MUSIQ: Multi-scale Image Quality Transformer

Junjie Ke[1], Qifei Wang[1], Yilin Wang[2], Peyman Milanfar[1], Feng Yang[1]
[1]Google Research, [2]Google
{junjiek, qfwang, yilin, milanfar, fengyang}@google.com

Abstract

Image quality assessment (IQA) is an important research topic for understanding and improving visual experience. The current state-of-the-art IQA methods are based on convolutional neural networks (CNNs). The performance of CNN-based models is often compromised by the fixed shape constraint in batch training. To accommodate this, the input images are usually resized and cropped to a fixed shape, causing image quality degradation. To address this, we design a multi-scale image quality Transformer (MUSIQ) to process native resolution images with varying sizes and aspect ratios. With a multi-scale image representation, our proposed method can capture image quality at different granularities. Furthermore, a novel hash-based 2D spatial embedding and a scale embedding is proposed to support the positional embedding in the multi-scale representation. Experimental results verify that our method can achieve state-of-the-art performance on multiple large scale IQA datasets such as PaQ-2-PiQ [41], SPAQ [11], and KonIQ-10k [16]. [1]

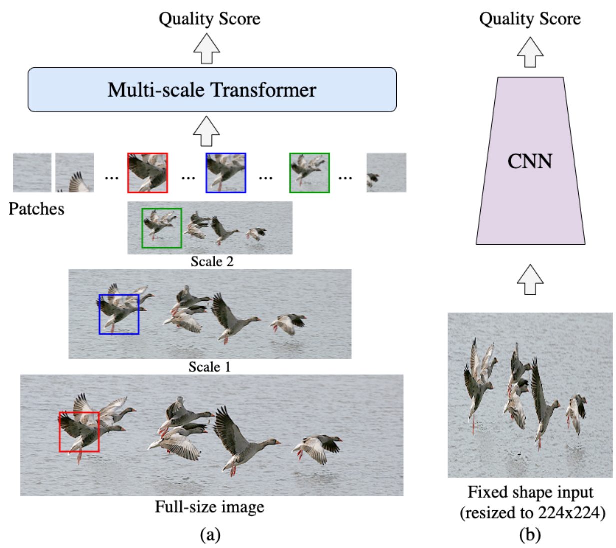

Figure 1. In CNN-based models (b), images need to be resized or cropped to a fixed shape for batch training. However, such preprocessing can alter image aspect ratio and composition, thus impacting image quality. Our patch-based MUSIQ model (a) can process the full-size image and extract multi-scale features, which aligns with the human visual system.

1. Introduction

The goal of image quality assessment (IQA) is to quantify perceptual quality of images. In the deep learning era, many IQA approaches [11, 33, 34, 41, 47] have achieved significant success by leveraging the power of convolutional neural networks (CNNs). However, the CNN-based IQA models are often constrained by the fixed-size input requirement in batch training, *i.e.*, the input images need to be resized or cropped to a fixed shape as shown in Figure 1 (b). This preprocessing is problematic for IQA because images in the wild have varying aspect ratios and resolutions. Resizing and cropping can impact image composition or introduce distortions, thus changing the quality of the image.

To learn IQA on the full-size image, the existing CNN-based approaches use either adaptive pooling or resizing to get a fixed-size convolutional feature map. MNA-CNN [24] processes a single image in each training batch which is not practical for training on a large dataset. Hosu et al. [15] extracts and stores fixed-size features offline, which costs additional storage for every augmented image. To keep aspect ratio, Chen et al. [7] proposes a dedicated convolution to preserve aspect ratio in the convolutional receptive field. Its evaluation verifies the importance of aspect-ratio-preserving (ARP) in the IQA tasks. But it still needs resizing and smart grouping for effective batch training.

In this paper, we propose a patch-based multi-scale image quality Transformer (MUSIQ) to bypass the CNN constraints on fixed input size and predict the quality effectively on the native resolution image as shown in Figure 1 (a). Transformer [36] is first proposed for natural language processing (NLP) and has recently been studied for various vision tasks [4–6, 10]. Among these, the Vision Transformer (ViT) [10] splits each image into a sequence of fixed-size patches, encodes each patch as a token, and then applies

[1]Checkpoints and code are available at https://github.com/google-research/google-research/tree/master/musiq

Transformer to the sequence for image classification. In theory, such kind of patch-based Transformer models can handle arbitrary numbers of patches (up to memory constraints), and therefore do not require preprocessing the input image to a fixed resolution. This motivates us to apply the patch-based Transformer on the IQA tasks with the full-size images as input.

Another aspect for improving IQA models is to imitate the human visual system which captures an image in a multi-scale fashion [1]. Previous works [15, 21, 46] have shown the benefit of using multi-scale features extracted from CNN feature maps at different depths. This inspires us to transform the native resolution image into a multi-scale representation, enabling the Transformer's self-attention mechanism to capture information on both fine-grained detailed patches and coarse-grained global patches. Besides, unlike the convolution operation in CNNs that has a relatively limited receptive field, self-attention can attend to the whole input sequence and it can therefore effectively capture the image quality at different granularities.

