mathematics

MDPI

Article

Full-Reference Image Quality Assessment with Transformer and DISTS

Pei-Fen Tsai, Huai-Nan Peng, Chia-Hung Liao and Shyan-Ming Yuan *

Computer Science Department, National Yang Ming Chiao Tung University, No. 1001, Daxue Rd., Est Dist., Hsinchu City 300093, Taiwan; nelson870708.cs09@nycu.edu.tw (H.-N.P.)
* Correspondence: smyuan@nycu.edu.tw; Tel.: +886-3-5712121 (ext. 56631)

Abstract: To improve data transmission efficiency, image compression is a commonly used method with the disadvantage of accompanying image distortion. There are many image restoration (IR) algorithms, and one of the most advanced algorithms is the generative adversarial network (GAN)-based method with a high correlation to the human visual system (HVS). To evaluate the performance of GAN-based IR algorithms, we proposed an ensemble image quality assessment (IQA) called ATDIQA (Auxiliary Transformer with DISTS IQA) to give weights on multiscale features global self-attention transformers and local features of convolutional neural network (CNN) IQA of DISTS. The result not only performed better on the perceptual image processing algorithms (PIPAL) dataset with images by GAN IR algorithms but also has good model generalization over LIVE and TID2013 as traditional distorted image datasets. The ATDIQA ensemble successfully demonstrates its performance with a high correlation with the human judgment score of distorted images.

Keywords: image quality assessment (IQA); full-reference IQA; deep image structure and texture similarity (DISTS); transformer IQA; PIPAL dataset; ensemble IQA

MSC: 68T07

Citation: Tsai, P.-F.; Peng, H.-N.; Liao, C.-H.; Yuan, S.-M. Full-Reference Image Quality Assessment with Transformer and DISTS. *Mathematics* **2023**, *11*, 1599. https://doi.org/10.3390/math11071599

Academic Editor: Konstantin Kozlov

Received: 6 March 2023
Revised: 19 March 2023
Accepted: 23 March 2023
Published: 26 March 2023

1. Introduction

Due to the vigorous development of the Internet of Things (IoT), the necessary compression and restoration of images in the transmission process have become increasingly hot topics. With the progress of deep learning models including the convolutional neural network (CNN) [1,2] and the generative and adversarial network (GAN) [3,4], the quality of image restoration [5], image compression, and super-resolution [6] has been greatly improved. Therefore, having a metric to quantify the image quality to the human visual system (HVS) is always a tough task. Image quality assessment (IQA) [7–9] can be subjective or objective.

Subjective IQA judged through human eyes is more reliable and accurate for HVS and is named by mean opinion scores (MOS). However, subjective IQA is time-consuming and expensive. Recently, more and more studies on objective IQA models have achieved outstanding results, including full-reference IQA (FR-IQA) [9], reduced-reference IQA (RR-IQA), and no-reference IQA (NR-IQA), which can be used to predict the image quality score.

FR-IQA compares the distance of the reference image and distorted image as a pair to measure the distorted image quality. NR-IQA predicts the distorted image quality without a reference image. That means NR-IQA has no prior knowledge for original information and limited uses for measuring a distorted image compared to a restored image.

However, in these years, with neural network contribution, NR-IQA dud has achieved a significant performance on metric distorted images, such as MetalIQA [10], proposed in 2020, and GraphIQA [11] and LIQA [12], proposed in 2022.

In this paper, we focus on FR-IQA since we are interested in not only distorted image quality but image quality after the restoration algorithm. Hence, we propose an ensemble

FR-IQA model named the auxiliary transformer with DISTS for image quality assessment (ATDIQA)-weighted average of the quality score from the models of the transformer IQA and CNN-based IQA.

ATDIQA was trained on the PIPAL [13] dataset to include distorted images and restoration images with ensemble methodology. Then, we tested it on the distorted-based dataset of LIVE [14] and TID2013 [15] for Pearson's linear correlation coefficient (PLCC) [16] correlation metric and Spearman's rank correlation (SRCC). In the LIVE dataset, PLCC and SRCC are over 0.9 with great correlation to human perception. In the TID2013 dataset, PLCC and SRCC are over 0.8 with a still-good correlation to human perception. In the PIPAL test set, we obtained over 0.83 PLCC and SRCC with good linearity. ATDIQA did demonstrate its good model generalization capability and capability for image quality assessment on both distorted images and images with various restored algorithms.

2. Background and Related Work

2.1. Full-Reference Image Quality Assessment (FR-IQA)

The FR-IQA algorithm as shown in Figure 1 inputs the original image, that is, the reference image, and the distorted image, that is, the quality image to be judged. After the image resolution is unified through preprocessing, the features of the images are extracted through feature extraction, and then, the distance between the distorted image and the reference image is calculated to obtain the image quality score of the distorted image.

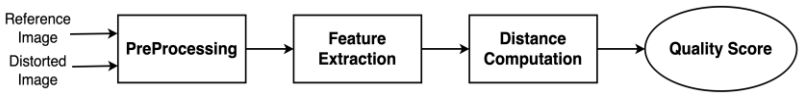

Figure 1. FR-IQA model.

In the distance computation, the mean square error (MSE) and peak signal-to-noise ratio (PSNR) are commonly used methodology. However, metrics of MSE and PSNR cannot obtain accurate image quality score for some images from deep-learning based image restoration algorithms.

To improve the poor performance of MSE and PSNR, structural similarity (SSIM) [17] was proposed to include image luminance, contrast, and structure as an index in FR-IQA to calculate the similarity of the image structure between a reference image and a distorted image. Then, SSIM was extended to a multiscale structural similarity index called MS-SSIM. Structural-based metric models of image luminance, contrast, and structure on multiple scales.

The convolutional neural network (CNN) architecture shows significant achievement in computer vision, including image classification, image segmentation, and object detection, due to its ability to extract different scales of image features. In FR-IQA, the use of CNN architecture to extract multiple levels of features for distance computation was proposed by learned perceptual image patch similarity (LPIPS) [18] and deep image structure and texture similarity (DISTS) [19].

