Perceptual Image Quality Assessment with Transformers

Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee

LG Electronics

Seoul, Korea

{manri.cheon, sungjun.yoon, byungyeon.kang, junwoo.lee}@lge.com

Abstract

In this paper, we propose an image quality transformer (IQT) that successfully applies a transformer architecture to a perceptual full-reference image quality assessment (IQA) task. Perceptual representation becomes more important in image quality assessment. In this context, we extract the perceptual feature representations from each of input images using a convolutional neural network (CNN) backbone. The extracted feature maps are fed into the transformer encoder and decoder in order to compare a reference and distorted images. Following an approach of the transformer-based vision models [18, 55], we use extra learnable quality embedding and position embedding. The output of the transformer is passed to a prediction head in order to predict a final quality score. The experimental results show that our proposed model has an outstanding performance for the standard IQA datasets. For a large-scale IQA dataset containing output images of generative model, our model also shows the promising results. The proposed IQT was ranked first among 13 participants in the NTIRE 2021 perceptual image quality assessment challenge [23]. Our work will be an opportunity to further expand the approach for the perceptual IQA task.

1. Introduction

Perceptual image quality assessment (IQA) is an important topic in the multimedia systems and computer vision tasks [11, 42, 56]. One of the goals of the image processing is to improve the quality of the content to an acceptable level for the human viewers. In this context, the first step toward generating acceptable contents is to accurately measure the perceptual quality of the content, which can be performed via subjective and objective quality assessment [49, 9, 25, 19]. 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 subjective ratings. However, it is time-consuming and expensive. Thus, objective quality as-

sessment performed by objective metrics is widely used to automatically predict perceived quality [51, 52, 40, 59, 53].

However, with the recent advances in deep learning-based image restoration algorithms, accurate prediction of the perceived quality has become more difficult. In particular, image restoration models based on generative adversarial network (GAN) [20] have been developed in order to improve the perceptual aspect of the result images [48, 2, 12, 8]. However, it sometimes generates output images with unrealistic artifacts. The existing objective metrics such as Peak Signal-to-Noise Ratio (PSNR), a structural similarity index (SSIM) [51], and conventional quality metrics are insufficient to predict the quality of this kind of outputs. In this respect, recent works [61, 16, 21, 39, 3] based on perceptual representation exhibit a better performance at the perceptual IQA task. As various image restoration algorithms are developed, however, it is still required to develop the IQA algorithm that accurately predicts the perceptual quality of images generated by emerging algorithms.

In recent years, based on the success in the natural language processing (NLP) field, the transformer [46] architecture has been applied in the computer vision field [27]. A wider research area in the computer vision has been improved based on the transformer, such as recognition task [4, 45, 18], generative modelling [37, 26, 7], low-level vision [6, 54, 30], etc. However, few attempts were made in the field of the image and video quality assessment. In a recent study, You and Korhonen proposed the application of transformer in image quality assessment [55]. They achieved outstanding performance on two publicly available large-scale blind image quality databases. With our knowledge, however, this study is the only transformer-based approach for image quality assessment. Therefore, it is urgently needed to investigate whether the transformer-based approach works well in the field of perceptual image quality assessment. Especially, it should be investigated whether this structure is applicable to a full-reference (FR) model aiming to measure the perceptual similarity between two images. In addition, it is also necessary to evaluate whether this approach can accurately predict the perceptual quality
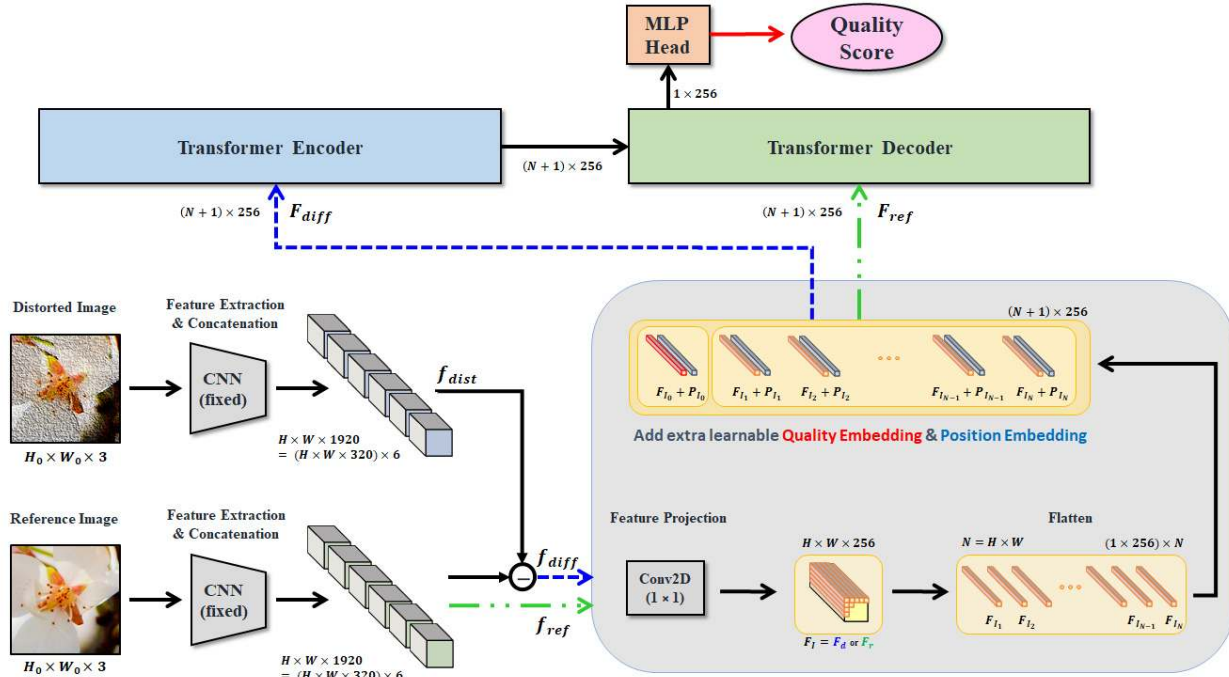
Figure 1. Model architecture of proposed image quality transformer (IQT). Note that F_I denotes F_d and F_r in Eqs. 1 and 2, respectively.