However, it is not straightforward to apply the Transformer on the multi-aspect-ratio multi-scale input. Although self-attention accepts arbitrary length of the input sequence, it is permutation-invariant and therefore cannot capture patch location in the image. To mitigate this, ViT [10] adds fixed-length positional embedding to encode the absolute position of each patch in the image. However, the fixed-length positional encoding fails when the input length varies. To solve this issue, we propose a novel hash-based 2D spatial embedding that maps the patch positions to a fixed grid to effectively handle images with arbitrary aspect ratios and resolutions. Moreover, since the patch locations at each scale are hashed to the same grid, it aligns spatially close patches at different scales so that the Transformer model can leverage information across multiple scales. In addition to the spatial embedding, a separate scale embedding is further introduced to help the Transformer distinguish patches coming from different scales in the multi-scale representation.

The main contributions of this paper can be summarized into three-folds:

- We propose a patch-based multi-scale image quality Transformer (MUSIQ), which supports processing full-size input with varying aspect ratios or resolutions, and allows multi-scale feature extraction.
- A novel hash-based 2D spatial embedding and a scale embedding are proposed to support positional encoding in the multi-scale representation, helping the Transformer capture information across space and scales.
- We apply MUSIQ on four large-scale IQA datasets. It consistently achieves the state-of-the-art performance on three technical quality datasets: PaQ-2-PiQ [41], KonIQ-10k [16], and SPAQ [11], and is on-par with the state-of-the-art on the aesthetic quality dataset AVA [29].

2. Related Work

Image Quality Assessment. Image quality assessment aims to quantitatively predict perceptual image quality. There are two important aspects for assessing image quality: technical quality [16] and aesthetic quality [29]. The former focuses on perceptual distortions while the latter also relates to image composition, artistic value and so on. In the past years, researchers proposed many IQA methods: early natural scene statistics based [13, 25, 27, 45], codebook-based [38, 40] and CNN-based [11, 33, 34, 41, 47]. CNN-based methods achieve the state-of-the-art performance. However they usually need to crop or resize images to a fixed size in batch training, which affects the image quality. Several methods have been proposed to mitigate the distortion from resizing and cropping in CNN-based IQA. An ensemble of multi-crops from the original image is proven to be effective for IQA [7, 15, 23, 32, 33], but it introduces non-negligible inference cost. In addition, MNA-CNN [24] handles full-size input by adaptively pooling the feature map to a fixed shape. However, it only accepts a single input image for each training batch to preserve the original resolution which is not efficient for large scale training. Hosu et al. [15] extracted and stored the fixed-sized features from the full-size image for model training which costs extra storage for every augmented image and is inefficient for large scale training. Chen et al. [7] proposed an adaptive fractional dilated convolution to adapt the receptive field according to the image aspect ratio. The method preserves aspect ratio but cannot handle full-size input without resizing. It also needs smart grouping strategy in mini-batch training.

Transformers in Vision. Transformers [36] were first applied to NLP tasks and achieved great performance [9, 22, 39]. Recent works applied transformers on various vision tasks [4–6, 10]. Among these, the Vision Transformer (ViT) [10] employs a pure Transformer architecture to classify images by treating an image as a sequence of patches. For batch training, ViT resizes the input images to a fixed squared size, *e.g.*, 224 $\times$ 224, where fixed number of patches are extracted and combined with fixed-length positional embedding. This constrains its usage for IQA since resizing will affect the image quality. To solve this, we propose a novel Transformer-based architecture that accepts the full-size image for IQA.

Positional Embeddings. Positional embeddings are introduced in Transformers to encode the order of the input sequence [36]. Without it, the self-attention operation is permutation-invariant [2]. Vaswani et al. [36] used deterministic positional embeddings generated from sinusoidal