The generative adversarial network (GAN) has been proven to be a good approximator for natural image restoration (IR) [20–24]. In 2021, the perceptual image processing algorithm (PIPAL) dataset [21] was developed with 116 distortions of traditional IR and GAN-based IR as a novel large-scale IQA dataset. The PIPAL dataset had been used in the CVPR NTIRE 2021 and 2022 challenge as a benchmark for IQA algorithms.

In the CVPR NTIRE 2021 challenge, the team that won first implemented the transformer-based IQA named IQT-C [25] to demonstrate the global attention of transformer-to-IQA capability. In the CVPR NTIRE 2022 challenge [26], the team that won first developed an ensemble method with transformer and CNN for feature extraction named attention-based hybrid image quality assessment network [27]. The ensemble deep learning model shows its power in IQA as well.

2.2. IQA Datasets

Numerous IQA datasets have been proposed over decades. The FR-IQA dataset contains reference images and different types of distortion images as pairs. Each distorted image is evaluated by MOS relative to the reference image by human judgment. The distortion types [28,29] vary to include traditional JPEGs such as image compression, white noise, Gaussian blur, and the traditional algorithm for image restoration.

Some of the well-known FR-IQA datasets are LIVE [14], TID2008 [30], TID2013 [15], and KADID-10k [31], as shown in Table 1 with the increase in reference images, distortion images, and distortion types yearly.

Table 1. Image Quality Assessment Datasets.

Image Restoration	Dataset	Year	No. Ref	No. Dist.	No. Dist. Type
Traditional Algorithm	LIVE [24]	2004	29	0.8k	5
	TID2013 [25]	2013	25	3k	24
	KADID-10k [27]	2019	81	10.1k	25
Trad. + GAN Algorithm	PIPAL [10]	2020	250	29k	40

PIPAL [13], first introduced in 2020 by Jinjin et al., increased the reference images to 250 and distortion types to 40 and introduced distorted images using the GAN-based restoration algorithm as the current state-of-the-art FR-IQA dataset today. It is also used to evaluate FR-IQA models in distorted and restored images to include the GAN-based algorithm in the CVPR NTIRE 2021 IQA challenge.

2.3. Feature Extraction Backbone in FR-IQA

Feature extraction backbones are key components in learning-based IQA algorithms as published in recent years. In 2018, Zhang et al. investigated different CNN-based models such as LPIPS [18] to demonstrate that deep features are representative of image semantics for evaluating image quality.

In 2021, Guo et al. proposed the IQA of multiscale features as an image quality multiscale assessment network (IQMA) [28] to demonstrate the power of the feature fusion module with feature pyramid network (FPN) backbone for multiscale feature fusion.

Shi et al. significantly improved WResNet IQA models (W stands for weighted averaging) [32] by modifying the residual block in the feature extraction backbone of the fourth ranking in the NTIRE 2021 IQA challenge.

2.4. Ensemble Methods in FR-IQA

Ensemble learning has been shown to be effective in performance on deep learning tasks. In FR-IQA, there are several models that use ensemble methodology to obtain high correlation metrics: the IQMA [33] network, which won the 2nd award in the NTIRE 2021 competition, proposed to average predicted quality scores from different levels of feature maps. EGB predicts image quality scores with three regressors and averages the three predicted scores to obtain the final quality score.

Ensemble methodology [34] is commonly used to solve statistical and computational problems while maintaining model generalization. It has been demonstrated successfully in various fields including face recognition [35], emotion recognition [36,37], medical treatment [38,39], etc. In the NTIRE 2021 IQA challenge, the team proposing an IQA metric named the gradient boosting (EGB) ensemble [40], with three regressors for the image quality score, won the second prize for outstanding performance of the IQA ensemble metric.

3. Methodology

In 2022, S.S. Lao et al. proposed an attention-based hybrid image quality assessment network (AHIQ) [27] to fuse the transformer-based IQA and CNN-based IQA with feature fusion module. This inspired us to ensemble the IQT and CNN-based DISTS IQA models as an ensemble IQA. In our model, we used VGG16 to reduce computational cost since the AHIQ model uses the deep backbone of Resnet50 as the CNN backbone. In that case, our ATDIQA is an ensemble model with the multi-metric of CNN IQA and transformer IQA, as shown in Figure 2.

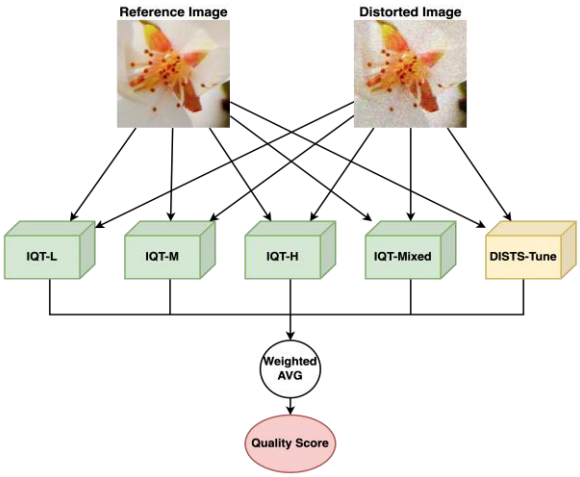

Figure 2. ATDIQA model architecture with a weighted average of IQT-L, IQT-M, IQT-H, IQT-Mixed, and DISTS-Tuned.

During training, we fine-tuned the weights of the feature extraction backbone for all models as transfer learning for the IQA task.

In 2021, Google proposed a multiscale image quality transformer (MUSIQ) [41] that trained their model on four scales of the input in IQA datasets. MUSIQ showed that multiscale training images can capture image quality of granularities.

Instead of increasing the training images with multiple scales, we tend to apply multi-level features. Hence, we built three different levels of feature extraction IQA models of IQT-L, IQT-M, and IQT-H models for the weighted sum score.

