VTAMIQ: Transformers for Attention Modulated Image Quality Assessment

Andrei Chubarau James Clark
Centre for Intelligent Machines, McGill University
Montreal, Canada
andrei.chubarau@mail.mcgill.ca clark@cim.mcgill.ca

Abstract

Following the major successes of self-attention and Transformers for image analysis, we investigate the use of such attention mechanisms in the context of Image Quality Assessment (IQA) and propose a novel full-reference IQA method, Vision Transformer for Attention Modulated Image Quality (VTAMIQ). Our method achieves competitive or state-of-the-art performance on the existing IQA datasets and significantly outperforms previous metrics in cross-database evaluations. Most patch-wise IQA methods treat each patch independently; this partially discards global information and limits the ability to model long-distance interactions. We avoid this problem altogether by employing a transformer to encode a sequence of patches as a single global representation, which by design considers interdependencies between patches. We rely on various attention mechanisms – first with self-attention within the Transformer, and second with channel attention within our difference modulation network – specifically to reveal and enhance the more salient features throughout our architecture. With large-scale pre-training for both classification and IQA tasks, VTAMIQ generalizes well to unseen sets of images and distortions, further demonstrating the strength of transformer-based networks for vision modelling.

1. Introduction

As consumers are more and more accustomed to high fidelity imagery and video, image quality has become a critical evaluation metric in many image processing and computer vision applications. Directly asking human observers to rate visual quality requires slow and expensive subjective experiments. As a practical alternative, objective Image Quality Assessment (IQA) methods automate this process by mimicking the behaviour of the Human Visual System (HVS) for image quality evaluation.

Given the highly subjective nature of human perception of image quality [69, 37, 47], naive computational IQA methods have limited success [68]; more perceptually accurate full-reference (FR) image quality metrics (IQMs) require modeling key characteristics of the HVS, for instance its sensitivity to various visual stimuli based on luminance, contrast, chrominance, frequency content, etc. [4, 33, 62], as well as natural image statistics that describe image patterns and structures [59, 57, 14]. That being said, conventional IQA methods are often constrained by their respective assumptions about the HVS and are prone to underperform for practical IQA tasks [67].

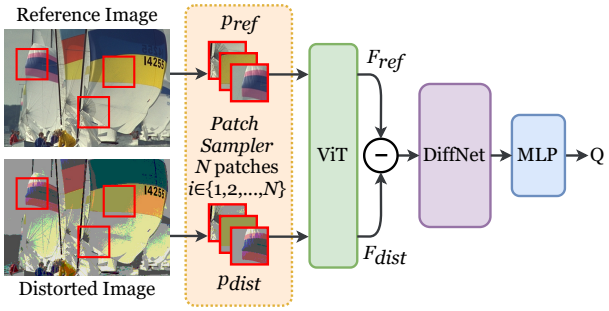

Figure 1. Simplified diagram of our proposed method, VTAMIQ. We apply context-aware probability sampling to extract sequences of N spatially aligned fixed-size patches from the reference and the distorted images, and encode these using a Vision Transformer (ViT) [13] modified for sparse patch extraction. Our Difference Modulation Network (DiffNet) then recalibrates the resulting difference signal between the two encoded representations using a stack of residual channel-attention layers. Finally, a conventional fully-connected Multilayer Perceptron (MLP) regresses the modulated difference into a single image quality prediction.

Following the major advances of deep Convolutional Neural Networks (CNNs) in many areas of computer vision research [32], the field of IQA has progressively gravitated towards deep learning [25, 27, 72, 8, 20, 36]. Jointly optimizing feature representations and inference directly from raw image data addresses the common limitations of classical approaches that rely on hand-crafted mechanisms. Granted, the available IQA datasets are typically modest in size and have limited image variety; this makes training deep IQMs from scratch a challenging task [34, 35, 49].

To facilitate training with limited data, many recent deep IQMs treat images as a combination of smaller-

sized patches. Instead of tackling the full-resolution inputs directly, patch-wise metrics are independently computed and then aggregated for the final image quality prediction [25, 27, 7, 49, 52]. An estimate of patch importance or attention – since attention essentially modulates the visibility of distortions [74, 18] – can be used for weighted averaging of the patch-wise scores. Most existing methods, however, do not focus on complex attention models and simply process each patch independently. Patch scores and weights are estimated without complete global knowledge; the ability to account for large multi-scale features and long-distance interactions in the full-resolution input is thus limited.

Furthermore, arbitrarily defined signals in neural networks require specialized attention mechanisms. Convolutional operators only exploit local information within a predefined receptive field; their ability to model spatial and channel-wise dependencies is limited. Various attention strategies were thus introduced to complement conventional architectures with attention [17, 60, 66]. For instance, channel attention (CA) [23, 77] and self-attention (SA) [65] explicitly evaluate the attention given to feature channels or elements of a sequence, respectively.

While CNNs dominate in computer vision, SA-based Transformers [65, 26] have taken over natural language processing (NLP) due to their excellent ability to model sequences and long-distance dependencies, which the previous methods often lack [3, 43, 9]. With enough data, appropriate pre-training, and clever architectural modifications, Transformers attain state-of-the-art results in vision tasks [50, 13] illustrating the flexibility and effectiveness of SA.

Our key observation is the similarity between patch-wise IQA and transformers-based architectures: the input is a sequence of tokens or image patches. We further investigate this analogy along with various attention mechanisms in the context of IQA and propose a novel full-reference IQA method, Vision Transformer for Attention Modulated Image Quality (VTAMIQ). Our approach builds upon the success of Vision Transformer (ViT) [13] for vision tasks and leverages the ability of self-attention to jointly encode a sequence of image patches, by design considering patch interdependencies unlike previous patch-wise IQA methods. Employing a transformer allows VTAMIQ to avoid computing independent patch-wise quality metrics, and instead to directly determine a single latent representation for each input image, followed by a single global image quality score. A simplified diagram of our method is depicted in Figure 1.