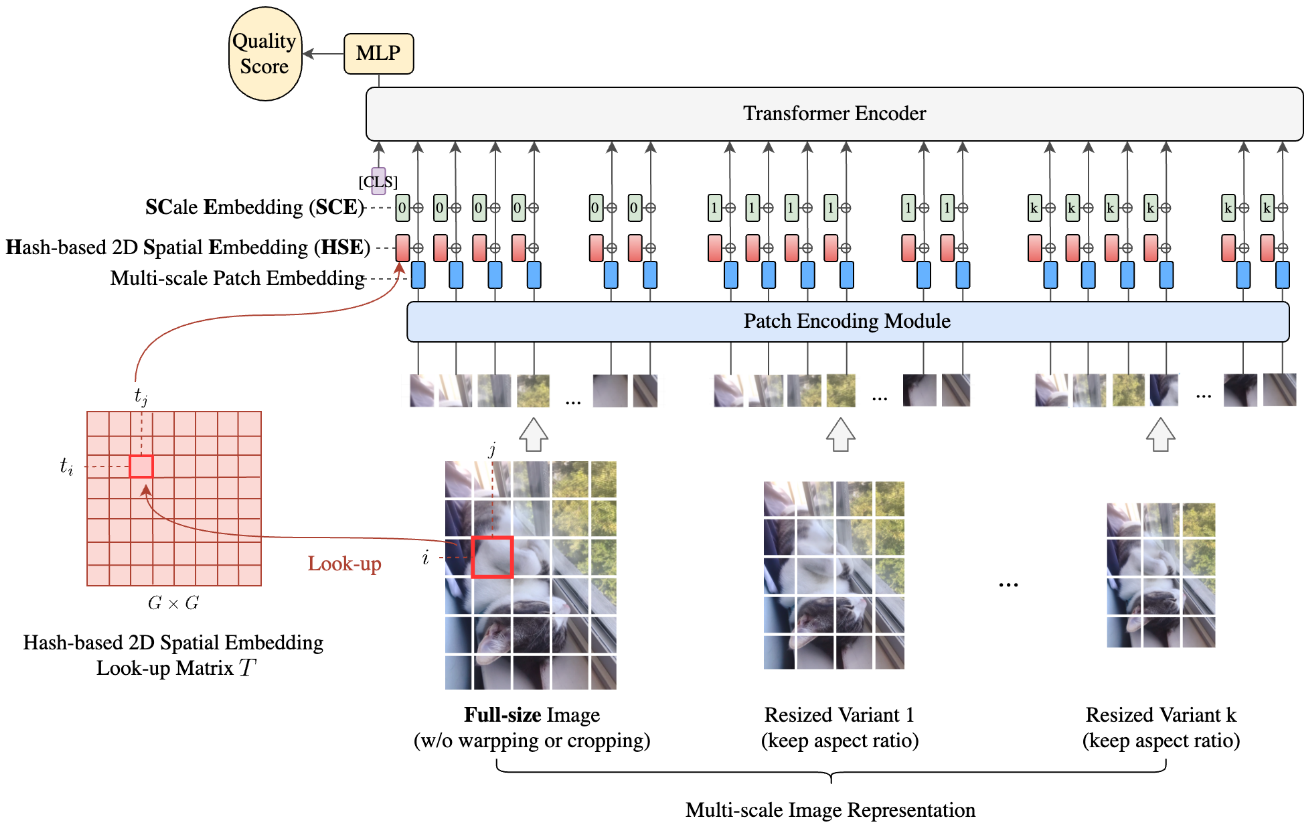

Figure 2. Model overview of MUSIQ. We construct a multi-scale image representation as input, including the native resolution image and its ARP resized variants. Each image is split into fixed-size patches which are embedded by a patch encoding module (blue boxes). To capture 2D structure of the image and handle images of varying aspect ratios, the spatial embedding is encoded by hashing the patch position (i, j) to (t_i, t_j) within a grid of learnable embeddings (red boxes). Scale Embedding (green boxes) is introduced to capture scale information. The Transformer encoder takes the input tokens and performs multi-head self-attention. To predict the image quality, we follow a common strategy in Transformers to add an [CLS] token to the sequence to represent the whole multi-scale input and use the corresponding Transformer output as the final representation.

functions. ViT [10] showed that the deterministic and learnable positional embeddings [12] works equally well. However, those positional embeddings are generated for fixed-length sequences. When the input resolution changes, the pre-trained positional embeddings is no longer meaningful. Relative positional embeddings [2, 31] is proposed to encode relative distance instead of absolute position. Although the relative positional embeddings can work for variable length inputs, it requires substantial modifications in Transformer attention and cannot capture multi-scale positions in our use case.

3. Multi-scale Image Quality Transformer

3.1. Overall Architecture

To tackle the challenge of learning IQA on full-size images, we propose a multi-scale image quality Transformer (MUSIQ) which can handle inputs with arbitrary aspect ratios and resolutions. An overview of the model is shown in Figure 2.

We first make a multi-scale representation of the input image, containing the native resolution image and its ARP resized variants. The images at different scales are partitioned into fixed-size patches and fed into the model. Since patches are from images of varying resolutions, we need to effectively encode the multi-aspect-ratio multi-scale input into a sequence of tokens (the small boxes in Figure 2), capturing both the pixel, spatial, and scale information.

To achieve this, we design three encoding components in MUSIQ, including: 1) A patch encoding module to encode patches extracted from the multi-scale representation (Section 3.2); 2) A novel hash-based spatial embedding module to encode the 2D spatial position for each patch (Section 3.3); 3) A learnable scale embedding to encode different scale (Section 3.4).

After encoding the multi-scale input into a sequence of tokens, we use the standard approach of prepending an extra learnable "classification token" (CLS) [9, 10]. The CLS token state at the output of the Transformer encoder serves as the final image representation. We then add a fully connected layer on top to predict the image quality score. Since MUSIQ only changes the input encoding, it is compatible with any Transformer variants. To demonstrate the effectiveness of the proposed method, we use the classic Transformer [36] (Appendix A) with a relatively lightweight setting to make model size comparable to ResNet-50 in our experiments.

3.2. Multi-scale Patch Embedding

Image quality is affected by both the local details and global composition. In order to capture both the global and local information, we propose to model the input image with a multi-scale representation. Patches from different scales enables the Transformer to aggregate information across multiple scales and spatial locations.

As shown in Figure 2, the multi-scale input is composed of the full-size image with height H, width W, channel C, and a sequence of ARP resized images from the full-size image using Gaussian kernel. The resized images have height h_k, width w_k, channel C, where $k = 1, ..., K$ and K is the number of resized variants for each input. To align resized images for a consistent global view, we fix the longer side length to L_k for each resized variant and yield:

$$\alpha_k = L_k / \max(H, W), \;\; h_k = \alpha_k H, \;\; w_k = \alpha_k W \quad (1)$$

α_k represents the resizing factor for each scale.

Square patches with size P are extracted from each image in the multi-scale representation. For images whose width or height are not multiples of P, we pad the image with zeros accordingly. Each patch is encoded into a D-dimension embedding by the patch encoder module. D is the latent token size used in the Transformer.