The approach is similar to the multi-level fusion of IQMA [33]. After training, ATDIQA can be observed regarding the weight of different scales of the IQA model's contribution to IQA, and we can compare the PLCC and SRCC to human perception.

3.1. DISTS-Based Methods

Deep image structure and texture similarity (DISTS) is a CNN-based FR-IQA, as shown in Figure 3, proposed by K. Ding et al. in 2020. DISTS has two components: a feature extraction backbone and image structure and texture similarity computation as the quality score. The feature extraction backbone is pre-trained by VGG16 [42] to extract feature maps from layer conv1_2, conv2_2, conv3_3, conv4_3, and conv5_3. It contains different levels of features including structure and texture information. The other component is the weighted summation of texture and structure similarity for each feature level and the output quality score.

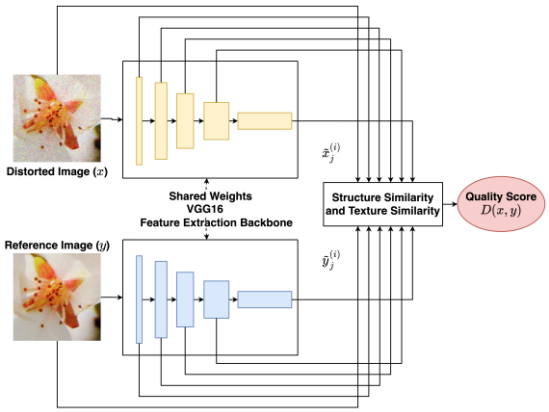

Figure 3. Deep image structure and texture similarity (DISTS) with feature extraction by VGG16 of pre-trained weight, including six stages of feature maps of five layers of 3, 64, 128, 256, 512 and 512 for similarity distance computation and output quality score.

DISTS is trained on the KADID-10k dataset and shows the improved performance of PLCC and KRCC on conventional image quality datasets as a baseline of SSIM, with its self-learned weight for structure and texture similarity with high tolerance to texture resampling. DISTS shows its capability with a pre-trained fixed VGG16 backbone to extract features from the reference image and distorted image as IQA input. The metric of weighted structural similarity and texture structure summation in multiple levels as a quality score of distorted images is given in Equations (1)–(3).

$$D(x, y; \alpha, \beta) = 1 - \sum_{i=0}^{m} \sum_{j=1}^{n_i} \left(\alpha_{ij} l\left(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}\right) + \beta_{ij} l\left(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}\right) \right) \tag{1}$$

$$l\left(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}\right) = \frac{2\mu_{\tilde{x}_j}^{(i)} \mu_{\tilde{y}_j}^{(i)} + c_1}{\left(\mu_{\tilde{x}_j}^{(i)}\right)^2 + \left(\mu_{\tilde{y}_j}^{(i)}\right)^2 + c_1} \tag{2}$$

$$s\left(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}\right) = \frac{2\sigma_{\tilde{x}_j \tilde{y}_j}^{(i)} + c_2}{\left(\sigma_{\tilde{x}_j}^{(i)}\right)^2 + \left(\sigma_{\tilde{y}_j}^{(i)}\right)^2 + c_2} \tag{3}$$

3.2. IQT-Based Methods

The transformer model [12] demonstrated its great power in NLP (Natural Language Processing) with a global self-attention module. In 2017, transformer was extended to computer vision, proposed as ViT (Vision Transformer) [43], with a self-attention algorithm to calculate weighted images as global image contents. Multi-head attention is functional as convolution kernels and can compute parallelly. The global self-attention module makes transformer models even more powerful compared to local feature extraction in CNN models.

ViT models also extend from image classification and object detection (DETR) [44] to IQA such as TRIQ (Transformer in Image Quality) [45] and IQT-C [25], proposed in 2021. The team that proposed IQT-C, as shown in Figure 4, won the first in the 2021 NTIRE IQA challenge in the PIPAL dataset.

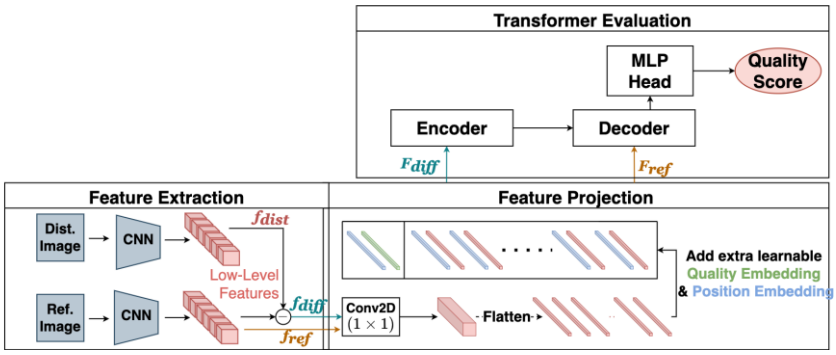

Figure 4. IQT-C Feature Embedding of Six Low-Level Features of Inception-Resenet-V2.

The IQT-C architecture consists of four main components:

- Feature extraction backbone: Inception-Resnet-V2 [46] of six layers of low-level features in Figure 5 with a fixed pre-trained weight from ImageNet [47];
- Feature projection: To reduce the dimension of the feature map with 1×1 convolution as an embedded feature;
- Transformer encoder-decoder: embedded feature to input the encoder and decoder for multiple self-attention;
- MLP Head + output of quality score.

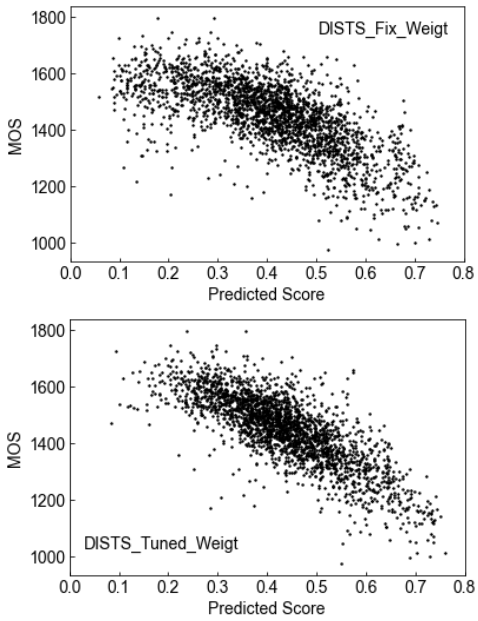

Figure 5. Scatter plot of DISTS models to MOS with no tuned and tuned backbone weight.