Furthermore, previous deep FR IQA methods can be typically broken down into several key steps: i) compute encoded representations (features) for the input images or patches, ii) compute difference between the encoded features, iii) correlate feature difference with image quality. Such approaches assume that the deep feature difference is immediately informative and can be directly correlated with image quality. We further propose an additional step: to intelligently modulate the difference between the two inputs such that more meaningful information is enhanced before regressing to a single quality score. We accomplish this by utilizing a specialized channel attention-based network inspired by the work in [77].

Our method, VTAMIQ, achieves competitive or state-of-the-art IQA performance on the popular IQA datasets LIVE [56], CSIQ [31], TID2013 [47], and KADID-10k [34]. We further emphasize the ability of our algorithm to generalize to unseen sets of images and distortions, demonstrated by VTAMIQ's vastly superior performance evaluated in cross-database tests. To further encourage the use of our metric as well as to facilitate reproducible research, our implementation and supplementary material will be made publicly available at https://github.com/ch-andrei/VTAMIQ.

2. Related Work

A brief overview of IQA methods and relevant attention mechanisms is provided. We focus primarily on FR IQA as our method specifically falls under this category.

Conventional IQA. Classical FR IQMs typically correlate image quality with some hand-crafted definition of perceptual difference between the inputs. The comparison can be based on error visibility [10, 19], structural similarity [69, 71, 70], information content [54, 53], contrast visibility [41, 42], or various other HVS-inspired feature similarities [75, 73, 44, 51, 46]. Combining IQA with visual saliency, i.e. models of human attention [30], further benefits prediction accuracy by targeting the more salient regions of the observed content [74].

Deep Learning for IQA. To address the limitations of classical IQA approaches, data-driven deep learning methods determine a model of image quality that does not rely on hand-crafted definitions. While many of the existing IQA datasets are limited in size [56, 31, 48, 47], deep learning for IQA has become more and more feasible [34, 35]. Moreover, with the major successes of CNNs in many areas of computer vision research, exploiting commonalities between different tasks is possible with transfer learning [6, 61]. For instance, as FR IQA can be performed by comparing deep feature responses for two images [1, 76, 63], it is possible to pre-train the underlying networks on large and widely available classification data [11, 28].

To further counter the lack of large IQA datasets, instead of directly training on full-resolution images, a common data augmentation approach is to explicitly subdivide the inputs into fixed-sized pixel patches. Patch-wise quality scores are then computed and aggregated into a single image quality prediction either by simple averaging [25] or by a weighted pooling scheme using a form of visual sensitivity map [27]. For instance, WaDIQaM [7] and PieAPP [49] directly estimate patch weights along with the patch

quality scores. As an improvement, JND-Salcar [52] further incorporates visual salience and just-noticeable-differences (JND) information for computing feature representations as well as for guiding patch-wise weight prediction.

Learning to rank – and not directly score – images based on quality is another possible interpretation of IQA. In [38], a network is trained to rank synthetically distorted images; the level of distortion is known, hence a ranking can be established. Ranking images can be further interpreted as the probability of preference of one image over another. Unlike in direct IQA measures, image quality is indirectly deduced from pairwise preference experiments. For instance, the work in [49] introduces a perceptual preference dataset PieAPP [40] along with the PieAPP metric for IQA, a pairwise-learning framework for training an error-estimating function using preference probabilities. Perceptual error scores are computed for images A and B, and the associated probability of preference is predicted; the trained error function can then be extracted as a baseline IQA model and further fine-tuned on the common IQA datasets.

Channel Attention. In the context of deep feature responses with multiple channels (kernels), Squeeze-and-Excitation (SE) modules [23, 22] explicitly model interdependencies between channels to adaptively rescale the more salient channel-wise feature responses. Existing network architectures, for instance variants of VGG [58] and ResNet [17], can be enhanced with SE modules to outperform their original baselines. In [77], SE is directly termed Channel Attention (CA); stacked CA modules are employed to build a deep residual network for super-resolution that specifically focuses on learning high-frequency information and significantly surpasses all previous approaches.

Self-Attention. A general attention function computes the relative importance of signals within a sequence. In self-attention (SA), the importance of each signal in the input is specifically estimated relative to the input sequence itself to compute its encoded representation. Conventional convolutional layers, by design, have a finite local receptive field; on the other hand, SA considers the entire input sequence and thus has a much stronger ability to model long-range interactions. While SA was first explored to favorably complement conventional CNNs [5, 3], it was later effectively applied in standalone applications averting the need for recurrence and convolutions [65, 12].

Transformers. Whereas prior methods apply recurrent neural networks for sequence modelling [3, 43, 9], Transformers utilize an encoder-decoder architecture fully relying on self-attention to compute latent representations without the need for recurrent mechanisms or convolutions. Transformers scale well for large datasets and complex models, and can be easily parallelized, making up for their natural applicability for a wide variety of tasks [26].

Vision Transformer (ViT). Inspired by the successes of Transformer-based architectures in NLP, the work in [13] applies a standard Transformer "with the fewest possible modifications" to vision tasks. Unlike NLP with word sequences, ViT operates on a sequence of flattened fixed-sized pixel patches extracted by tiling the input image. Although ViT only yields modest performance when trained directly on ImageNet [11], with additional pre-training on large amounts of data and increased model complexity, ViT attains state-of-the-art classification performance at a fraction of training computational cost when compared to competitive methods.

3. VTAMIQ: Vision Transformer for IQA

As illustrated in Figure 1, our proposed FR IQA method, VTAMIQ, utilizes a modified Vision Transformer (ViT) to encode each input image as a single latent representation, computes the difference between encoded representations for the reference and the distorted images, intelligently modulates this difference, and finally, interprets the modulated difference as an image quality score.