Instead of encoding the patches with a linear projection as in [10], we choose a 5-layer ResNet [14] with a fully connected layer of size D as the patch encoder module to learn a better representation for the input patch. We find that encoding the patch with a few convolution layers performs better than linear projection when pre-training on ILSVRC-2012 ImageNet [30] (see Section 4.4). Since the patch encoding module is lightweight and shared across all the input patches whose size P is small, it only adds a small amount of parameters.

The sequence of patch embeddings output from the patch encoder module are concatenated together to form a multi-scale embedding sequence for the input image. The number of patches from the original image and the resized ones are calculated as $N = HW/P^2$ and $n_k = h_k w_k / P^2$.

Since each input image has a different resolution and aspect ratio, H and W are different for each input and therefore N and n_k are different. To get fixed-length input during training, we follow the common practice in NLP [36] to zero-pad the encoded patch tokens to the same length. An input mask is attached to indicate the effective input, which will be used in the Transformer to perform masked self-attention (Appendix A.3). Note that the padding operation will not change the input because the padding tokens are ignored in the multi-head attention by masking them.

As previously mentioned, we fix the longer length to L_k for each resized variant. Therefore $n_k \leq L_k^2/P^2 = m_k$ and we can safely pad to m_k. For the native resolution image, we simply pad or cut the sequence to a fixed length l. The padding is not necessary during single-input evaluation because the sequence length can be arbitrary.

3.3. Hash-based 2D Spatial Embedding

Spatial positional embedding is important in vision Transformers to inject awareness of the 2D image structure in the 1D sequence input [10]. The traditional fixed-length positional embedding assigns an embedding for every input location. This fails for variable input resolutions where the number of patches are different and therefore each patch in the sequence may come from an arbitrary location in the image. Besides, the traditional positional embedding models each position independently and therefore it cannot align the spatially close patches from different scales.

We argue that an effective spatial embedding design for MUSIQ should meet the following requirements: 1) effectively encode patch spatial information under different aspect ratios and input resolutions; 2) spatially close patches at different scales should have close spatial embeddings; 3) efficient and easy to implement, non-intrusive to the Transformer attention.

Based on that, we propose a novel hash-based 2D spatial embedding (HSE) where the patch locating at row i, column j is hashed to the corresponding element in a $G \times G$ grid. Each element in the grid is a D-dimensional embedding.

We define HSE by a learnable matrix $T \in \mathbb{R}^{G \times G \times D}$. Suppose the input resolution is $H \times W$. The input image will be partitioned into $\frac{H}{P} \times \frac{W}{P}$ patches. For the patch at position (i, j), its spatial embedding is defined by the element at position (t_i, t_j) in T where

$$t_i = \frac{i \times G}{H/P}, \; t_j = \frac{j \times G}{W/P} \quad (2)$$

The D-dimensional spatial embedding T_{t_i, t_j} is added to the patch embedding element-wisely as shown in Figure 2. For fast lookup, we simply round (t_i, t_j) to the nearest integers. HSE does not require any changes in the Transformer attention module. Moreover, both the computation of t_i and t_j and the lookup are lightweight and easy to implement.

To align patches across scales, patch locations from all scales are mapped to the same grid T. As a result, patches located closely in the image but from different scales are mapped to spatially close embeddings in T, since i and H as well as j and W change proportionally to the resizing factor α. This achieves spatial alignment across different images from the multi-scale representation.

There is a trade-off between expressiveness and trainability with the choice hash grid size G. Small G may result in a lot of collision between patches which makes the model unable to distinguish spatially close patches. Large G wastes memory and may need more diverse resolutions to train. In our IQA setting where rough positional information is sufficient, we find once G is large enough, changing G only results in small performance differences (see Appendix B). We set $G = 10$ in the experiments.

3.4. Scale Embedding

Since we reuse the same hashing matrix for all images, HSE does not make a distinction between patches from different scales. Therefore, we introduce an additional scale embedding (SCE) to help the model effectively distinguish information coming from different scales and better utilize information across scales. In other words, SCE marks which input scale the patch is coming from in the multi-scale representation.

We define SCE as a learnable scale embedding $Q \in \mathbb{R}^{(K+1)\times D}$ for the input image with K-scale resized variants. Following the spatial embedding, the first element $Q_0 \in \mathbb{R}^D$ is added element-wisely to all the D-dimensional patch embeddings from the native resolution image. $Q_k \in \mathbb{R}^D, k = 1, ..., K$ are also added element-wisely to all the patch embeddings from the resized image at scale k.

3.5. Pre-training and Fine-tuning

Typically, the Transformer models need to be pre-trained on the large datasets, *e.g.* ImageNet, and fine-tuned on the downstream tasks. During the pre-training, we still keep random cropping as an augmentation to generate images of different sizes. However, instead of doing square resizing like the common practice in image classification, we intentionally skip resizing to prime the model for inputs with different resolutions and aspect ratios. We also employ common augmentations such as RandAugment [8] and mixup [44] in pre-training.

When fine-tuning on IQA tasks, we do not resize or crop the input image to preserve the image composition and aspect ratio. In fact, we only use random horizontal flipping for augmentation in fine-tuning. For evaluation, our method can be directly applied on the original image without aggregating multiple augmentations (*e.g.* multi-crops sampling).