for the latest GAN-based artifacts.

In this study, we propose an Image Quality Transformer (IQT), which is the FR image quality assessment method as shown in Fig. 1. To tackle the perceptual aspects, a convolutional neural network (CNN) backbone is used to extract perceptual representations from an input image. Based on the transformer encoder-decoder architecture, the proposed model is trained to predict the perceptual quality accurately. The proposed model was ranked in the first place among 13 participants in the NTIRE 2021 challenge on perceptual image quality assessment [23] at the CVPR 2021.

The rest of this article is organized as follows. The following section presents the related work. Section 3 describes the proposed method and the experiments are given in Section 4. Finally, conclusions are given in Section 5.

2. Related Work

Image Quality Assessment. The most important goal of the developing objective IQA is to accurately predict the perceived quality by human viewers. In general, the objective IQA methods can be classified into three categories according to the existence of reference information: FR [51, 52, 40, 32, 5], reduced-reference (RR) [43], and no-reference (NR) [36, 35] IQA methods. The NR method is useful for the system because of its feasibility. However, the absence of a reference makes it challenging to predict image quality accurately compared to the FR method. The FR

method focuses more on visual similarity or dissimilarity between two images, and this method still plays an important role in the development of image processing system.

The representative of the commonly and widely used quality FR metric is the PSNR. It has the advantage of convenience for optimization; however, it tends to poorly predict perceived visual quality. Wang *et al.* proposed the SSIM [51] that is based on the fact that the human visual system (HVS) is highly correlated to structural information. Since that, various FR metrics have been developed to take into account various aspects of human quality perception, e.g., information-theoretic criterion [40, 41], structural similarity [52, 59], etc. Recently, CNN-based IQA methods as well as other low-level computer vision tasks have been actively studied [61, 3, 39, 17, 24]. Zhang *et al.* proposed a learned perceptual image patch similarity (LPIPS) metric [61] for FR-IQA. The LPIPS showed that trained deep features that are optimized by the Euclidean distance between distorted and reference images are effective for IQA compared to the conventional IQA methods. Ding *et al.* proposed the metric that is robust to texture resampling and geometric transformation based on spatial averages of the feature maps [16]. Various IQA methods including aforementioned metrics are included in our experiments for performance comparison.

The primary criterion of performance measurement is the accuracy of the metrics. Pearson linear correlation coefficient (PLCC) followed by the third-order polynomial non-linear regression [42] is usually used in order to evaluate

the accuracy of the methods. Spearman rank order correlation coefficient (SRCC) and the Kendall rank order correlation coefficient (KRCC) are used to estimate the monotonicity and consistency of the quality prediction. Additional statistical method [29] and an ambiguity based approach have also been proposed in [10]. In our study, we select the SRCC, KRCC, and PLCC as performance evaluation metrics.

Vision Transformer. The transformer [46] consists of multi-head attentions (MHAs), multi-layer perceptrons (MLPs), layer normalizations (LNs) [1], and residual connections. Unlike the CNN, the transformer has a minimum inductive bias and can scale with the length of the input sequence without limiting factors. Recently, it has emerged that the transformer has combined with the CNN using the self-attention [4], and some of which have completely replaced CNN [47].

The transformer is mainly self-attention based approach. Since the self-attention layer aggregates global information from the entire input sequence, therefore, the model can capture the entire image for measuring the perceptual quality of the whole image. Vision Transformer (ViT) [18] is a representative success model among transformer-based vision models. A hybrid architecture was proposed for image recognition using a concord of CNN and the transformer encoder. It replaces the pixel patch embedding with the patches extracted from the CNN feature map. This architecture could be applied well in the IQA task, because the effectiveness of the deep features on the perceptual IQA task was demonstrated in recent studies [61, 16, 55]. In DETR [4], the encoder-decoder architecture is employed and the decoder takes learned positional embeddings as object queries for object detection. This approach could be applied to the FR-IQA model that compares two images and measures the similarity. To measure similarity, one of the two images can be adopted as the query information in the self-attention layer. From the successful approaches using the transformer, we learn the direction to develop the perceptual IQA method with the transformer.

Vision Transformer based IQA. Inspired by ViT, TRIQ [55] naturally attempts to solve the blind IQA task using the transformer with the MLP head. In order to exploit ViT and handle images with different resolution, the TRIQ model defines the positional embedding with sufficient length to cover the maximal image resolution. The transformer encoder employs adaptive positional embedding, which handles a image with arbitrary resolutions. The output of the encoder is fed into the MLP head and the MLP head predicts the perceived image quality.