While many previous deep FR IQA algorithms independently compute and aggregate patch-wise quality metrics, we directly avoid this by employing a Transformer to encode each input image as a single global representation. Although we still extract a sequence of spatially aligned patches to initially describe the reference and the distorted images, our model processes the input patches collectively and considers interdependencies between patches. More specifically, this is achieved with ViT computing self-attention with respect to the complete sequence of input patches. In this way, global information and long-distance interactions from the full-resolution image are taken into account and thereafter implicitly employed in the context of image quality assessment.

The goal of our difference modulation network is then to enhance the more salient features of the difference between the encoded representations. Most deep IQMs directly use fully-connected layers to regress the difference vector; this assumes that the feature difference can be immediately correlated with image quality. Our results emphasize the importance of modulating the difference vector before applying regression: Channel Attention [77] (also known as Squeeze-and-Excitation [23, 22]) perfectly satisfies the requirements for this task.

3.1. Feature Extraction

We use a Vision Transformer modified for sparse unordered patch input sequences to encode each input image as a vector of dimension D corresponding to the hidden size of the transformer. As in ViT, the input image $I \in \mathbb{R}^{H \times W \times 3}$ is represented by N fixed-sized RGB color patches $p_i \in \mathbb{R}^{N \times p \times p \times 3}$, where p is the patch dimension.

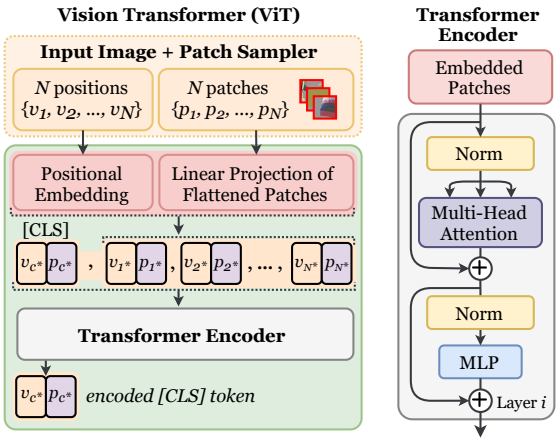

Figure 2. Vision Transformer (ViT) modified for VTAMIQ. Unlike the original ViT implementation where the raster order of tiled patches is relevant, we use patch center coordinates v to index into the positional embeddings. After transforming the sequence, the encoded "classification token" `[CLS]` acts as the global representation of the input sequence. The architecture of the Transformer Encoder is shown on the right (figure adapted from [13, 65]).

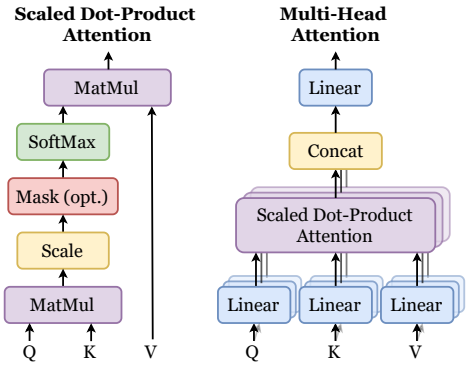

Figure 3. Scaled Dot-Product Attention (left) and Multi-Head Attention (right) used in Vision Transformer. Adapted from [65].

Patch embeddings are computed by applying a trainable linear projection to map from patch dimension to the latent Transformer dimension. Learned positional embeddings are then injected into the patch embeddings to account for positional information. The overall structure and components of ViT are illustrated in Figures 2-3.

Although we reuse the overall structure of positional embeddings as in ViT, we employ a different strategy to assign positional embeddings to each patch. Instead of tiling the input image as in ViT, we use random sampling to extract N patches; the resulting sequence is thus unordered and may contain overlapping patches. To address this, we use the positional uv coordinate of each patch (available at the time of sampling), to index into the array of positional embeddings. This little modification complements the original ViT architecture allowing us to fine-tune a pre-trained ViT.

3.2. Patch Sampling Strategies

The simplest method to extract pixel patches is to tile an image. In the ideal scenario, we want to completely cover the full-resolution input, but this is often impractical due to memory and computational constraints associated with operating deep neural networks. Similarly to previous work, we instead opt to randomly sample a potentially sparse set of patches from the input image. This provides the hidden benefit of data augmentation which favors model training and reduces overfitting: each time an image is processed, a new and unseen random set of patches is extracted.

Previous patch-wise IQA methods employ naive tiling or random sampling to draw patch samples uniformly across the image; we improve on this by employing a special context-aware patch sampling strategy to extract more informative patches. We leverage key properties of human visual perception as well as the nature of full-reference IQA in our sampling scheme. First, to account for the centerbias of human visual attention [30, 29], we sample the more salient central regions with higher probability. Secondly, given the observation that extracting patches that are similar between the reference and the distorted images adds little information about image quality distortions, we bias the sampling towards regions with higher perceptual difference between the reference and the distorted images. An estimate of perceptual difference can be computed by simple metrics such as MSE or SSIM [69] with similar results.

Altogether, our context-aware patch sampling (CAPS) strategy allow us to use fewer patch samples for the same level of accuracy, effectively reducing memory and computational requirements. CAPS requires minimal overhead and speeds up model training, e.g. VTAMIQ can be trained about 15% faster (fewer epochs) with CAPS when training with 256 patch samples per image. Note that the benefit of CAPS becomes less apparent when using larger patch counts as sampling variance is then naturally reduced.