When fine-tuning on the IQA datasets, we use common regression losses such as L1 loss for single mean opinion score (MOS) and Earth Mover Distance (EMD) loss to predict the quality score distribution [34]:

$$EMD(p, \hat{p}) = \left(\frac{1}{N}\sum_{m=1}^{N}|CDF_p(m) - CDF_{\hat{p}}(m)|^r\right)^{\frac{1}{r}} \quad (3)$$

where p is the normalized score distribution and $CDF_p(m)$ is the cumulative distribution function as $\sum_{i=1}^{m} p_i$.

4. Experimental Results

4.1. Datasets

We run experiments on four large-scale image quality datasets including three technical quality datasets (PaQ-2-PiQ [41], SPAQ [11], KonIQ-10k [16]) and one aesthetics quality dataset (AVA [29]).

PaQ-2-PiQ is the largest picture technical quality dataset by far which contains 40k real-world images and 120k cropped patches. Each image or patch is associated with a MOS. Since our model does not make a distinction between image and extracted patches, we simply use all the 30k full-size images and the corresponding 90k patches from the training split to train the model. We then run the evaluation on the 7.7k full-size validation and 1.8k test set.

SPAQ dataset consists of 11k pictures taken by 66 smartphones. For a fair comparison, we follow [11] to resize the raw images such that the shorter side is 512. We only use the image and its corresponding MOS for training, not including the extra tag information in the dataset.

KonIQ-10k contains 10k images selected from a large public multimedia database YFCC100M [35].

AVA is an image aesthetic assessment dataset. It contains 250k images with 10-scale score distribution for each.

For KonIQ-10k, we follow [33, 48] to randomly sample 80% images for each run and report the results on the remaining 20%. For other datasets, we use the same split as the previous literature.

4.2. Implementation Details

For MUSIQ, the multi-scale representation is constructed as the native resolution image and two ARP resized input ($L_1 = 224$ and $L_2 = 384$) by default. It therefore uses 3-scale input. Our method also works on 1-scale input using just the full-size image without resized variants. We report the results of this single-scale setting as MUSIQ-single.

We use patch size $P = 32$. The dimensions for Transformer input tokens are $D = 384$, which is also the dimension for pixel patch embedding, HSE and SCE. The grid size of HSE is set to $G = 10$. We use the classic Transformer [36] with lightweight parameters (384 hidden size, 14 layers, 1152 MLP size and 6 heads) to make the model size comparable to ResNet-50. The final model has around 27 million total parameters.

We pre-train our models on ImageNet for 300 epochs, using Adam with $\beta_1 = 0.9, \beta_2 = 0.999$, a batch size of 4096, 0.1 weight decay and cosine learning rate decay from 0.001. We set the maximum number of patches from full-size image l to 512 in training. For fine-tuning, we use SGD with momentum and cosine learning rate decay from 0.0002, 0.0001, 0.0001, 0.12 for 10, 30, 30, 20 epochs on PaQ-2-PiQ, KonIQ-10k, SPAQ, and AVA, respectively. Batch size is set to 512 for AVA, 96 for KonIQ-10k, and

method	Validation Set SRCC	Validation Set PLCC	Test Set SRCC	Test Set PLCC
BRISQUE [25]	0.303	0.341	0.288	0.373
NIQE [26]	0.094	0.131	0.211	0.288
CNNIQA [17]	0.259	0.242	0.266	0.223
NIMA [34]	0.521	0.609	0.583	0.639
Ying *et al.* [41]	0.562	0.649	0.601	0.685
MUSIQ-single	**0.563**	**0.651**	**0.640**	**0.721**
MUSIQ (Ours)	**0.566**	**0.661**	**0.646**	**0.739**
std	$\pm$0.002	$\pm$0.003	$\pm$0.005	$\pm$0.006

Table 1. Results on PaQ-2-PiQ full-size validation and test sets. Blue and black numbers in bold represent the best and second best respectively. We take numbers from [41] for the results of the reference methods.

method	SRCC	PLCC
BRISQUE [25]	0.665	0.681
ILNIQE [45]	0.507	0.523
HOSA [37]	0.671	0.694
BIECON [18]	0.618	0.651
WaDIQaM [3]	0.797	0.805
PQR [42]	0.880	0.884
SFA [20]	0.856	0.872
DBCNN [47]	0.875	0.884
MetaIQA [48]	0.850	0.887
BIQA [33] (**25** crops)	**0.906**	0.917
MUSIQ-single	0.905	**0.919**
MUSIQ (Ours)	**0.916**	**0.928**
std	$\pm$0.002	$\pm$0.003

Table 2. Results on KonIQ-10k dataset. Blue and black numbers in bold represent the best and second best respectively. We take numbers from [33, 48] for results of the reference methods.

128 for the rest. For AVA, we use the EMD loss with $r = 2$. For other datasets, we use the $L1$ loss.

The models are trained on TPUv3. All the results from our method are averaged across 10 runs. Spearman rank ordered correlation (SRCC), Pearson linear correlation (PLCC), and the standard deviation (std) are reported.