IQT-C obtained a significant improvement in test results in PLCC and SRCC to beat the baseline of DISTS on PIPAL. It demonstrates the power of global self-attention for distorted image quality in the GAN-based IR algorithm. The ablation study shows that the difference of features between the reference and the distorted image is key as an input of the encoder to predict human perception of a distorted image.

3.3. Testing Protocol

In the test phase, the image is cropped into five overlapping H × W patches: left-top, right-top, left-bottom, right-bottom, and center. The predicted score of a distorted image is an average of five patches of the test image. Each test image has five predicted scores of five models, namely IQT-L, IQT-M, IQT-H, IQT-Mixed, and DISTS-Tuned. Then, the sum

of weights is obtained with the predicted score of each distorted image as the ATDIQA predicted score.

The same test is applied to the PIPAL test set, LIVE, and TID2013 for the generalization capability of the model with PLCC, SRCC, and KRCC metrics.

3.4. Weighted Averaging

DISTS-based methods use the distance to represent the image quality score. The shorter the distance between the pair of images, the higher the IQA score of the distorted image.

However, IQT-based methods calculate the score directly: the higher the score of the distorted image, the better the quality compared to the reference image.

In that case, we performed the normalization of the predicted IQA score with third-order polynomial nonlinear regression to the score of MOS score for all models and optimize to obtain the best weight average for the five models.

4. Experiments

4.1. IQA Datasets

The PIPAL [13] dataset is the main dataset used for training with 250 reference images of 288 × 288 using 40 distortion types in 29k distorted images, as shown in Table 1. However, the public PIPAL dataset does not include a validation and testing set for the NTIRE 2021 IQA competition. Therefore, we divided the public training dataset with ratios of 80%, 10%, and 10% as our training set, validation set, and test set, respectively, as described in Table 2. Overall, 80% of the public data was used for model training, and a 10% validation set was used to optimize the hyperparameters. Then, 10% of the test set was used to predict the quality score and evaluate the IQA model with correlation metrics. We also included the LIVE and TID2013 as test datasets to compare the correlation metrics in different datasets for the model's capability for data generalization.

Table 2. IQA datasets for model training, validation, and test.

	Training Set	Validation Set	Test Set
Public PIPAL [13]	80%	10%	10%
LIVE [14]	-	-	100%
TID2013 [15]	-	-	100%

4.2. Evaluation Metrics

To evaluate the objective IQA performance, a correlation between MOS and the objective IQA quality score is commonly used. There are three correlation metrics: PLCC, SRCC, and KRCC.

SRCC and KRCC measure the monotonicity of the predicted quality score, while PLCC measures the linearity of the predicted quality score. The higher the correlation metrics, the better the performance of the IQA prediction.

On the other hand, PLCC is sensitive to the difference between the ground truth score of MOS and the predicted score. In that case, we employed a nonlinear third-order polynomial regression to normalize the predicted scores on the MOS range scale [13,15]. In that case, the PLCC was calculated between the MOS scores and the normalized predicted scores.

4.3. DISTS-Based Methods

4.3.1. Ablation Study for Dataset and Weight Tuning

In this section, we perform the ablation study to discuss the performance with fine-tuning of weights, as DISTIS in Figure 3 uses the backbone of VGG16 with fixed weight for feature extraction with the pre-trained weight of ImageNet [47,48], which is commonly used as a feature extraction in an image classification task.

However, fine-tuning weight [49] to fit a specific task [50] is commonly used to improve performance to fit a specific task and accelerate training to convergence. The training dataset is 80% the public PIPAL instead of KADID-10k since PIPAL has 29k image pairs with 40 distortion types. Diversity datasets are known to improve model generalization ability.

In the feature extraction backbone, we kept the VGG16 and the weights backpropagate: not only structure and texture similarity parameters but also the VGG16 feature extraction parameters.

Then, we calculated the three correlations of DISTS-Tuned and MOS on the 10% PIPAL test datasets, as shown in Table 3. It shows that DISTS with fine-tuned backbone weights improves performance by up to 5% in correlation coefficients. Weights' fine tuning will improve the model to fit the IQA task.

Table 3. Ablation Study for DISTS with Fixed and Tuned Weight on PIPAL test datasets.

			Correlation Result		
Model	Training Dataset	Tuned Weight	PLCC	SRCC	KRCC
DISTS	PIPAL 80%	✘	0.661	0.648	0.466
		✓	0.723	0.703	0.518

We also applied the same methodology to the IQT-C model [25] to observe how it affects performance on the transformer-based IQA as in Table 4. It showed no significant improvement in the performance of the correlation with MOS. This may be the reason that global attention has good performance in the IQA task as well. That is not surprising since transformer-based IQA won the top four spots in the past two years of the NTIRE IQA competition in 2021 and 2022 [26].

Table 4. Ablation Study for IQT-C with Fixed and Tuned Weight on PIPAL Test Datasets.

			Correlation Result		
Model	Training Dataset	Tuned Weight	PLCC	SRCC	KRCC
IQT-C	PIPAL 80%	✘	0.839	0.810	0.621
		✓	0.845	0.814	0.629

4.3.2. DISTS-Tuned Model Performance

In Figure 5, the scatter plots show the comparison between, with, and without DISTS backbone weight. It can be observed that the backbone-weight-tuned DISTS model obtained better correlation with MOS with smaller variance, shown by the 5% model performance improvement given in Table 4.

It is shown in Figure 6 that the DISTS-Tuned weight model will be able to adopt different kinds of image restoration (IR), including traditional SR, CNN-based SR, and GAN-based SR with its model generalization capability. Moreover, the model can adopt the IQA metric for GAN-based SR images with same trend, which is viewed as a tough IQA task.