3.3. Deep Feature Difference Modulation

We use Residual Channel Attention Blocks (RCABs) and Residual Groups (RG) proposed in [77] to modulate the difference between the encoded representations of the reference and the distorted images. Our Difference Modulation Network – DiffNet for simplicity – consists of a stack of N RGs. Similarly, each RG consists of M stacked RCABs, with the additional residual skip connection. Each RCAB further contains a skip connection linking the output of multiplicative Channel Attention (CA) and the input. The overall structures of CA, RCAB, and RG are depicted in Figure 4. The underlying mathematical computations for DiffNet are defined in Equations 1-4, where $Attention$ corresponds to the learned "excitation" stage of CA responsible for computing channel weights [23, 77], and U_{RG} and U_{RCAB} are learnable linear transformations.

$$CA(x) = x \times Attention(x) \tag{1}$$

$$RCAB(x) = x + U_{RCAB}(x) \times CA(x) \tag{2}$$

$$RG(x) = x + U_{RG}(RCAB_N(...(RCAB_1(x))) \tag{3}$$

$$DiffNet(x) = RG_N(...(RG_1(x))) \tag{4}$$

Previously in IQA, CA modules were used for feature extraction [52] but never for modulating the difference vector. Moreover, in [77], a construct similar to our DiffNet – referred to as Residual in Residual (RIR) – is shown to contribute to the extraction of more informative high-frequency features. Unlike the original RIR, we do not include the final residual skip connection in our variant to emphasize its signal rescaling properties. Lastly, unlike the original implementation with 10 RGs and a total of 200 RCABS [77], our best IQA prediction accuracy was achieved with a shallower network with only 4 RGs each with 4 RCABs.

3.4. Model Training

Mean Absolute Error (MAE) and Mean Squared Error (MSE) are the commonly used loss functions for training IQA models [25, 27, 7, 52]. Both functions compute the difference between the expected and the predicted values; due to the additional squaring operation, MSE penalizes larger differences more harshly than MAE. We evaluated training our model with MAE and MSE and found that more consistent results were obtained with MAE loss.

In addition to the above, we also found that optimizing with ranking loss [38, 76, 52] benefits our model. Knowing that images can be ranked based on their quality – e.g. image A has higher quality than image B – ranking loss explicitly penalizes predictions that do not agree with the expected ordering. Ranking loss thus complements the more general MAE loss by offering a different guidance signal; it does not directly expose the magnitude of the expected predictions but specifically their relative ranking, which is arguably equally meaningfull in IQA. For a pair of image quality predictions, pairwise ranking loss can be computed as per Equation 5, where $y_1, y_2, \hat{y}_1, \hat{y}_2$ correspond to the two predicted and the two expected scores, respectively, and ϵ is a small stabilizing constant. When the predicted ranking of the two images agrees with the expected result, the numerator value is negative, and hence the result is clamped to zero by the max operator; when the ranking does not agree, ranking loss corresponds to the absolute difference between the two scores, thus pointing towards the correct ordering.

$$L_{rank}(y_1, y_2, \hat{y}_1, \hat{y}_2) = max\left(0, \frac{-(\hat{y}_1 - \hat{y}_2) \times (y_1 - y_2)}{|\hat{y}_1 - \hat{y}_2| + \epsilon}\right) \tag{5}$$

Naturally, in the context of training with mini-batches of size N, pairwise ranking loss can be computed for $\binom{N}{2}$ possible pairwise combinations of the predictions; the total ranking loss then corresponds to the sum or the mean of the pairwise values. We experimented with combining MSE and ranking losses using different strategies and found that simple summation led to the optimal results.

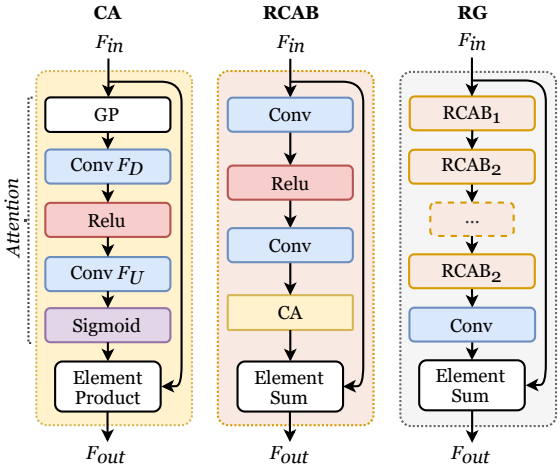

Figure 4. Components of VTAMIQ's Difference Modulation Network (DiffNet). Adapted from [77].

4. Experiments and Results

4.1. Datasets

We assess the performance of our method using the common major IQA datasets: LIVE [56], CSIQ [31], TID2013 [47], and KADID-10k [34]. PieAPP [49] and KADIS-700k [35] datasets are used for pre-training our model. Table 1 compares the listed datasets in more detail. Note that we omit the BAPPS dataset [76] as it targets perceptual patch similarity and JND as opposed to image quality.

As recommended in [34, 16], we randomly split the datasets into training (∼60%), validation (∼20%), and test sets (∼20%) along the reference image dimension. This ensures that the test and validation images are not seen by the network during training; validation set is then used for picking the best performing model, and the test set to evaluate the final performance. For LIVE, reference images are split into 17 training, 6 validation, and 6 test images; analogously for CSIQ, TID2013 and KADID-10k, the datasets are randomly split into 18-6-6, 15-5-5, and 49-16-16 subsets, respectively. PieAPP and KADIS-700k datasets are divided onto 160-40 and 120k-20k train-validation splits, respectively, and only used for pre-training our model.

Table 1. Comparison of the available IQA datasets.

Dataset	Ref. images	Distortions	Dist. images
LIVE [56]	29	5	779
CSIQ [31]	30	6	866
TID2013 [47]	25	24	3,000
KADID-10k [34]	81	25	10125
PieAPP [49]	200	75	3,000
KADIS-700k [35]	140,000	25*	700,000

4.2. Pre-training a Baseline Model

As reported in [13], Vision Transformer strongly benefits from pre-training on larger datasets prior to transfer to down-stream tasks. To circumvent the lengthy requirements for training a Transformer from scratch, for all our experiments, we use ViT initialized with weights pre-trained on ImageNet and ImageNet-21k [11] classification databases[1].