Basically, similar to the TRIQ, our proposed model applies the transformer architecture for the IQA task. How-
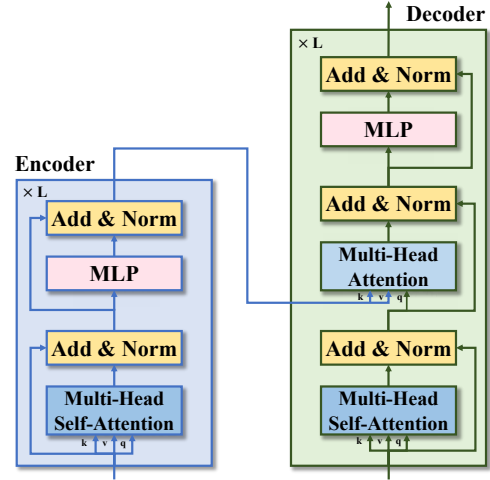

Figure 2. The transformer encoder and decoder.

ever, additional aspects are considered in order to design the perceptual FR-IQA with the transformer. First, the transformer encoder-decoder architecture is an important point in our approach. The reference information and the difference information between the distorted and reference images are employed as an input into the transformer. Second, we adopt the Siamese architecture to extract both the input feature representations from the reference and distorted images. For each image, by concatenating multiple feature maps extracted from intermediate layers, we obtained sufficient information for the model.

3. Proposed Method

The proposed method that is illustrated in Fig. 1 consists of three main components: a feature extraction backbone, a transformer encoder-decoder, and a prediction head. First, we use a CNN backbone to extract feature representations from both reference and distorted input images. The extracted feature maps are projected to fixed size of vectors and flattened. In order to predict perceived quality, the trainable extra `[quality]` embedding is added to the sequence of embedded feature. It is similar to approach using `[class]` token in previous transformer models [15, 18, 45]. The position embedding is also added in order to maintain the positional information. We pass this input feature embedding into the transformer encoder and decoder. The transformer encoder and decoder are based on the standard architecture of the transformer [46], where the structure is briefly illustrated in Fig. 2. The first vector of the output embedding of the decoder is fed into the MLP head in order to predict a single perceptual quality score.

Feature Extraction Backbone. A conventional CNN network, Inception-Resnet-V2 [44], is employed as the

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

Database	# Ref.	# Dist.	Dist. Type	# Dist. Type	# Rating	Rating Type	Env.
LIVE [42]	29	779	traditional	5	25k	MOS	lab
CSIQ [32]	30	866	traditional	6	5k	MOS	lab
TID2013 [38]	25	3,000	traditional	25	524k	MOS	lab
KADID-10k [33]	81	10.1k	traditional	25	30.4k	MOS	crowdsourcing
PIPAL [22]	250	29k	trad.+alg. outputs	40	1.13m	MOS	crowdsourcing

feature extraction backbone network. Pretrained weights on ImageNet [14] is imported and frozen. Feature maps from six intermediate layers of Inception-Resnet-V2, i.e., $\{mixed\_5b, block35\_2, block35\_4, block35\_6, block35\_8, block35\_10\}$, are extracted. The extracted feature maps have the same shape $f_{layer} \in \mathbb{R}^{H \times W \times c}$, where $c = 320$, and they are concatenated into feature map. In other words, for an input image $I \in \mathbb{R}^{H_0 \times W_0 \times 3}$, the feature map $f \in \mathbb{R}^{H \times W \times C}$, where $C = 6 \times c$, is extracted.

Both *reference* and *distorted* images are used; therefore, the two input feature maps, f_{ref} and f_{dist}, are employed for the transformer, respectively. In order to obtain difference information between reference and distorted images, a *difference* feature map, f_d, is also used. It can be simply obtained by subtraction between two feature maps of reference and distorted images, i.e., $f_{diff} = f_{ref} - f_{dist}$.

Transformer encoder. A *difference* feature embedding, $F_d \in \mathbb{R}^{N \times D}$, is used as the input of the transformer encoder. We first reduce the channel dimension of the f_d to the transformer dimension D using a 1×1 convolution. Then, we flatten the spatial dimensions, which means the number of patches in the feature map is set as $N = H \times W$. As often used in the vision transformer models [18, 55], we append extra quality embedding at the beginning of the input feature embedding as F_{d_0}. And the trainable position embedding $P_d \in \mathbb{R}^{(1+N) \times D}$ are also added in order to retain the positional information. The calculation of the encoder can be formulated as

$$
\begin{aligned}
y_0 &= [F_{d_0} + P_{d_0}, F_{d_1} + P_{d_1}, ..., F_{d_N} + P_{d_N}], \\
q_i &= k_i = v_i = y_{i-1}, \\
y'_i &= LN(MHA(q_i, k_i, v_i) + y_{i-1}), \\
y_i &= LN(MLP(y'_i) + y'_i), \quad i = 1, ..., L \\
[F_{E_0}, F_{E_1}, ..., F_{E_N}] &= y_L,
\end{aligned}
\tag{1}
$$

where L denotes the number of the encoder layers. The output of the encoder $F_E \in \mathbb{R}^{(1+N) \times D}$ has the same size to that of the input feature embedding.

