing 2022		
	Main Score ↑	PLCC	SRCC	Main Score ↑	PLCC	SRCC
Brisque [39]	0.075	0.059	0.015	0.184	0.097	0.087
NIQE [40]	0.120	0.115	0.005	0.142	0.112	0.030
PI [4]	0.213	0.133	0.079	0.276	0.153	0.123
Ma [38]	0.261	0.131	0.129	0.398	0.224	0.174
PSNR [26]	0.533	0.284	0.250	0.572	0.303	0.269
SSIM [60]	0.718	0.386	0.332	0.785	0.407	0.377
FSIM [68]	1.048	0.575	0.473	1.138	0.610	0.528
LPIPS-AlexNet [70]	1.197	0.616	0.581	1.176	0.592	0.584
Ours	1.410	0.710	0.700	1.490	0.752	0.733

Table 7. Performance comparison of IQA methods on the PIPAL NTIRE 2022 No-Reference benchmark [19,20]. Our method outperforms traditional and learnt methods by large margin. See also Table 6, where we compare our method with other outstanding approaches.



PIPAL [19] Full-Reference

PIPAL [19] No-Reference

Figure 8. Predicted MOS scores against Elo MOS subjective scores on the validation set of PIPAL [19] NTIRE 2022 IQA Challenge [20].

This inference time of approximately 0.22ms per image for a NR single model represents a beneficial approach for adversarial networks [17], where this model can be plugged-in as discriminator or differentiable loss function for direct perceptual quality optimization.

5. Conclusion

In this paper, we propose a method for IQA knowledge distillation from Full-Reference (FR) teacher models into Referenceless student models. First, we explore different IQA Full-Reference models, including transformer-based approaches. Next, we apply a semi-supervised noisy student approach: we annotate unlabeled reference-distorted image pairs using the FR model, we expand the original training set of distorted images using such pseudo-labeled data, and we finally train a Blind noisy student model.

Our methods achieved competitive performance on the latest PIPAL dataset, which contains new algorithm-based distorted images, and our predictions are well correlated with subjective human mean opinion scores of the images. Our methods achieved the 4th and 3rd place at the NTIRE 2022 Perceptual Image Quality Assessment Challenge for Full-reference and No-Reference respectively. Moreover, our approach can successfully generalize to other datasets like TID2013 or LIVE. As future work, we will study our performance using massively augmented datasets via semi-supervised noisy pseudo-labels.

Acknowledgments This work was supported by the Humboldt Foundation. We thank the organizers of the NTIRE 2022 Perceptual IQA Challenge for their support.

References

[1] Sewoong Ahn, Yeji Choi, and Kwangjin Yoon. Deep learning-based distortion sensitivity prediction for full-reference image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1, 2

[2] Seyed Mehdi Ayyoubzadeh and Ali Royat. (asna) an attention-based siamese-difference neural network with surrogate ranking loss function for perceptual image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 388–397, 2021. 3, 4, 5, 6

[3] Goutam Bhat, Martin Danelljan, Radu Timofte, et al. NTIRE 2021 challenge on burst super-resolution: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1

[4] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor. The 2018 PIRM challenge on perceptual image super-resolution. In *Eur. Conf. Comput. Vis. Worksh.*, pages 1–22, 2018. 1, 2, 5, 8

[5] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 1, 2, 5

[6] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017. 1, 2, 3, 4

[7] Maxime Burchi and Valentin Vielzeuf. Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition. *arXiv preprint arXiv:2109.01163*, 2021. 3

[8] M. Cheon, J.-H. Kim, J.-H. Choi, and J.-S. Lee. Generative adversarial network-based image super-resolution using perceptual content losses. In *Eur. Conf. Comput. Vis. Worksh.*, pages 1–12, 2018. 1

[9] M. Cheon and J.-S. Lee. Subjective and objective quality assessment of compressed 4k uhd videos for immersive experience. *IEEE TCSVT*, 28(7):1467–1480, 2017. 1

[10] Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee. Perceptual image quality assessment with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2021. 1, 2, 3, 4, 5

[11] Niranjan Damera-Venkata, Thomas D Kite, Wilson S Geisler, Brian L Evans, and Alan C Bovik. Image quality assessment based on a degradation model. *IEEE transactions on image processing*, 9(4):636–650, 2000. 5

[12] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 6, 7

[13] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 2020. 1, 4

[14] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 129(4):1258–1281, 2021. 2, 4, 5

[15] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 1, 2

[16] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang. Perceptual quality assessment of smartphone photography. In *CVPR*, pages 3677–3686, 2020. 1

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, page 2672–2680, 2014. 1, 2, 8

[18] J. Gu, H. Cai, H. Chen, X. Ye, J. Ren, and C. Dong. Image quality assessment for perceptual image restoration: A new dataset, benchmark and metric. *arXiv preprint arXiv:2011.15002*, 2020. 1, 2, 3, 4, 5