Our further pre-training for IQA consists of several steps. First, we pre-train VTAMIQ on the large KADIS-700k [35] dataset, which contains 700,000 image pairs along with the corresponding scores given by 11 conventional IQMs as "weak supervision" signal for IQA. Out of the 11 available metrics, we select VSI [74] scores as the prediction target. Unlike the conventional IQA datasets with limited number of reference images, KADIS-700k contains a wide variety of image data which benefits VTAMIQ's ability to generalize. Secondly, the resulting pre-trained model is fine-tuned on the subjective perceptual preference data from the PieAPP dataset [49] using the pairwise training framework proposed by the same authors. We find that this second pre-training stage serves as a further improvement for our baseline. We train for five epochs on each dataset and do not leave out a test set as no evaluations are performed. Note that as recommended in [35], we apply histogram equalization on the scores prior to training.

4.3. Experimental Setup

Since we use ViT models pre-trained on ImageNet [11] images, we normalize all input images with ImageNet's mean and variance from their original range. At training time, we extract 256 spatially aligned patches from the reference and the distorted images; for validation and testing, we use 1024 patches per input image to decrease the sparsity of sampling and improve the overall accuracy of image quality predictions. Patch sampling is redone each time an image is seen, thus even though the same source images are used, the extracted patches are different. We train with batch size set to 20 images, hence the pairwise rank loss is computed for 20 images in the batch. For evaluations, we run 20 epochs of training, where an epoch consists of seeing each image in the dataset once. For optimization, we use the AdamW optimizer [39] with the recommended parameters and an initial learning rate (LR) of 10^{-5}, decayed by a factor of 10 at epoch 12. Note that pre-training is done for 5 epochs: 3 with LR of 10^{-4}, followed by 2 more at 10^{-5} LR. VTAMIQ converges to the optimal solution within the scheduled period as defined above.

We implemented our proposed method VTAMIQ in Pytorch [45] and trained using a single NVIDIA GeForce RTX 3080 GPU with 10GB video memory. The practical training runtimes differ across datasets as the number of images in each dataset is different. On our setup and with the configurations defined above, training 20 epochs on KADID-10k dataset requires one hour, whereas a single epoch on KADIS-700k lasts 4 hours.

[1]We did not implement or train ViT but re-used a publicly available pre-trained model from https://github.com/jeonsworld/ViT-pytorch [24]

4.4. Model Configurations

While exploring model design, we found that we may discard several top-most layers and only use the stem of ViT with little reduction to performance. This strategy is relatively common in transfer learning [61]; specifically for layered architectures, earlier layers encode the general structure of images, whereas deeper layers typically extract more task-dependent information. Furthermore, we tested ViT with 32×32 and 16×16 patch sizes and determined that the latter exhibits superior IQA performance and generalization. As such, our best results are reported for VTAMIQ using a variant of ViT with 16×16 patch size and the "Base" configuration (see [13]), keeping 6 out of the 12 available layers. We configure our DiffNet with 4 RGs each with 4 RCABS, and refer to this variant as VTAMIQ-16-6-4-4, a model with 57.3M parameters (43.6M from ViT).

4.5. Performance Evaluation and Comparison

Performance of our model is assessed with the commonly used Pearson linear correlation coefficient (PLCC), Spearman rank order coefficient (SRCC), and Kendall rank-order correlation coefficient (KROCC). PLCC assesses the linear correlation between the expected and the predicted quality scores, whereas SROCC and KROCC describe the level of monotonic correlation between the two trends. The results are presented in Table 2 where we report the mean of 20 evaluation runs. As with most recent IQA methods, we follow the suggestion in [55] and apply non-linear fitting to our model's predictions prior to computing PLCC.

Comparison Against State-of-the-Art. For all our tests, we fine-tune a baseline VTAMIQ pre-trained on KADIS-700k and PieAPP datasets. Our experiments demonstrate that VTAMIQ outperforms or is competitive with WaDIQaM [7], PieAPP [49], and JND-SalCAR [52] for all tested datasets. Especially on the larger KADID-10k database, we observe a solid improvement over previous work. Since VTAMIQ is built with ViT, it is a data-hungry Transformer-based model that improves with more training data. Besides, even conventional IQMs perform well on the smaller LIVE and CSIQ, but fall short on the more complex datasets such as TID2013 and KADID-10k. Furthermore, unlike JND-SalCAR [52], VTAMIQ does not require JND probability and salience maps and is directly applicable to the simple FR IQA problem involving a pair of images, without any additional inputs. As such, we both improve on the state-of-the-art, as well as reduce the input requirements of the previously best method.

Cross-Database Performance Evaluation. Due to limited

Table 2. Performance comparison on LIVE, CSIQ, TID2013, and KADID-10k Databases. Performance scores of other methods are as reported in the corresponding original papers. Best scores are **bolded**, second best are underlined, missing scores are shown as "–" dash.