4.3. Comparing with the State-of-the-art (SOTA)

Results on PaQ-2-PiQ. Table 1 shows the results on the PaQ-2-PiQ dataset. Our proposed MUSIQ outperforms other methods on both the validation and test sets. Notably, the test set is entirely composed of pictures having at least one dimension exceeding 640 [41]. This is very challenging for traditional deep learning approaches where resizing is inevitable. Our method is able to outperform previous methods by a large margin on the full-size test set which verifies its robustness and effectiveness.

Results on KonIQ-10k. Table 2 shows the results on the KonIQ-10k dataset. Our method outperforms the SOTA methods. In particular, BIQA [33] needs to sample 25 crops from each image during training and testing. This kind of multi-crops ensemble is a way to mitigate the fixed shape constraint in the CNN models. But since each crop is only

method	SRCC	PLCC
DIIVINE [27]	0.599	0.600
BRISQUE [25]	0.809	0.817
CORNIA [40]	0.709	0.725
QAC [38]	0.092	0.497
ILNIQE [45]	0.713	0.721
FRIQUEE [13]	0.819	0.830
DBCNN [47]	0.911	0.915
Fang *et al.* [11] (w/o extra info)	0.908	0.909
MUSIQ-single	**0.917**	**0.920**
MUSIQ (Ours)	**0.917**	**0.921**
std	$\pm$0.002	$\pm$0.002

Table 3. Results on SPAQ dataset. Blue and black numbers in bold represent the best and second best respectively. We take numbers from [11] for results of the reference methods.

method	cls. acc.	MSE $\downarrow$	SRCC	PLCC
MNA-CNN-Scene [24]	0.765	-	-	-
Kong *et al.* [19]	0.773	-	0.558	-
AMP [28]	0.803	0.279	0.709	-
A-Lamp [23] (**50** crops)	0.825	-	-	-
NIMA (VGG16) [34]	0.806	-	0.592	0.610
NIMA (Inception-v2) [34]	0.815	-	0.612	0.636
MP_{ada} [32] ($\geq$ **32** crops)	0.830	-	-	-
Zeng *et al.* (ResNet101) [43]	0.808	0.275	0.719	0.720
Hosu *et al.* [15] (**20** crops)	0.817	-	**0.756**	**0.757**
AFDC + SPP (single warp) [7]	**0.830**	0.273	0.648	-
AFDC + SPP (**4** warps) [7]	**0.832**	0.271	0.649	0.671
MUSIQ-single	0.814	**0.247**	0.719	0.731
MUSIQ (Ours)	0.815	**0.242**	**0.726**	**0.738**
std	$\pm$0.121	$\pm$0.001	$\pm$0.001	$\pm$0.001

Table 4. Results on AVA dataset. Blue and black numbers in bold represent the best and second best respectively. cls. acc. stands for classification accuracy. MSE stands for mean square error. We take numbers from [7] for results of the reference methods.

a sub-view of the whole image, the ensemble is still an approximate approach. Moreover, it adds additional inference cost for every crop and sampling can introduce randomness in the result. Since MUSIQ takes the full-size image as input, it can directly learn the best aggregation of information across the full image and only one evaluation is involved.

Results on SPAQ. Table 3 shows the results on the SPAQ dataset. Overall, our model is able to outperform other methods in terms of both SRCC and PLCC.

Results on AVA. Table 4 shows the results on the AVA dataset. Our method achieves the best MSE and has top SRCC and PLCC. As previously discussed, instead of multi-crops sampling, our model can accurately predict image aesthetics by directly looking at the full-size image.

4.4. Ablation Studies

Importance of Aspect-Ratio-Preserving (ARP). CNN-based IQA models usually resize the input image to a square resolution without preserving the original aspect ratio. We argue that such preprocessing can be detrimental to IQA because it alters the image composition. To verify that, we

method	# Params	SRCC	PLCC
NIMA(Inception-v2) [34] (224 square input)	56M	0.612	0.636
NIMA(ResNet50)* (384 square input)	24M	0.624	0.632
ViT-Base 32* (384 square input) [10]	88M	0.654	0.664
ViT-Small 32* (384 square input) [10]	22M	0.656	0.665
MUSIQ w/ square resizing (512, 384, 224)	27M	0.706	0.720
MUSIQ w/ ARP resizing (512, 384, 224)	27M	0.712	0.726
MUSIQ w/ ARP resizing (full, 384, 224)	27M	**0.726**	**0.738**

Table 5. Comparison of ARP resizing and square resizing on AVA dataset. * means our implementation. ViT-Small* is constructed by replacing the Transformer backbone in ViT with our 384-dim lightweight Transformer. The last group of rows show our method with different resizing methods. Numbers in the bracket show the resolution used in the multi-scale representation.

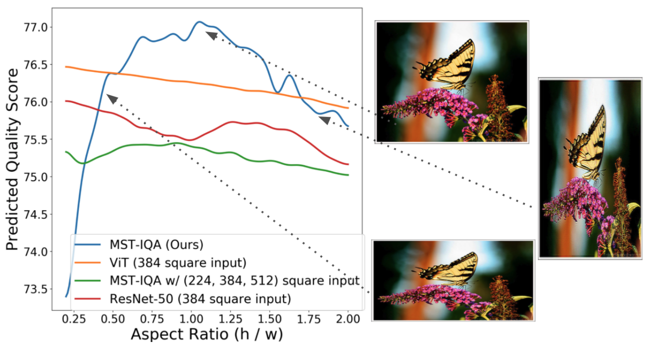

Figure 3. Model predictions for an image resized to different aspect ratios. The blue curve shows MUSIQ with ARP resizing. The green curve shows our model trained and evaluated with square input. Orange and red curves show the ViT and ResNet-50 with square input. MUSIQ can detect quality degradation due to unnatural resizing while other methods are not sensitive.

compare the performance of the proposed model with either square or ARP resizing. As shown in Table 5, ARP resizing performs better than square resizing, demonstrating the importance of ARP when assessing image quality.