Transformer decoder. The decoder takes the *reference* feature embedding $F_r \in \mathbb{R}^{N \times D}$, obtained through the channel reduction and flattening. The extra quality embedding and position embedding are also added to it. The output of the encoder, F_E, is used as an input of the decoder, and it is

used as a key-value in the second MHA layer. The calculation of the decoder can be formulated as:

$$
\begin{aligned}
y_L &= [F_{E_0}, F_{E_1}, ..., F_{E_N}], \\
z_0 &= [F_{r_0} + P_{r_0}, F_{r_1} + P_{r_1}, ..., F_{r_N} + P_{r_N}], \\
q_i &= k_i = v_i = z_{i-1}, \\
z'_i &= LN(MHA(q_i, k_i, v_i) + z_{i-1}), \\
q'_i &= z'_i, \quad k'_i = v'_i = y_L, \\
z''_i &= LN(MHA(q'_i, k'_i, v'_i) + z'_i), \\
z_i &= LN(MLP(z''_i) + z''_i), \quad i = 1, ..., L \\
[F_{D_0}, F_{D_1}, ..., F_{D_N}] &= z_L,
\end{aligned}
\tag{2}
$$

where L denotes the number of decoder layers. The output embedding $F_D \in \mathbb{R}^{(1+N) \times D}$ of the decoder is finally obtained.

Head. The final quality prediction is computed in the prediction MLP head. The first vector of the decoder output, $F_{D_0} \in \mathbb{R}^{1 \times D}$ in Eq. 2, is fed into the MLP head, which contains the quality information. The MLP head consists of two fully connected (FC) layers, and the first FC layer is used followed by the ReLU activation. The second FC layer has one channel to predict a single score.

4. Experiments

4.1. Datasets

We employ five databases that are commonly used in the research of perceptual image quality assessment. The LIVE Image Quality Assessment Database (LIVE) [42], the Categorical Subjective Image Quality (CSIQ) database [32], and the TID2013 [38] are the databases that serve as baselines for full-reference IQA studies. These datasets only include traditional distortion types and the subjective scores are measured in the controlled laboratory environment. KADID-10k [33] is a large-scale IQA dataset and is chosen as the training dataset in our experiment. It is three times larger compared to the TID2013 [38] and the ratings are collected from crowdsourcing. The PIPAL [22] dataset is used for both the training and evaluation of the model in this study. A large quantity of distorted images including GAN based algorithms' outputs and following human ratings are included in the PIPAL dataset. It is challenging for existing metrics to predict perceptual quality accurately

Table 2. Performance comparison of the IQA methods on three standard IQA databases, i.e., LIVE [42], CSIQ [32], and TID2013 [38], in terms of SRCC and KRCC. The top three performing methods are highlighted in bold face. Some results are borrowed from [16, 21].

Method	LIVE[42]		CSIQ[32]		TID2013[38]	
	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
PSNR	0.873	0.680	0.810	0.601	0.687	0.496
SSIM [51]	0.948	0.796	0.865	0.680	0.727	0.545
MS-SSIM [52]	0.951	0.805	0.906	0.730	0.786	0.605
VSI [58]	0.952	0.806	0.943	0.786	**0.897**	**0.718**
MAD [32]	**0.967**	**0.842**	**0.947**	**0.797**	0.781	0.604
VIF [40]	0.964	0.828	0.911	0.743	0.677	0.518
FSIMc [59]	**0.965**	**0.836**	0.931	0.769	0.851	0.667
NLPD [31]	0.937	0.778	0.932	0.769	0.800	0.625
GMSD [53]	0.960	0.827	**0.950**	**0.804**	0.804	0.634
WaDIQaM [3]	0.947	0.791	0.909	0.732	0.831	0.631
PieAPP [39]	0.919	0.750	0.892	0.715	**0.876**	**0.683**
LPIPS [61]	0.932	0.765	0.876	0.689	0.670	0.497
DISTS [17]	0.954	0.811	0.929	0.767	0.830	0.639
SWD [21]	-	-	-	-	0.819	0.634
IQT (ours)	**0.970**	**0.849**	**0.943**	**0.799**	**0.899**	**0.717**
IQT-C (ours)	0.917	0.737	0.851	0.649	0.804	0.607

[21]. Table 1 summarizes the characteristics of the datasets employed in this study.

4.2. Implementation details

We denote our model trained on the KADID-10k as IQT. The hyper-parameters for the model are set as follow: i) the number of encoder and decoder layer is set to 2 (i.e., $L = 2$), ii) the number of heads in the MHA is set to 4 (i.e., $H = 4$), iii) the transformer dimension is set to 256 (i.e., $D = 256$), iv) dimension of the MLP in the encoder and decoder is set to 1024 (i.e., $D_{feat} = 1024$), v) the dimension of the first FC layer in MLP head is set to 512 (i.e., $D_{head} = 512$).

In the training phase, a given image is cropped to obtain image patches. The dimension of the patch fed into the proposed IQT is $256 \times 256 \times 3$. The number of patches in the feature map is set to $N = 891$. In the testing phase, image patches are also acquired from the given image pair. We extract M overlapping patches and predict final quality score by averaging M individual quality scores of the patches. The stride size is set as large as possible to cover the entire image with fewer patches.