Method	LIVE[56]			CSIQ[31]			TID2013[47]			KADID-10k[34]		
	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC
SSIM[69]	0.942	0.948	0.812	0.861	0.876	0.691	0.691	0.637	0.464	0.723	0.724	–
MS–SSIM[71]	0.946	0.956	0.829	0.899	0.913	0.739	0.833	0.786	0.608	0.801	0.802	–
FSIMc[75]	0.953	0.963	0.846	0.919	0.931	0.769	0.877	0.851	0.667	0.851	0.854	–
MDSI[44]	0.966	0.967	0.840	0.953	0.957	0.813	0.909	0.890	0.712	0.873	0.872	–
VSI[74]	0.946	0.948	0.807	0.928	0.942	0.786	0.900	0.897	0.718	0.878	0.879	–
SCQI[2]	0.937	0.948	0.810	0.927	0.943	0.787	0.907	0.905	0.733	0.853	0.854	–
DOG–SSIMc[46]	0.966	0.963	0.844	0.943	0.954	0.813	0.934	0.926	0.768	–	–	–
DeepFL–IQA[34]	0.978	0.972	–	0.946	0.930	–	0.876	0.858	–	0.938	0.936	–
DeepQA[27]	0.982	0.981	–	0.965	0.961	–	0.947	0.939	–	–	–	–
DualCNN[64]	–	–	–	–	–	–	0.924	0.926	0.761	0.949	0.941	0.802
WaDIQaM–FR[7]	0.980	0.970	–	–	–	–	0.946	0.940	0.780	–	–	–
PieAPP[49]	0.986	0.977	0.894	0.975	0.973	**0.881**	0.946	0.945	0.804	–	–	–
JND–SalCAR[52]	**0.987**	**0.984**	**0.899**	0.977	0.976	0.868	**0.956**	0.949	**0.812**	0.960	0.959	–
VTAMIQ (ours)	0.984	0.982	0.891	**0.982**	**0.979**	0.875	0.954	**0.950**	0.809	**0.964**	**0.963**	**0.838**

Table 3. Performance comparison (SROCC) for cross-database evaluations.

Trained on:	**LIVE**		**TID2013**		**KADID-10k**	
Tested on:	**TID2013**	**KADID-10k**	**LIVE**	**KADID-10k**	**LIVE**	**TID2013**
DOG-SSIMc [46]	0.751	-	0.948	-	-	-
WaDIQaM-FR [7]	0.751	-	0.936	-	-	-
DualCNN [64]	-	-	-	-	0.751	0.729
DeepFL-IQA [35]	0.577	0.689	0.780	0.698	0.894	0.71
VTAMIQ (ours)	**0.874**	**0.899**	**0.957**	**0.907**	**0.974**	**0.897**

training data as well as the high dimensionality of the input space, deep learning-based IQA methods tend to overfit on the datasets they are trained with, motivating the use of cross-database tests. We evaluate the cross-database performance of our method and demonstrate that VTAMIQ generalizes to unseen images and distortions more effectively than most previous methods. We compare VTAMIQ's cross-database performance to several prominent IQMs in Table 3, clearly demonstrating the superior performance of our approach. Analogous improvement can be seen for all possible combinations of database cross-testing, with at least 10% higher SROCC achieved by VTAMIQ.

Full cross-database performance results of VTAMIQ are reported in Table 4. Immediately after our pre-training stage (see Section 4.2), VTAMIQ already outperforms most conventional metrics. Furthermore, we observe that VTAMIQ does not "forget" its pre-training; its performance on unseen content only improves with more training data. For instance, a model pre-trained on KADIS-700k and LIVE performs better on the unseen TID2013 than a model pre-trained on KADIS-700k alone. What is more, fine-tuning on the larger KADID-10k dataset offers more improvement than the smaller LIVE (despite LIVE and TID2013 partially sharing reference images). With that in mind, we foresee that the performance of our method can be further improved as more datasets are included in the training stage. Currently, we only train on FR IQA datasets; large NR IQA datasets (e.g. [21, 15]), can be leveraged in the future.

Performance on Unseen Distortions. Recently with deep learning, some practical applications for IQA have shifted towards evaluating image restoration algorithms, e.g. deblurring, super-resolution, etc.; a lineup of "novel" generative distortions are now relevant. PieAPP dataset specifically contains a variety of such novel distortions and, unlike other IQA datasets, PieAPP's predefined test set purposefully holds out a disjoint set of distortions unseen during training. In Table 5, we assess VTAMIQ on the PieAPP

Table 4. Cross-database performance (SROCC) evaluation for VTAMIQ. Baseline is pre-trained on KADIS-700k and PieAPP.

Trained on	**Tested on**			
	LIVE	**CSIQ**	**TID2013**	**KADID-10k**
Baseline	0.949	0.936	0.891	0.895
LIVE	**0.982**	0.955	0.885	0.899
CSIQ	0.965	**0.979**	0.893	0.905
TID2013	0.957	0.947	**0.950**	0.907
KADID-10k	0.974	0.967	0.897	**0.963**

Table 5. Performance (SROCC) comparison for models trained on PieAPP training set using the pairwise learning method from [49]. All evaluations use unseen test sets. *Results reported in [49].

Method	**Dataset**		
	PieAPP (test)	**CSIQ**	**TID2013**
DeepQA* [27]	0.632	0.873	0.837
WaDIQaM-FR* [7]	0.748	0.898	0.859
PieAPP* [49]	**0.831**	0.907	0.875
VTAMIQ (ours)	0.829	**0.945**	**0.908**

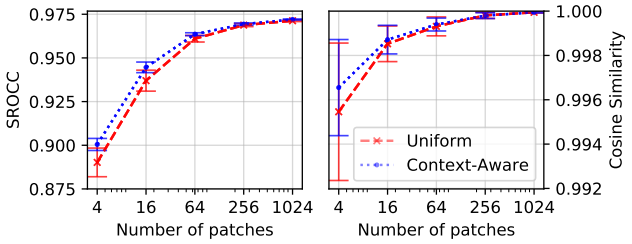

Figure 5. Influence of context-aware patch sampling and patch counts on VTAMIQ's performance on the LIVE dataset. On the left, the mean of SROCC for 10 runs is reported. On the right, we illustrate cosine similarity between the encoded representations for the same image acquired by resampling N patches from the same image 128 times (averaged across all images in the LIVE dataset). Error bars denote 95% confidence intervals.