[19] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy S Ren, and Chao Dong. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[20] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy Ren, Radu Timofte, et al. NTIRE 2022 challenge on perceptual image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2, 3, 4, 5, 8

[21] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S. Ren, Yu Qiao, Shuhang Gu, Radu Timofte, et al. NTIRE 2021 challenge on perceptual image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1, 2, 3, 4, 5, 6

[22] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020. 3, 4

[23] Haiyang Guo, Yi Bin, Yuqing Hou, Qing Zhang, and Hengliang Luo. Iqma network: Image quality multi-scale assessment network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[24] Dounia HAMMOU, Sid Ahmed FEZZA, and Wassim Hamidouche. Egb: Image quality assessment based on ensemble of gradient boosting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 4, 5

[25] 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*, pages 770–778, 2016. 7

[26] Alain Hore and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *2010 20th International Conference on Pattern Recognition*. IEEE, Aug. 2010. 4, 5, 8

[27] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE TIP*, 29:4041–4056, 2020. 1, 2

[28] Weilong Hou, Xinbo Gao, Dacheng Tao, and Xuelong Li. Blind image quality assessment via deep learning.

IEEE transactions on neural networks and learning systems, 26(6):1275–1286, 2014. 3

[29] B. Hu, L. Li, J. Wu, and J. Qian. Subjective and objective quality assessment for image restoration: A critical survey. *Signal Processing: Image Communication*, 85:115839, 2020. 1

[30] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 4

[31] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli. Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging*, 2016(16):1–6, 2016. 4

[32] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 2, 6

[33] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network, 2016. 2

[34] Xin Li. Blind image quality assessment. In *Proceedings. International Conference on Image Processing*, volume 1, pages I–I, 2002. 3

[35] A. Liu, W. Lin, and M. Narwaria. Image quality assessment based on gradient similarity. *IEEE TIP*, 21(4):1500–1512, 2012. 5

[36] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022. 4, 5

[37] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *European conference on computer vision*, pages 715–732. Springer, 2020. 1, 2

[38] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 1, 2, 8

[39] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 1, 2, 3, 8

[40] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 8

[41] Seungjun Nah, Sanghyun Son, Suyoung Lee, Radu Timofte, Kyoung Mu Lee, et al. NTIRE 2021 challenge on image deblurring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[42] 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*, 30:57–77, 2015. 1, 2, 3, 4

[43] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. Tid2008 - a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10:30–45, 01 2009. 1, 2

[44] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *CVPR*, pages 1808–1817, 2018. 1, 2, 4

[45] Ali Royat Seyed Mehdi Ayyoubzadeh. (asna) an attention-based siamese-difference neural network with surrogate ranking loss function for perceptual image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2, 4

[46] H. R. Sheikh, A. C. Bovik, and G. De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE TIP*, 14(12):2117–2128, 2005. 2

[47] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 1, 2, 3, 4, 7

[48] Shuwei Shi, Qingyan Bai, Mingdeng Cao, Weihao Xia, Jiahao Wang, Yifan Chen, and Yujiu Yang. Region-adaptive deformable network for image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2021. 3, 4, 5

[49] Shuwei Shi, Qingyan Bai, Mingdeng Cao, Weihao Xia, Jiahao Wang, Yifan Chen, and Yujiu Yang. Region-adaptive deformable network for image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 4

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 5

[51] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[52] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 3, 4, 5

[53] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6, 7

[54] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–125, 2017. 2

[55] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1865–1873, 2016. 4, 7

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017. 3, 4

[57] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision*, pages 63–79. Springer, 2018. 1, 6

[58] Z. Wang. Applications of objective image quality assessment methods [applications corner]. *IEEE Signal Processing Magazine*, 28(6):137–142, 2011. 1

[59] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, 2002. 5

[60] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 2, 4, 5, 8

[61] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *Proc. IEEE Asilomar Conf. Signals, Systems and Computers*, volume 2, pages 1398–1402, 2003. 2, 4

[62] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 6

[63] W. Xue, L. Zhang, X. Mou, and A. C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE TIP*, 23(2):684–695, 2014. 4

[64] Ren Yang, Radu Timofte, et al. NTIRE 2021 challenge on quality enhancement of compressed video: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1, 2

[65] G. Zhai and X. Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63:1–52, 2020. 1

[66] L. Zhang and H. Li. SR-SIM: A fast and high performance iqa index based on spectral residual. In *ICIP*, pages 1473–1476, 2012. 5

[67] L. Zhang, Y. Shen, and H. Li. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE TIP*, 23(10):4270–4281, 2014. 5

[68] L. Zhang, D. Zhang, and X. Mou. FSIM: a feature similarity index for image quality assessment. *IEEE TIP*, 20(8):2378–2386, 2011. 2, 8

[69] Lin Zhang, Lei Zhang, and Xuanqin Mou. Rfsim: A feature based image quality assessment metric using riesz transforms. In *2010 IEEE International Conference on Image Processing*, pages 321–324. IEEE, 2010. 5

[70] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1, 2, 4, 5, 8

[71] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021. 3