Data augmentation including horizontal flip and random rotation is applied during the training. The training is conducted using an ADAM [28] optimizer with a batch size of 16. Initial learning rate 2×10^{-4} and cosine learning rate decay are set. The training loss is computed using a mean squared error (MSE) loss function. Our network is implemented using Tensorflow framework. It roughly takes a half day with a single NVIDIA TITAN RTX to train our model.

4.3. Results

The proposed IQT shows that the transformer based model is sufficiently competitive compared to existing approaches for the dataset that has traditional distortions. Our model is trained on KADID-10k and, then, the performance
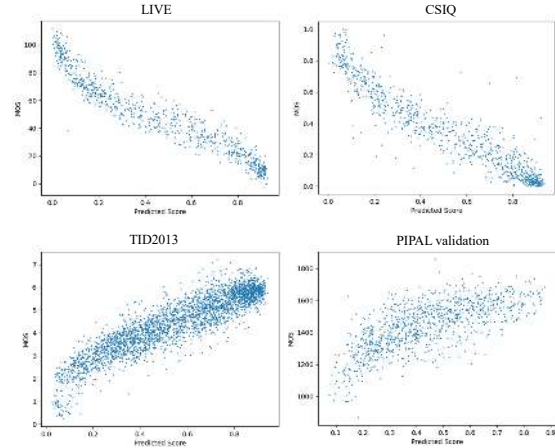

Figure 3. Scatter plots of ground-truth mean opinion scores (MOSs) against predicted scores of proposed IQT on LIVE, CSIQ, TID2013, and PIPAL datasets. The predicted scores are obtained from the model trained on KADID-10k dataset.

on the three standard IQA datasets is evaluated. The performance comparison result is reported in Table 2 and the scatter plots of the predicted scores of IQT and the ground-truth MOS are also presented in Fig. 3. For LIVE and TID2013 databases, the proposed IQT shows the best performance in terms of SRCC. Also, it is ranked in the top three in all benchmarks in terms of SRCC and KRCC. In particular, our method shows better performance than recent deep learning-based methods [3, 39, 61, 16, 21] for most cases.

Example images of the PIPAL validation dataset and following PSNR, SSIM [51], MS-SSIM [52], LPIPS [61], and proposed IQT are illustrated in Fig. 4. From the left to the right, the perceptually better images to worse images are listed based on MOS. Our proposed IQT predicts the quality scores similar to MOS in terms of the superiority. There
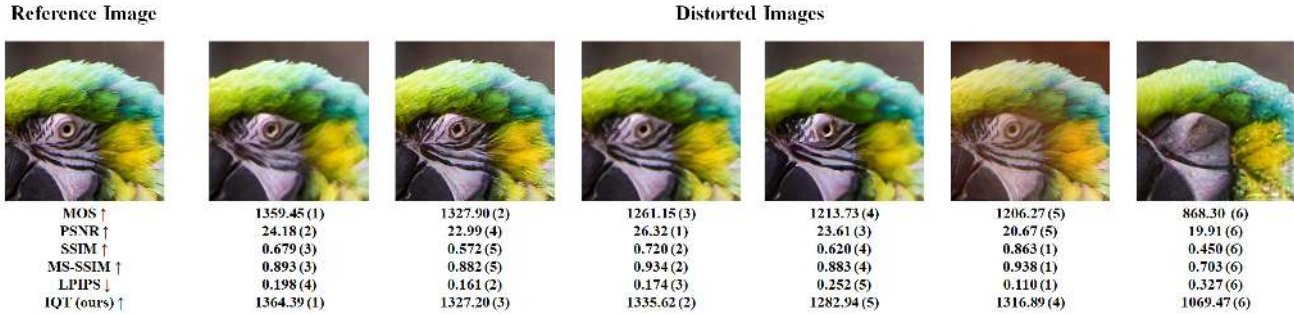
Figure 4. Example images from validation dataset of the NTIRE 2021 challenge. For each distorted image, predicted scores of PSNR, SSIM [51], MS-SSIM [52], LPIPS [61], and proposed IQT are listed. MOS denotes the ground-truth human rating. The number in the parenthesis denotes the rank among considered distorted images in this figure.

exist the images that are clearly distinguished by all methods, however, it is difficult to accurately predict perceptual quality for some images.

Our model is also evaluated on PIPAL [22] dataset. The IQT trained on KADID-10k dataset shows the best performance among all metrics. The benchmark results comparing the existing IQA methods on PIPAL validation and testing datasets are shown in Table 3. Corresponding scatter plots of the predicted scores of IQT and the ground-truth

MOS for validation dataset are also presented in Fig. 3. It is shown that our method could also be a promising approach in the field of quality assessment on various datasets including generative models' output images. Moreover, as shown in the results on the standard IQA datasets, our model shows a robust performance on different dataset.