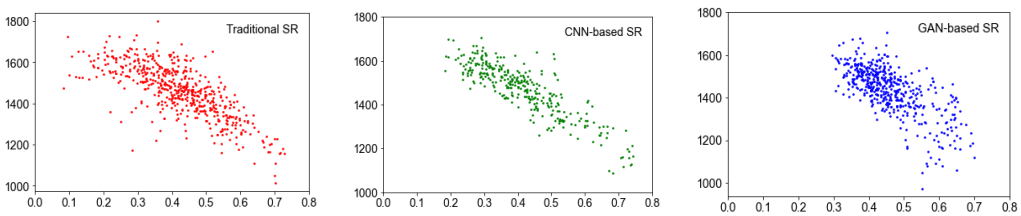

Figure 6. Scatter plot for three kinds of super-resolution methodology.

4.4. IQT-Based Methods

4.4.1. Multiscale Feature Projection in IQT

In CNN-style transfer [51], low-level features of the image are used to extract information about image texture, such as color, edge and corner, and pixel, while high-level features are used to extract information about image structure, such as image semantics. For GAN-based distortion in PIPAL, IQT-C uses low-level features as the input of the self-attention module. However, high-level features that contain information about the image structure may be useful as difference information between reference and distorted images.

In this hypothesis, we built four feature projection modules, namely low-level feature projection of IQT-L, medium-level feature projection of IQT-M, high-level feature projection of IQT-H, and three levels of feature projection of IQT-Mixed to improve IQT-C with only low-level features.

The feature projection module in IQT-Mixed includes low-level features, medium-level features, and high-level features in Figure 7 as the corresponding layers of Inception-ResNet-V2 for IQT-Mixed multiscale feature extraction layers.

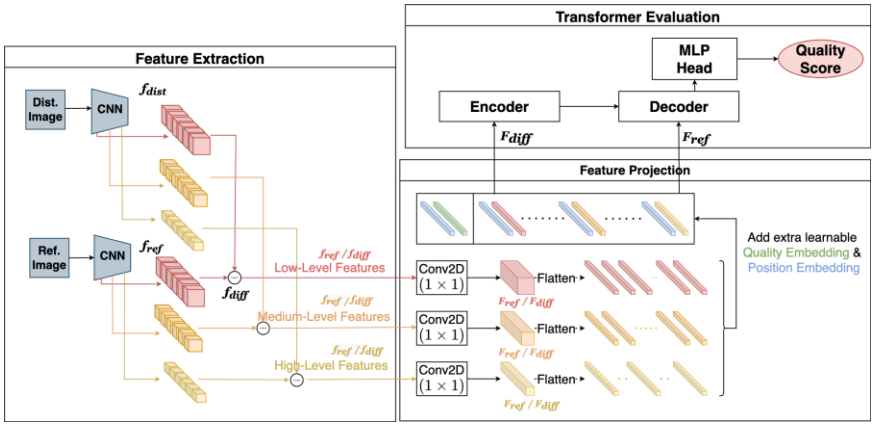

Figure 7. Feature projection of IQT-Mixed.

Parameters of model training with 80% training data are as follows:

- Training epochs = 200;
- Learning rate = 1×10^{-4};
- Loss function = MSE;
- Optimizer = Adam with batch size 16;
- Pre-train weight of ImageNet;
- Weight-tuned in backpropagation.

The training converges for each model within 100 epochs with pre-trained weight from ImageNet.

Next, 10% validation data were used to optimize hyperparameters in the transformer structure for IQT-L, IQT-M, IQT-H, and IQT-Mixed.

4.4.2. Performance of the Multiscale Feature Extraction Backbone

We compared the multiscale backbone of IQT models in three datasets including PIPAL, LIVE, and TID2013 in Table 5. Transformer-based IQA showed a huge improvement for its self-attention mechanism compared to CNN-based DISTS-Tuned IQA. Multi-level feature extraction models of IQT-L, IQT-M, IQT-H, and IQT-Mixed showed about 4% improvement in LIVE and TID2013 datasets rather than in PIPAL, which showed good model generalization.

Table 5. Correlation Comparison for IQT-C and Multiscale Feature-Extraction-Tuned Backbone of IQT in PLCC and SRCC.

Model	PIPAL Test Set [10]			LIVE [24]			TID2013 [26]		
	PLCC	SRCC	PLCC + SRCC	PLCC	SRCC	PLCC + SRCC	PLCC	SRCC	PLCC + SRCC
IQT-C (Baseline)	0.839	0.810	1.649	0.873	0.863	1.736	0.801	0.757	1.558
IQT-L	0.845↑	0.814↑	1.659↑	0.901↑	0.910↑	1.811 (2)	0.808↑	0.751↑	1.559 (4)
IQT-M	0.834	0.807	1.641	0.890↑	0.902↑	1.792 (4)	0.834↑	0.790↑	1.624 (2)
IQT-H	0.822	0.795	1.617	0.899↑	0.910↑	1.809 (3)	0.821↑	0.784↑	1.605 (3)
IQT-Mixed	0.840↑	0.817↑	1.656↑	0.912↑	0.922↑	1.834 (1)	0.855↑	0.826↑	1.681 (1)

"↑" shows improvement in PLCC and SRCC comparing to IQT-C (Baseline).

The high correlation of models with MOS shows that human perception is sensitive to low-level features such as colors and brightness-related textures as well as high-level features such as structural patterns. Human perception to image quality is a combination of multilevel features with different sensitivity. This inspired us to combine these transformer-based multiscale IQA models into one ensemble model.

4.5. Ensemble Model as an Auxilary Transformer with DISTS IQA(ATDIQA)

To combine global and local features of images, we tested the ensemble multiscale IQT, as in Section 4.4, and DISTS-Tuned, described in Section 4.3, as an ensemble for evaluation. Using differential evolution [52] for the determination of the weights of the five models of IQT-L, IQT- M, IQT-H, IQT-Mixed, and DISTS-Tuned, we then performed normalization to make them sum up to 1. The results were weights of 0.365, 0.072, 0.08, 0.396, and 0.088, respectively, according to our ATDIQA model in Table 6.