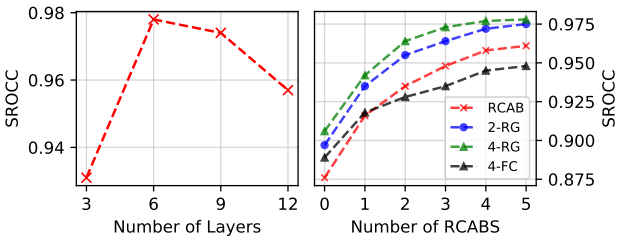

Figure 6. VTAMIQ's performance with different configurations for ViT and DiffNet as assessed on LIVE. On the left, we vary the number of retained layers in ViT. On the right, we evaluate DiffNet with varying number of Residual Groups (RG) and Residual Channel Attention Blocks (RCABs): i) no RGs and only RCABs (red), ii) two RGs (blue), and iii) four RGs (green), iv) four RGs with fully-connected layers instead of CA (black).

test set to directly compare against PieAPP IQA metric and evaluate the ability of our method to generalize to unseen distortions. The results show that VTAMIQ is essentially on par with the PieAPP IQA metric but significantly outperforms all other previous work. That being said, the very same VTAMIQ model largely surpasses PieAPP metric in cross-dataset tests on unseen CSIQ and TID2013.

4.6. Patch Sampling Strategies

We evaluate the effect of the number of patches and our context-aware patch sampling (CAPS) on VTAMIQ's performance and report the results in Figure 5 (using 16×16 patches and tested on the 512×384 images from LIVE dataset). Naturally, using more patches produces lower sampling variance and thus higher prediction accuracy. Especially for low sample counts, CAPS improves the consistency and the overall performance as shown by higher SROCC and smaller standard deviation values for the same number of extracted patches. That being said, using more than 512 patches results in increasingly minimal improvements. We further evaluate the cosine similarity between the encoded representations for images given randomized sequences of patches and similarly observe that CAPS improves the convergence of the encoded representations. As expected, the benefit of our sampling scheme becomes less apparent as the number of patches increases.

4.7. Ablation Study

We perform ablations for VTAMIQ to validate our architectural choices, namely varying the configurations used for ViT and our Difference Modulation Network. The results are summarized in Figure 6.

ViT Configurations. We investigate the influence of number of layers in ViT on VTAMIQ's performance and find that discarding several topmost layers results in very minimal performance loss. We determine that for the "Base" ViT configuration, keeping 6 layers in the Transformer is a good trade-off between model complexity and performance. Moreover, we observe that larger models are not as easily trained and often result in suboptimal IQA solutions despite having stronger performance for classification tasks.

Number of Residual Groups and RCABs. VTAMIQ's quality estimation strongly relies on the employed difference modulation network. We ran ablations on the number of Residual Groups (RG) and the number of RCABs per RG in the modulation network. Judging by the results of our evaluation, the improvement of using a specialized difference modulating network is quite substantial: we find that without our modulation network, the performance deteriorates to that of early deep learning IQMs. We further determine that four RRs each with four RCABs offers a good balance between performance and model complexity.

Use of Channel Attention. In order to directly validate the need for CA in our architecture, we replaced all CA modules with fully-connected layers (black plot in Figure 6) and observe a significantly decreased SROCC on LIVE. We also validated VTAMIQ's performance with various configurations of RGs and RCABs (CA is part of RCAB): when using no RGs and zero RCABs – and thus no CA – VTAMIQ's performance is severely reduced.

5. Conclusion

In this paper, we presented VTAMIQ, a Transformer-based difference modulation network for FR IQA that relies on various attention mechanisms to expose the more salient and informative features. We leverage a transformer to avoid the limitations of patch-wise IQA and instead operate with global feature representations. Our specialized difference modulation network then enhances the deep feature difference prior to predicting the final image quality score. With thorough pre-training, VTAMIQ outperforms the state-of-the-art on the common IQA datasets and robustly generalizes to unseen images and distortions with superior cross-database performance, further proving the effectiveness of attention mechanisms in vision modelling.

References

[1] Seyed Ali Amirshahi, Marius Pedersen, and Stella Yu. Image quality assessment by comparing cnn features between images. *Journal of Imaging Science and Technology*, 60:604101–6041010, 11 2016. 2

[2] Sung-Ho Bae and Munchurl Kim. A novel image quality assessment with globally and locally consilient visual quality perception. *IEEE Transactions on Image Processing*, 25(5):2392–2406, 2016. 7

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2, 3

[4] Peter G. J. Barten. Contrast sensitivity of the human eye and its effects on image quality. 1999. 1

[5] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks. *CoRR*, abs/1904.09925, 2019. 3

[6] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27*, UTLW'11, page 17–37. JMLR.org, 2011. 2

[7] 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, 2018. 2, 5, 6, 7

[8] Sebastian Bosse, D. Maniry, T. Wiegand, and W. Samek. A deep neural network for image quality assessment. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3773–3777, Sep. 2016. 1

[9] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. *CoRR*, abs/1601.06733, 2016. 2, 3

[10] Niranjan Damera-Venkata, T. D. Kite, Wilson S. Geisler, Brian L. Evans, and Alan C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4):636–650, April 2000. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2, 3, 6

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 3

[13] 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, 2020. 1, 2, 3, 4, 6

[14] Zhengfang Duanmu, Wentao Liu, Zhongling Wang, and Zhou Wang. Quantifying visual image quality: A bayesian view, 2021. 1

[15] Deepti Ghadiyaram and Alan C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2016. 7

[16] Franz Götz-Hahn, Vlad Hosu, and Dietmar Saupe. Critical analysis on the reproducibility of visual quality assessment using deep features, 2021. 5

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2, 3

[18] Lihuo He, Yanzhe Zhong, Wen Lu, and Xinbo Gao. A visual residual perception optimized network for blind image quality assessment. *IEEE Access*, 7:176087–176098, 2019. 2

[19] Damon M. Chandler; Sheila S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16(9):2284–2298, Sep. 2007. 2