Fig. 5 shows the examples of attention maps from the IQT model. It refers to the area where the model focuses more when predicting the perceptual quality. From our
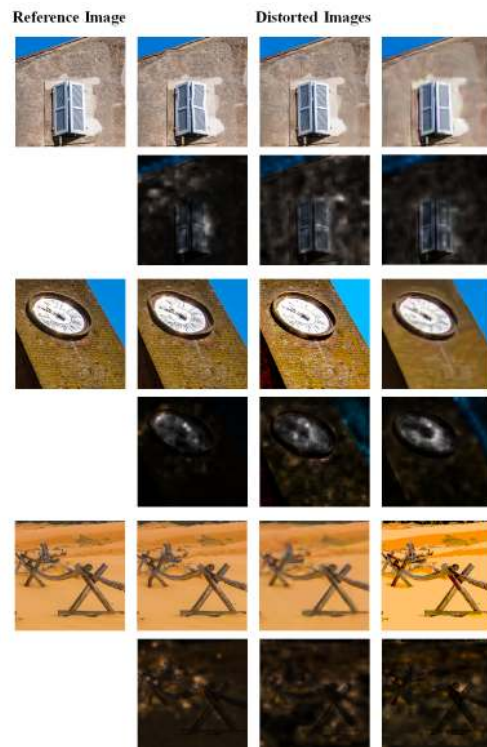
Table 3. Performance comparison of IQA methods on PIPAL [22] dataset. Main score is calculated with summation of PLCC and SRCC. The top performing method is highlighted in bold. Some results are provided from the NTIRE 2021 IQA challenge report [23].

Method	Validation		Testing	
	PLCC	SRCC	PLCC	SRCC
PSNR	0.292	0.255	0.277	0.249
SSIM [51]	0.398	0.340	0.394	0.361
MS-SSIM [52]	0.563	0.486	0.501	0.462
VIF [40]	0.524	0.433	0.479	0.397
VSNR [5]	0.375	0.321	0.411	0.368
VSI [58]	0.516	0.450	0.517	0.458
MAD [32]	0.626	0.608	0.580	0.543
NQM [13]	0.416	0.346	0.395	0.364
UQI [50]	0.548	0.486	0.450	0.420
IFC [41]	0.677	0.594	0.555	0.485
GSM [34]	0.469	0.418	0.465	0.409
RFSIM [60]	0.304	0.266	0.328	0.304
SRSIM [57]	0.654	0.566	0.636	0.573
FSIM [59]	0.561	0.467	0.571	0.504
FSIMc [59]	0.559	0.468	0.573	0.506
NIQE [36]	0.102	0.064	0.132	0.034
MA [35]	0.203	0.201	0.147	0.140
PI [2]	0.166	0.169	0.145	0.104
LPIPS-Alex [61]	0.646	0.628	0.571	0.566
LPIPS-VGG [61]	0.647	0.591	0.633	0.595
PieAPP [39]	0.697	0.706	0.597	0.607
WaDIQaM [3]	0.654	0.678	0.548	0.553
DISTS [17]	0.686	0.674	0.687	0.655
SWD [22]	0.668	0.661	0.634	0.624
IQT (ours)	**0.741**	**0.718**	-	-
IQT-C (ours)	**0.876**	**0.865**	**0.790**	**0.799**

Figure 5. Visualization of attention maps from the proposed IQT. The center-cropped images are randomly sampled from the PI-PAL [22] dataset. Attention maps are averaged over all attention weights in the encoder and decoder.

Table 4. Comparison of performance on three standard IQA databases depending on the inputs to the transformer encoder and decoder. The top performing method is highlighted in bold face.

No.	Encoder			Decoder			LIVE	CSIQ	TID2013
	F_{dist}	F_{ref}	F_{diff}	F_{dist}	F_{ref}	F_{diff}	SRCC/KRCC	SRCC/KRCC	SRCC/KRCC
(1)	✓			✓			0.901 / 0.713	0.768 / 0.575	0.646 / 0.468
(2)	✓				✓		0.934 / 0.767	0.855 / 0.670	0.739 / 0.548
(3)		✓		✓			0.954 / 0.805	0.865 / 0.680	0.755 / 0.564
(4)	✓					✓	0.967 / 0.838	0.944 / 0.803	0.884 / 0.698
(5)		✓				✓	0.967 / 0.837	0.945 / 0.803	0.881 / 0.694
(6)			✓	✓			0.969 / 0.843	0.945 / 0.803	0.897 / **0.714**
(7)			✓		✓		**0.970 / 0.845**	**0.947 / 0.805**	**0.896** / 0.712
(8)			✓			✓	0.968 / 0.840	0.942 / 0.795	0.889 / 0.704

model architecture, the learned attention weights exist in the MHA layers of the encoder and decoder. We visualize the attention maps by averaging all of the attention weights and resizing to the image size. It is observed that the attention is spatially localized or spread uniformly across whole image depending on the image and distortion type. It is important to see the entire image and, then, focus on a localized region when one perceives the quality of the image. Our approach to determine the important region based on the self-attention mechanism will be useful to predict the quality.

4.4. Ablations

The use of the difference information between reference and distorted images is one of the important factors in the proposed architecture. As mentioned in the previous section 3, the input into the encoder or decoder is a feature embedding and there are three types available, i.e., F_{ref}, F_{dist}, and F_{diff}. To investigate the effect of input types and location, we conduct ablation experiment and the results of performance comparison are shown in Table 4.

First, it is found that the use of the difference feature embedding as the input is a better choice than using only reference and distorted feature embeddings directly on the input. It is shown that the models (4)-(8) have better performance than models (1)-(3) in Table 4. From this experiment, the model (7) is selected for our model design and this means that the F_{diff} and F_{ref} are used into the encoder and decoder, respectively. When difference information enters the encoder or decoder, there is no significant performance difference between putting distorted or reference feature embedding in the other side. We can find the similar results between the models (4) and (5), and between the models (6) and (7). From this experiment, it is concluded that the difference information is an important factor in the proposed architecture for the IQA task.