Table 6. Weights in ensemble models of ATDIQA.

IQT-L	IQT-M	IQT-H	IQT-Mixed	DISTS-Tuned
0.365	0.072	0.08	0.396	0.088

The results show that human perception is sensitive to low-level features of images such as color and fine texture. However, medium- and high-level features still play an important role in human perception of image quality judgment.

5. Discussion

The evaluation results comprise the three datasets of the PIPAL test set, LIVE and TID2013 for PLCC, and SRCC. We compared the correlation MOS with IQT-C (transformer-based IQA) [25], which is our baseline; DISTS-Tuned (CNN-based IQA) [17]; and our ATDIQA (ensemble IQA) with performance comparison and improvement.

5.1. Model Performance on Super Resolution Image Restoration Algorithm

Our goal was to build an ensemble IQA model with capability for not only the traditional SR algorithm but the GAN-based SR algorithm since to build an objective IQA for the GAN-based SR algorithm is a tough task regarding achieving a good correlation with human perception.

To compare the performance, we sampled images from PIPAL test dataset since PIPAL is the only dataset with a GAN-based and traditional SR algorithm.

We sampled one of the reference images and obtained scores with different distorted levels, as listed in Table 7 from left to right and as MOS from high to low. The distortion type is super-resolution, including GAN and traditional algorithms. The calculated quality score includes MSE, PSNR, SSIM [7], DISTS [9], IQT-C [13], and the ATDIQA metric as FR-IQA with the reference image from the left side.

Table 7. Example image from the PIPAL test dataset for different IQA models (Data from the PIPAL dataset 2020 [13]).

	Super-Resolution Algorithm for Image Restoration					
	RankSRGAN X4 PI oriented	BOE X4 R2	EPSR X4 R3	ESRGAN X6	AR-CNN + EDSR 10 X2	BM3D + EDSR 50 X3
IQA	GAN SR[1]	GAN SR[1]	GAN SR[1]	GAN SR[1]	Traditional SR[2]	Traditional SR[2]
MOS↑	1571.249	1492.574	1394.923	1324.865	1235.043	1133.264
MSE↓	211.725 (4)	144.199 (2)	135.347 (1)	294.221 (6)	202.091 (3)	294.158 (5)
PSNR↑	24.873 (4)	26.541 (2)	26.816 (1)	23.444 (6)	25.075 (3)	23.445 (5)
SSIM↑ [17]	0.617 (4)	0.704 (1)	0.703 (2)	0.521 (5)	0.622 (3)	0.517 (6)
DISTS↓ [19]	0.126 (1)	0.127 (2)	0.131 (3)	0.185 (4)	0.268 (5)	0.308 (6)
IQT-C↑ [25]	0.604 (1)	0.534 (2)	0.500 (3)	0.449 (4)	0.321 (5)	0.151 (6)
ATDIQA↑	0.639 (1)	0.580 (2)	0.552 (3)	0.480 (4)	0.364 (5)	0.226 (6)

GAN-SR[1], GAN-based super-resolution algorithm; Traditional-SR[2], traditional super-resolution algorithm. The images listed from left to right based on image quality after distortion or restoration. "↑" is IQA scores of list of images from low to high; "↓" is IQA scores of list of images from high to low.

DISTS, IQT-C, and ATDIQA have a high correlation to MOS, and even when the distorted type of distorted images is GAN-based super-resolution, these algorithms can predict the quality scores accurately. This shows the power of the feature extraction backbone for image content extraction.

IQT-C and ATDIQA have better resolution to different levels of distortion in GAN-based super-resolution images with the global attention power of the transformer-based algorithm.

ATDIQA shows the best overall resolution for distorted levels from left to right. This is quite impressive in terms in terms of the power of the IQA group for image quality. ATDIQA not only performs linearly to MOS, but also has the highest resolution for distortion level on the GAN-based algorithm.

5.2. Linearity of ATDIQA to Human Perception

The ATDIQA scatter plot shows linearity with linear regression adjustment in the PIPAL dataset, as given in Figure 8. The good R-squared indicates that ATDIQA can measure distorted levels of an image from low to high as linear to human perception.

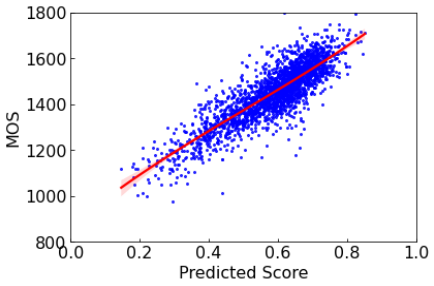

Figure 8. ATDIQA predicted the MOS score on the PIPAL dataset.

5.3. Model Generalization in Different Datasets

To compare different feature extraction backbones, we tested each model over the three datasets. We list PLCC, SRCC, and KRCC in Table 8.

Table 8. Comparison of ATDIQA and baseline models of IQT_C and DISTS on the PIPAL test set, LIVE, and TID2013.

Model	PIPAL Test Dataset [10]			LIVE [24]			TID2013 [26]		
	PLCC	SRCC	PLCC + SRCC	PLCC	SRCC	PLCC + SRCC	PLCC	SRCC	PLCC + SRCC
DISTS-Tuned	0.799	0.703	0.518	0.948	0.929	0.776	0.741	0.637	0.468
IQT_C	0.839	0.810	1.649	0.873	0.863	1.736	0.801	0.757	1.558
ATDIQA	0.854	0.830	1.684	0.931	0.931	1.862	0.837	0.795	1.632

IQT-Mixed with a fusion of low-level, medium-level, and high-level feature projection significantly improved the correlation metrics compared to IQT-C of low-level feature projection, with a weight of 36.5% in IQT-L. This gives evidence that human perception is quite sensitive to changes in image texture. However, high-level features such as image structure still impact human perception of image quality by a weight of 8%.