To intuitively understand the importance of keeping aspect ratios in IQA, we follow [7] to artificially resize the same image into different aspect ratios and run models to predict quality scores. Since aggressive resizing will cause image quality degradation, a good IQA model should give lower scores to such unnatural looking images. As shown in Figure 3, MUSIQ (blue curve) is discriminative to the change of aspect ratios while scores from the other ones trained with square resizing are not sensitive to the change. This shows that ARP resizing is important and MUSIQ can effectively detect quality degradation due to resizing.

Effect of Full-size Input and the Multi-scale Input Composition. In Table 1 2 3 4, we compare using only the full-size input (MUSIQ-single) and the multi-scale input (MUSIQ). MUSIQ-single achieves promising results, showing the importance of preserving full-size input in IQA. The performance is further improved using multi-scale and the gain is larger on PaQ-2-PiQ and AVA because these two datasets have much more diverse resolutions than KonIQ-10k and SPAQ. This shows that multi-scale is important for effectively capturing quality information on real-world images with varying sizes.

Multi-scale Composition	SRCC	PLCC
(224)	0.600	0.667
(384)	0.618	0.695
(512)	0.620	0.691
(384, 224)	0.620	0.707
(512, 384, 224)	0.629	0.718
(full)	0.640	0.721
(full, 224)	0.643	0.726
(full, 384)	0.642	0.730
(full, 384, 224)	**0.646**	**0.739**
Average ensemble of (full), (224), (384)	0.640	0.710

Table 6. Comparison of multi-scale representation composition on PaQ-2-PiQ full-size test set. The multi-scale representation is composed of the resolutions shown in the brackets. Numbers in brackets indicate the longer side length L for ARP resizing. "full" means full-size input image.

Figure 4. Visualization of attention from the output tokens to the multi-scale representation (original resolution image and two ARP resized variants). Note that images here are resized to fit the grid, the model inputs are 3 different resolutions. The model is focusing on details in higher resolution image and on global area in lower resolution ones.

We also vary the multi-scale composition and show in Table 6 that multi-scale consistently improves performance on top of single-scale models. The performance gain of multi-scale is more than a simple ensemble of individual scales because an average ensemble of individual scales actually under-performs using only the full-size image. Since MUSIQ has full receptive field of the multi-scale input sequences, it can more effectively aggregate quality information across scales.

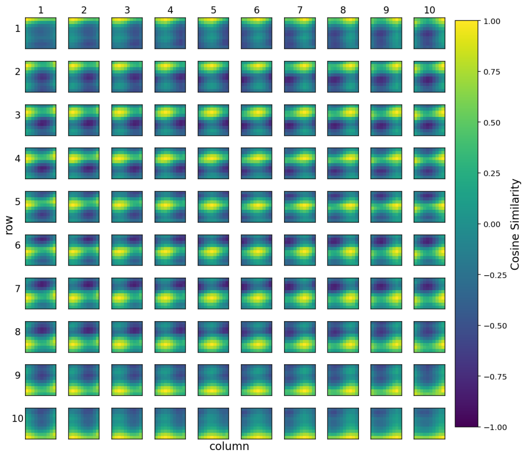

Figure 5. Visualization of the grid of hash-based 2D spatial embedding with $G = 10$. Each subplot (i, j) is of size $G \times G$, showing the cosine similarity between $T_{i,j}$ and every element in T. Visualizations for different G are available in Appendix B.3.

To further verify that the model captures different information at different scales, we visualize the attention weights on each image in the multi-scale representation as Figure 4. We observe that the model tends to focus on more detailed areas on full-size high-resolution images and on more global areas on the resized ones. This shows that the model learns to capture image quality at different granularities.

Effectiveness of Proposed Hash-based Spatial Embedding (HSE) and Scale Embedding (SCE). We run ablations on different ways to encode spatial information and scale information using positional embeddings. As shown in Table 7, there is a large gap between adding and not adding spatial embeddings. This aligns with the finding in [10] that spatial embedding is crucial for injecting 2D image structure. To further verify the effectiveness of HSE, we try to add a fixed length spatial embedding as ViT [10]. This is done by treating all input tokens as a fixed length sequence and assigning a learnable embedding for each position. The performance of this method is unsatisfactory compared to HSE because of two reasons: 1) the inputs are of different aspect ratios. So each patch in the sequence can come from a different location from the image. Fixed positional embedding fails to capture this change; 2) since each position is modeled independently, there is no cross-scale information, meaning that the model cannot locate spatially close patches from different scales in the multi-scale representation. Moreover, the method is inflexible because fixed length spatial embedding cannot be easily applied to the large images with more patches. On the contrary, HSE is meaningful under all conditions.