An additional experiment is conducted to prove that the difference information in the feature level is more effective than that in the image level. The comparison results are shown in Table 5. Application of the difference information in the feature level that is important in our model design results in a better performance for all datasets. In other words,

Table 5. Comparison of performance on the three standard IQA databases according to the method of using difference information. "Feature" refers to a difference operation conducted between feature maps extracted from the backbone. "Image" refers to the difference operation on RGB images.

Diff. Info.	LIVE	CSIQ	TID2013
	SRCC/KRCC	SRCC/KRCC	SRCC/KRCC
Feature	0.970 / 0.845	0.947 / 0.805	0.896 / 0.712
Image	0.954 / 0.809	0.946 / 0.798	0.862 / 0.671

the difference information in perceptual space is more useful to predict an image quality score compared to the RGB color space.

4.5. NTIRE 2021 Perceptual IQA Challenge

This work is proposed to participate in the NTIRE 2021 perceptual image quality assessment challenge [23]. The objective of this challenge is to develop a model predicting a value with high accuracy comparable to the ground-truth MOS. The PIPAL [22] dataset is used for the NTIRE 2021 challenge. For this challenge, we train our model on training dataset provided in the NTIRE 2021 challenge. The same model structure and both training and testing strategies are applied for the challenge. The model hyper-parameters are set as follow: $L = 1$, $D = 128$, $H = 4$, $D_{feat} = 1024$, and $D_{head} = 128$. The input image size of the model is set to $192 \times 192 \times 3$; therefore, we set the number of patches in feature map $N = 441$. The model for the NTIRE 2021 challenge is denoted as IQT-C to distinguish from the previously mentioned model IQT in Tables 2, 3 and 6.

The benchmark results of the IQT-C on validation and testing datasets of the NTIRE 2021 challenge are shown in Table 3. The scatter plot is also illustrated in Fig. 6. The IQT-C shows the best performance among all metrics. In addition, a better performance than the IQT model trained on the KADID-10k is also found. Final result of the challenge in testing phase is reported in Table 6. The rankings of the entries are determined in terms of main score, which is calculated with summation of PLCC and SRCC. Our model won the first place in terms of the main score among all participants. In terms of PLCC and SRCC, we obtain the first

Table 6. Performance comparison of the participants on testing dataset of the NTIRE 2021 challenge. Main score is calculated as the sum of PLCC and SRCC. The number in the parenthesis denotes the rank. Only a few of the teams are shown in this Table. This result is provided from the NTIRE 2021 IQA challenge report [23].

Entries	PLCC	SRCC	Main Score ↑
IQT-C (ours)	0.7896 (1)	0.7990 (2)	1.5885 (1)
Anonymous 1	0.7803 (2)	0.8009 (1)	1.5811 (2)
Anonymous 2	0.7707 (4)	0.7918 (3)	1.5625 (3)
Anonymous 3	0.7709 (3)	0.7770 (4)	1.5480 (4)
Anonymous 4	0.7615 (5)	0.7703 (6)	1.5317 (5)
Anonymous 5	0.7468 (7)	0.7744 (5)	1.5212 (6)
Anonymous 6	0.7480 (6)	0.7641 (7)	1.5121 (7)

Figure 6. Scatter plots of ground-truth mean opinion scores (MOSs) against predicted scores of IQT-C on the PIPAL validation dataset.

and second highest scores, respectively.

The model trained on the PIPAL shows outstanding performance for the validation and testing dataset in Table 3. However, on the other hands, it tends to increase the risk of over-fitting. When we evaluate the IQT-C model on the three standard IQA datasets, it shows much lower performance than the IQT trained on KADID-10k (Table 2). It is noteworthy noting that the IQT-C is the special case of our approach for the NTIRE 2021 challenge. However, there is a room for improvement in terms of robustness for any other distortion types when we train the IQT on PIPAL dataset. In addition, future work is needed to improve the model to solve this problem.

5. Conclusion

We proposed an IQT and it is appropriately applied to a perceptual image quality assessment task by taking an advantage of transformer encoder-decoder architecture. The IQT demonstrated the outstanding performance on the three standard IQA databases compared to many existing methods. Our method also showed the best performance for the latest IQA dataset that contains deep learning-based distorted images. The IQT showed another promising example that the transformer based approach can achieve a high performance even in the perceptual quality assessment task.

Despite the success of our model, there exists a room for improvement. Further investigation of the transformer based approach, especially considering more diverse resolutions and distortion types, is needed. In addition, developing a no-reference metric for perceptual quality assessment will be desirable that can be used in real-world scenarios.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[2] 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, 6
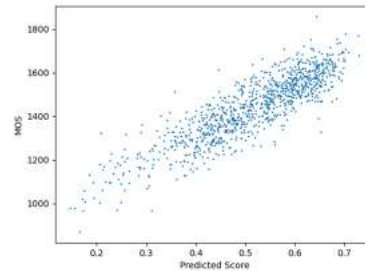
[3] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. Image Process.*, 27(1):206–219, 2017. 1, 2, 5, 6

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pages 213–229, 2020. 1, 3

[5] D. M. Chandler and S. S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Trans. Image Process.*, 16(9):2284–2298, 2007. 2, 6