The bar chart in Figure 9 compares the IQT-C of transformer-based IQA, DISTS of CNN-based IQA, and ATDIQA as a group IQA. It shows that ATDIQA has a stable performance across the three datasets and provides good model generalization to score distorted image quality on different types of data in different IQA datasets.

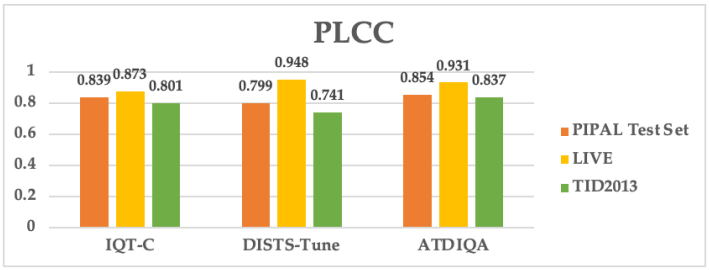

Figure 9. PLCC in IQT-C (baseline), DISTS-Tuned (CNN-based IQA), and ATDIQA (ensemble IQA) in the PIPAL test set, LIVE, and TID2013 dataset.

6. Conclusions

In this paper, we propose the auxiliary transformer of DISTS IQA (ATDIQA) as a FR-IQA ensemble. In Table 7, we compare the performance of different IQAs of the quality assessment methodology with the corresponding type of feature extraction. It is not surprising that human perception is a complicated system that requires a deep neural network for IQA.

There are mainly three-fold improvements and findings in our study:

First, by fine-tuning the weight of the feature extraction backbone, the model can better fit the quality score prediction task and improve performance compared to fixed backbone weight.

Second, fusing different levels of feature maps can further improve attention to the image content of structure and texture with the feature projection module. ATDIQA with IQC-L, IQT-M, IQT-H, IQT-Mixed, and DISTS-Tuned with a weighted sum for global and local feature attention provides outstanding model generalization to PIPAL, TID2013, and LIVE datasets and shows improvement over IQT-C [25] as well.

Third, the weights of different submodules in ATDIQA show that human perception is extremely sensitive to low-level features of image texture, while medium-level and high-level features of image structures are considered at the same time when scoring image quality. The fusion of feature extraction IQA was proposed by Guo et al. [33] as well as IQMA.

These are the three key reasons that ATDIQA showed a significant improvement correlation in predicting distorted image quality. In the future, we can try to modify the feature extraction from the pyramidal feature hierarchy of ATDIQA to the feature pyramid network as IQMA and give global attention to fusion features with transformer. Fusion

feature extraction for global attention may reduce the model parameters with prediction efficiency improvement and provide the same or even better scoring performance for GAN-based algorithm-distorted images.

Since the deep learning model is data-driven, augmented data with GAN [53] to increase training data diversity are another possible method, as more and more GAN-based super-resolution algorithms [53–56] have successfully demonstrated their ability for image restoration in the recent years.

Author Contributions: Conceptualization, H.-N.P. and C.-H.L.; methodology, H.-N.P.; software, H.-N.P.; writing—original draught preparation, P.-F.T. and H.-N.P.; writing—review and editing, P.-F.T.; visualization, P.-F.T.; supervision, S.-M.Y.; funding acquisition, S.-M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by NSTC, grant number 111-2410-H-A49-070-MY2.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the datasets used in this article are openly available as below link. PIPAL dataset: https://www.jasongt.com/projectpages/pipal.html; https://drive.google.com/drive/folders/1G4fLeDcq6uQQmYdkjYUHhzyel4Pz81p-; LIVE dataset: http://live.ece.utexas.edu/research/quality/subjective.htm; TID2013 dataset: https://ponomarenko.info/tid2013.htm.

Conflicts of Interest: The authors declare that they have no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
2. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6.
3. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
4. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative Adversarial Networks: An Overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
5. Banham, M.R.; Katsaggelos, A.K. Digital image restoration. *IEEE Signal Process. Mag.* **1997**, *14*, 24–41. [CrossRef]
6. van Ouwerkerk, J. Image super-resolution survey. *Image Vis. Comput.* **2006**, *24*, 1039–1052. [CrossRef]
7. Wang, Z.; Bovik, A.C.; Lu, L. Why is image quality assessment so difficult? In Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, 13–17 May 2002; Volume 4, pp. IV-3313–IV-3316.
8. Zhai, G.; Min, X. Perceptual image quality assessment: A survey. *Sci. China Inf. Sci.* **2020**, *63*, 1–52. [CrossRef]
9. Sheikh, H.; Sabir, M.; Bovik, A. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Trans. Image Process.* **2006**, *15*, 3440–3451. [CrossRef]
10. Zhu, H.; Li, L.; Wu, J.; Dong, W.; Shi, G. MetaIQA: Deep meta-learning for no-reference image quality assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14143–14152.
11. Sun, S.; Yu, T.; Xu, J.; Zhou, W.; Chen, Z. GraphIQA: Learning Distortion Graph Representations for Blind Image Quality Assessment. *IEEE Trans. Multimedia* **2022**. [CrossRef]
12. Liu, J.; Zhou, W.; Li, X.; Xu, J.; Chen, Z. LIQA: Lifelong Blind Image Quality Assessment. *IEEE Trans. Multimedia* **2022**, 1–16. [CrossRef]
13. Jinjin, G.; Haoming, C.; HaoYu, C.; Xiaoxing, Y.; Ren, J.S.; Chao, D. PIPAL: A Large-Scale Image Quality Assessment Dataset for Perceptual Image Restoration. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 633–651. [CrossRef]
14. Sheikh, H. LIVE Image Quality Assessment Database Release 2. 2005. Available online: http://live.ece.utexas.edu/research/quality/subjective.htm (accessed on 5 June 2021).
15. Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; et al. Image database TID2013: Peculiarities, results and perspectives. *Signal Process. Image Commun.* **2015**, *30*, 57–77. [CrossRef]
16. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–4.
17. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]

18. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 586–595.
19. Ding, K.; Ma, K.; Wang, S.; Simoncelli, E.P. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2567–2581. [CrossRef]
20. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.P.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
21. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
22. Wang, X.; Yu, K.; Dong, C.; Loy, C.C. Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 606–615. [CrossRef]
23. Zhang, W.; Liu, Y.; Dong, C.; Qiao, Y. RankSRGAN: Generative Adversarial Networks with Ranker for Image Super-Resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019. [CrossRef]
24. Cai, H.; He, J.; Qiao, Y.; Dong, C. Toward interactive modulation for photo-realistic image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 294–303.
25. Cheon, M.; Yoon, S.-J.; Kang, B.; Lee, J. Perceptual image quality assessment with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 433–442.
26. Gu, J.; Cai, H.; Dong, C.; Ren, J.S.; Timofte, R.; Gong, Y.; Lao, S.; Shi, S.; Wang, J.; Yang, S.; et al. NTIRE 2022 challenge on perceptual image quality assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 951–967.
27. Lao, S.; Gong, Y.; Shi, S.; Yang, S.; Wu, T.; Wang, J.; Xia, W.; Yang, Y. Attentions Help CNNs See Better: Attention-based Hybrid Image Quality Assessment Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1139–1148. [CrossRef]
28. Larson, E.C.; Chandler, D.M. Most apparent distortion: A dual strategy for full-reference image quality assessment. In *Image Quality and System Performance VI*; SPIE: Sydney, Australia, 2009; Volume 7242, pp. 270–286. [CrossRef]
29. Larson, E.C.; Chandler, D.M. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electron. Imaging* **2010**, *19*, 011006.
30. Ponomarenko, N.; Lukin, V.; Zelensky, A.; Egiazarian, K.; Carli, M.; Battisti, F. TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Adv. Mod. Radioelectron.* **2009**, *10*, 30–45.
31. Lin, H.; Hosu, V.; Saupe, D. KADID-10k: A large-scale artificially distorted IQA database. In Proceedings of the 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), Berlin, Germany, 5–7 June 2019; pp. 1–3.
32. Shi, S.; Bai, Q.; Cao, M.; Xia, W.; Wang, J.; Chen, Y.; Yang, Y. Region-adaptive deformable network for image quality assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 324–333.
33. Guo, H.; Bin, Y.; Hou, Y.; Zhang, Q.; Luo, H. Iqma network: Image quality multi-scale assessment network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 443–452.
34. Zhou, Z.-H. *Ensemble Methods: Foundations and Algorithms*; CRC Press: Boca Raton, FL, USA, 2012.
35. Mu, X.; Lu, J.; Watta, P.; Hassoun, M.H. Weighted voting-based ensemble classifiers with application to human face recognition and voice recognition. In Proceedings of the 2009 International Joint Conference on Neural Networks, Atlanta, GA, USA, 14–19 June 2009; pp. 2168–2171. [CrossRef]
36. Rieger, S.A.; Muraleedharan, R.; Ramachandran, R.P. Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers. In Proceedings of the 9th International Symposium on Chinese Spoken Language Processing, Singapore, 12–14 September 2014; pp. 589–593. [CrossRef]
37. Krajewski, J.; Batliner, A.; Kessel, S. Comparing Multiple Classifiers for Speech-Based Detection of Self-Confidence - A Pilot Study. In Proceedings of the 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 3716–3719. [CrossRef]
38. Savio, A.; García-Sebastián, M.; Chyzyk, D.; Hernandez, C.; Graña, M.; Sistiaga, A.; de Munain, A.L.; Villanúa, J. Neurocognitive disorder detection based on feature vectors extracted from VBM analysis of structural MRI. *Comput. Biol. Med.* **2011**, *41*, 600–610. [CrossRef] [PubMed]
39. Ayerdi, B.; Savio, A.; Graña, M. Meta-ensembles of classifiers for Alzheimer's disease detection using independent ROI features. In *International Work-Conference on the Interplay Between Natural and Artificial Computation, Mallorca, Spain, 10–14 June 2013*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 122–130.
40. Hammou, D.; Fezza, S.A.; Hamidouche, W. Egb: Image quality assessment based on ensemble of gradient boosting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 541–549.
41. Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; Yang, F. Musiq: Multi-scale image quality transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 5148–5157.

42. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
43. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
44. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
45. You, J.; Korhonen, J. Transformer for image quality assessment. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 1389–1393.
46. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
47. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
48. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2017**, *60*, 84–90. [CrossRef]
49. Cetinic, E.; Lipic, T.; Grgic, S. Fine-tuning Convolutional Neural Networks for fine art classification. *Expert Syst. Appl.* **2018**, *114*, 107–118. [CrossRef]
50. Kumar, A.; Kim, J.; Lyndon, D.; Fulham, M.; Feng, D. An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification. *IEEE J. Biomed. Heal. Informatics* **2016**, *21*, 31–40. [CrossRef] [PubMed]
51. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 2414–2423.
52. Storn, R.; Price, K. Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* **1997**, *11*, 341–359. [CrossRef]
53. Zhu, X.; Zhang, L.; Liu, X.; Shen, Y.; Zhao, S. GAN-Based Image Super-Resolution with a Novel Quality Loss. *Math. Probl. Eng.* **2020**, *2020*, 1–12. [CrossRef]
54. Park, S.-J.; Son, H.; Cho, S.; Hong, K.-S.; Lee, S. Srfeat: Single image super-resolution with feature discrimination. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 439–455.
55. Liu, T.; de Haan, K.; Rivenson, Y.; Wei, Z.; Zeng, X.; Zhang, Y.; Ozcan, A. Deep learning-based super-resolution in coherent imaging systems. *Sci. Rep.* **2019**, *9*, 1–13. [CrossRef] [PubMed]
56. Lugmayr, A.; Danelljan, M.; Van Gool, L.; Timofte, R. SRFlow: Learning the Super-Resolution Space with Normalizing Flow. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 715–732. [CrossRef]