A visualization of the learned HSE cosine similarity is provided as Figure 5. As depicted, the HSE of spatially close locations are more similar (yellow color) and it corresponds well to the 2D structure. For example, the bottom

Spatial Embedding	SRCC	PLCC
w/o	0.704	0.716
Fixed-length (no HSE)	0.707	0.722
HSE	**0.726**	**0.738**

Table 7. Ablation study results for spatial embeddings on AVA. For "Fixed length (not HSE)", we consider the input as a fixed-length sequence and assign a learnable embedding for each position.

Scale Embedding	SRCC	PLCC
w/o	0.717	0.729
w/	**0.726**	**0.738**

Table 8. Ablation study results for scale embedding on AVA.

	# Params	SRCC	PLCC
Linear projection	22M	0.634	0.714
Simple Conv	23M	0.639	0.726
5-layer ResNet	27M	**0.646**	**0.739**

Table 9. Comparison of different patch encoding modules on PaQ-2-PiQ full-size test set. For simple conv, we use the root of ResNet (a 7x7 conv followed by a 3x3 conv). For 5-layer ResNet, we stack a residual block on top of Simple Conv.

HSEs are brightest at the bottom. This shows that HSE can effectively capture the 2D structure of the image.

In Table 8, we show that adding SCE can further improve performance when compared with not adding SCE. This shows that SCE is helpful for the model to capture scale information independently of the spatial information.

Choice of Patch Encoding Module. We tried different designs for encoding the patch, including linear projection as [10] and small numbers of convolutional layers. As shown in Table 9, using a simple convolution based patch encoding module can boost the performance. Adding more conv layers has diminishing returns and we find a 5-layer ResNet can provide satisfactory representation for the patch.

5. Conclusion

We propose a multi-scale image quality Transformer (MUSIQ), which can handle full-size image input with varying resolutions and aspect ratios. By transforming the input image to a multi-scale representation with both global and local views, the model is able to capture the image quality at different granularities. To encode positional information in the multi-scale representation, we propose a hash-based 2D spatial embedding and a scale embedding strategy. Although MUSIQ is designed for IQA, it can be applied to other scenarios where task labels are sensitive to the image resolutions and aspect ratios. Moreover, MUSIQ is compatible with any type of Transformers that accept input as a sequence of tokens. Experiments on the four large-scale IQA datasets show that MUSIQ can consistently achieve state-of-the-art performance, demonstrating the effectiveness of the proposed method.

References

[1] Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. Pyramid methods in image processing. *RCA engineer*, 29(6):33–41, 1984. 2

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3286–3295, 2019. 2, 3

[3] 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. 6

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1, 2

[5] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. *arXiv preprint arXiv:2012.00364*, 2020.

[6] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020. 1, 2

[7] Qiuyu Chen, Wei Zhang, Ning Zhou, Peng Lei, Yi Xu, Yu Zheng, and Jianping Fan. Adaptive fractional dilated convolution network for image aesthetics assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14114–14123, 2020. 1, 2, 6, 7

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 5

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. 2, 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`. 1, 2, 3, 4, 7, 8

[11] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020. 1, 2, 5, 6

[12] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR, 2017. 3

[13] Deepti Ghadiyaram and Alan C Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of Vision*, 17(1):32–32, 2017. 2, 6

[14] 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. 4

[15] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9375–9383, 2019. 1, 2, 6

[16] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 1, 2, 5

[17] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014. 6

[18] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):206–220, 2016. 6

[19] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, pages 662–679. Springer, 2016. 6

[20] Dingquan Li, Tingting Jiang, Weisi Lin, and Ming Jiang. Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Transactions on Multimedia*, 21(5): 1221–1234, 2018. 6

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 2

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`. 2

[23] Shuang Ma, Jing Liu, and Chang Wen Chen. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4535–4544, 2017. 2, 6

[24] Long Mai, Hailin Jin, and Feng Liu. Composition-preserving

deep photo aesthetics assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 497–506, 2016. 1, 2, 6

[25] 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. 2, 6

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

[27] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20 (12):3350–3364, 2011. 2, 6

[28] Naila Murray and Albert Gordo. A deep architecture for unified aesthetic prediction. *arXiv preprint arXiv:1708.04890*, 2017. 6

[29] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012. 2, 5

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. 4

[31] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 3

[32] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 879–886, 2018. 2, 6

[33] 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*, pages 3667–3676, 2020. 1, 2, 5, 6

[34] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8): 3998–4011, 2018. 1, 2, 5, 6, 7

[35] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 5

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1, 2, 4, 5

[37] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, 25(9):4444–4457, 2016. 6

[38] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 995–1002, 2013. 2, 6

[39] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5754–5764. Curran Associates, Inc., 2019. 2

[40] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1098–1105. IEEE, 2012. 2, 6

[41] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3585, 2020. 1, 2, 5, 6

[42] Hui Zeng, Lei Zhang, and Alan C Bovik. A probabilistic quality representation approach to deep blind image quality prediction. *arXiv preprint arXiv:1708.08190*, 2017. 6

[43] Hui Zeng, Zisheng Cao, Lei Zhang, and Alan C Bovik. A unified probabilistic formulation of image aesthetic assessment. *IEEE Transactions on Image Processing*, 29:1548–1561, 2019. 6

[44] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 5

[45] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 2, 6

[46] 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. 2

[47] 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*, 30(1):36–47, 2018. 1, 2, 6

[48] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Metaiqa: deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14143–14152, 2020. 5, 6