[6] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao. Pre-trained image processing transformer. *arXiv preprint arXiv:2012.00364*, 2020. 1

[7] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In *Int. Conf. Mach. Learn.*, pages 1691–1703, 2020. 1

[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 Trans. Circuit Syst. Video Technol.*, 28(7):1467–1480, 2017. 1

[10] M. Cheon, T. Vigier, L. Krasula, J. Lee, P. Le Callet, and J.-S. Lee. Ambiguity of objective image quality metrics: A new methodology for performance evaluation. *Signal Processing: Image Communication*, 93:116150, 2021. 3

[11] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Trans. Broadcasting*, 57(2):165–182, 2011. 1

[12] J.-H. Choi, J.-H. Kim, M. Cheon, and J.-S. Lee. Deep learning-based image super-resolution considering quantitative and perceptual quality. *Neurocomputing*, 398:347–359, 2020. 1

[13] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik. Image quality assessment based on a degradation model. *IEEE Trans. Image Process.*, 9(4):636–650, 2000. 6

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In

IEEE/CVF Conf. Comput. Vis. Pattern Recog., pages 248–255, 2009. 4

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[16] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 1, 2, 3, 5

[17] 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, 5, 6

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3, 4

[19] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang. Perceptual quality assessment of smartphone photography. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 3677–3686, 2020. 1

[20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, page 2672–2680, 2014. 1

[21] 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, 5

[22] J. Gu, H. Cai, H. Chen, X. Ye, J. Ren, and C. Dong. PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In *Eur. Conf. Comput. Vis.*, pages 633–651, 2020. 4, 6, 7

[23] 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 Conf. Comput. Vis. Pattern Recog. Worksh.*, 2021. 1, 2, 6, 7,

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

[25] 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

[26] Y. Jiang, S. Chang, and Z. Wang. Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074*, 2021. 1

[27] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021. 1

[28] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*, 2015. 5

[29] Lukáš Krasula, Karel Fliegel, Patrick Le Callet, and Miloš Klíma. On the accuracy of objective image and video quality models: New methodology for performance evaluation.

In *Int. Conf. Quality of Multimedia Experience*, pages 1–6, 2016. 3

[30] M. Kumar, D. Weissenborn, and N. Kalchbrenner. Colorization transformer. *arXiv preprint arXiv:2102.04432*, 2021. 1

[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. 5

[32] E. C. Larson and D. M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006, 2010. 2, 4, 5, 6

[33] H. Lin, V. Hosu, and D. Saupe. KADID-10k: A large-scale artificially distorted iqa database. In *Int. Conf. Quality of Multimedia Experience*, pages 1–3, 2019. 4

[34] A. Liu, W. Lin, and M. Narwaria. Image quality assessment based on gradient similarity. *IEEE Trans. Image Process.*, 21(4):1500–1512, 2012. 6

[35] 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. 2, 6

[36] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 2, 6

[37] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran. Image transformer. In *Int. Conf. Mach. Learn.*, pages 4055–4064, 2018. 1

[38] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al. Image database TID2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015. 4, 5

[39] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 1808–1817, 2018. 1, 2, 5, 6

[40] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Trans. Image Process.*, 15(2):430–444, 2006. 1, 2, 5, 6

[41] H. R. Sheikh, A. C. Bovik, and G. De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans. Image Process.*, 14(12):2117–2128, 2005. 2, 6

[42] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 15(11):3440–3451, 2006. 1, 2, 4, 5

[43] Rajiv Soundararajan and Alan C Bovik. RRED indices: Reduced reference entropic differencing for image quality assessment. *IEEE Trans. Image Process.*, 21(2):517–526, 2011. 2

[44] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 1–7, 2017. 3

[45] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablay-rolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 1, 3

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 1, 3

[47] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *Eur. Conf. Comput. Vis.*, pages 108–126. Springer, 2020. 3

[48] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Eur. Conf. Comput. Vis. Worksh.*, pages 1–16, 2018. 1

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

[50] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, 2002. 6

[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 1, 2, 5, 6

[52] 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. 1, 2, 5, 6

[53] W. Xue, L. Zhang, X. Mou, and A. C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Trans. Image Process.*, 23(2):684–695, 2014. 1, 5

[54] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo. Learning texture transformer network for image super-resolution. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 5791–5800, 2020. 1

[55] J. You and J. Korhonen. Transformer for image quality assessment. *arXiv preprint arXiv:2101.01097*, 2020. 1, 3, 4

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

[57] L. Zhang and H. Li. SR-SIM: A fast and high performance iqa index based on spectral residual. In *IEEE Int. Conf. Image Process.*, pages 1473–1476, 2012. 6

[58] L. Zhang, Y. Shen, and H. Li. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Trans. Image Process.*, 23(10):4270–4281, 2014. 5, 6

[59] L. Zhang, D. Zhang, and X. Mou. FSIM: a feature similarity index for image quality assessment. *IEEE Trans. Image Process.*, 20(8):2378–2386, 2011. 1, 2, 5, 6

[60] L. Zhang, L. Zhang, and X. Mou. RFSIM: A feature based image quality assessment metric using riesz transforms. In *IEEE Int. Conf. Image Process.*, pages 321–324, 2010. 6

[61] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018. 1, 2, 3, 5, 6