[20] Peyman Milanfar Hossein Talebi. NIMA: neural image assessment. *CoRR*, abs/1709.05424, 2017. 1

[21] 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. 7

[22] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. *CoRR*, abs/1810.12348, 2018. 3

[23] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017. 2, 3, 4

[24] Eunkwang Jeon "Jeonsworld". Vit-pytorch. https://github.com/jeonsworld/ViT-pytorch. Accessed on: March 17, 2021. 6

[25] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014. 1, 2, 5

[26] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey, 2021. 2, 3

[27] Jongyoo Kim and Sanghoon. Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1969–1977, 2017. 1, 2, 5, 7

[28] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012. 2

[29] Matthias Kummerer, Thomas S. A. Wallis, and Matthias Bethge. Saliency benchmarking made easy separating models, maps and metrics. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 798–814. Springer International Publishing. 4

[30] Matthias Kummerer, Thomas S. A. Wallis, Leon A. Gatys, and Matthias Bethge. Understanding low- and high-level contributions to fixation prediction. In *Proceedings of the*

IEEE International Conference on Computer Vision (ICCV), Oct 2017. 2, 4

[31] Eric Larson and Damon Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electronic Imaging*, 19:011006, 01 2010. 2, 5, 7

[32] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015. 1

[33] Gordon E. Legge and John M. Foley. Contrast masking in human vision. *J. Opt. Soc. Am.*, 70(12):1458–1471, Dec 1980. 1

[34] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 1, 2, 5, 7

[35] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Deepfl-iqa: Weak supervision for deep iqa feature learning. *arXiv preprint arXiv:2001.08113*, 2020. 1, 2, 5, 6, 7

[36] Kwan-Yee Lin and Guanxiang Wang. Hallucinated-IQA: No-reference image quality assessment via adversarial learning. *CoRR*, abs/1804.01681, 2018. 1

[37] Weisi Lin and C.-C. Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297 – 312, 2011. 1

[38] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. *CoRR*, abs/1707.08347, 2017. 3, 5

[39] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. 6

[40] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo Exploration Database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, Feb. 2017. 3

[41] Rafal Mantiuk, Scott J. Daly, Karol Myszkowski, and Hans-Peter Seidel. Predicting visible differences in high dynamic range images: model and its calibration, 2005. 2

[42] Rafał Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4):40:1–40:14, July 2011. 2

[43] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *CoRR*, abs/1406.6247, 2014. 2, 3

[44] Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam, and Mohamed Cheriet. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *CoRR*, abs/1608.07433, 2016. 2, 7

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[46] Soo-Chang Pei and Li-Heng Chen. Image quality assessment using human visual dog model fused with random forest. *IEEE Transactions on Image Processing*, 24(11):3282–3292, 2015. 2, 7

[47] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015. 1, 2, 5, 7

[48] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. TID2008 - a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10:30–45, 01 2009. 2

[49] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. PieAPP: Perceptual image-error assessment through pairwise preference. *CoRR*, abs/1806.02067, 2018. 1, 2, 3, 5, 6, 7

[50] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models, 2019. 2

[51] Rafael Reisenhofer, Sebastian Bosse, Gitta Kutyniok, and Thomas Wiegand. A haar wavelet-based perceptual similarity index for image quality assessment. *CoRR*, abs/1607.06140, 2016. 2

[52] Soomin Seo, Sehwan Ki, and Munchurl Kim. A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2020. 2, 3, 5, 6, 7

[53] Hamid Rahim Sheikh and Alan C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, Feb 2006. 2

[54] Hamid Rahim Sheikh, Alan C. Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, Dec 2005. 2

[55] Hamid Rahim Sheikh, Muhammad Farooq Sabir, and Alan C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, Nov 2006. 6

[56] Hamid Rahim Sheikh, Zhou Wang, Lawrence Cormack, and Alan C. Bovik. LIVE image quality assessment database release 2. http://live.ece.utexas.edu/research/quality, 2005. 2, 5, 7

[57] Eero P. Simoncelli and Bruno A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001. PMID: 11520932. 1

[58] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 3

[59] Anuj Srivastava, A.B. Lee, Eero Simoncelli, and Song Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18:17–33, 01 2003. 1

[60] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent

Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 2

[61] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. *CoRR*, abs/1808.01974, 2018. 2, 6

[62] Hans-Peter Seidel Tunç O. Aydın, Rafal Mantiuk. Extending quality metrics to full luminance range images. In *Human Vision and Electronic Imaging*, pages 68060B–10. Spie, 2008. 1

[63] Domonkos Varga. A combined full-reference image quality assessment method based on convolutional activation maps. *Algorithms*, 13(12):313, Nov 2020. 2

[64] Domonkos Varga. Composition-preserving deep approach to full-reference image quality assessment. *Signal, Image and Video Processing*, 14, 09 2020. 7

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 2, 3, 4

[66] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. *CoRR*, abs/1704.06904, 2017. 2

[67] Zhou Wang and Alan Bovik. *Modern Image Quality Assessment*. Morgan & Claypool Publishers, 1st edition, 2006. 1

[68] Zhou Wang, Alan C. Bovik, and Ligang Lu. Why is image quality assessment so difficult? In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3313–IV–3316, 2002. 1

[69] Zhou Wang, Alan C. Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. 1, 2, 4, 7

[70] Zhou Wang and Qiang Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198, May 2011. 2

[71] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, Nov 2003. 2, 7

[72] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 995–1002, June 2013. 1

[73] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *CoRR*, abs/1308.3052, 2013. 2

[74] Lin Zhang, Ying Shen, and Hongyu Li. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, Oct 2014. 2, 6, 7

[75] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, Aug 2011. 2, 7

[76] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, June 2018. 2, 5

[77] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. *CoRR*, abs/1807.02758, 2018. 2, 3, 4, 5