Adaptive Feature Selection for No-Reference Image Quality Assessment by Mitigating Semantic Noise Sensitivity

Xudong Li [*1] **Timin Gao** [*1] **Runze Hu** [*2] **Yan Zhang** [1] **Shengchuan Zhang** [1] **Xiawu Zheng** [1] **Jingyuan Zheng** [3] **Yunhang Shen** [4] **Ke Li** [4] **Yutao Liu** [5] **Pingyang Dai** [1] **Rongrong Ji** [1]

Abstract

The current state-of-the-art No-Reference Image Quality Assessment (NR-IQA) methods typically rely on feature extraction from upstream semantic backbone networks, assuming that all extracted features are relevant. However, we make a key observation that not all features are beneficial, and some may even be harmful, necessitating careful selection. Empirically, we find that many image pairs with small feature spatial distances can have vastly different quality scores, indicating that the extracted features may contain quality-irrelevant noise. To address this issue, we propose a Quality-Aware Feature Matching IQA Metric (QFM-IQM) that employs an adversarial perspective to remove harmful semantic noise features from the upstream task. Specifically, QFM-IQM enhances the semantic noise distinguish capabilities by matching image pairs with similar quality scores but varying semantic features as adversarial semantic noise and adaptively adjusting the upstream task's features by reducing sensitivity to adversarial noise perturbation. Furthermore, we utilize a distillation framework to expand the dataset and improve the model's generalization ability. Extensive experiments conducted on eight standard IQA datasets have demonstrated the effectiveness of our proposed QFM-IQM.

[*]Equal contribution [1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China. [2]School of Information and Electronics, Beijing Institute of Technology, Beijing 100080, China. [3]School of Medicine, Xiamen University. [4]Tencent Youtu Lab. [5]School of Computer Science and Technology, Ocean University of China.. Correspondence to: Yan Zhang <bzhy986@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

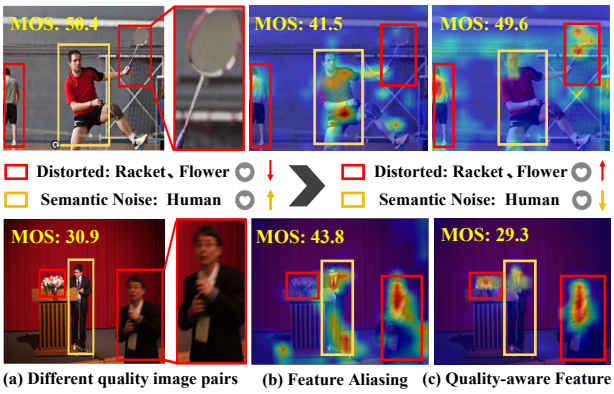

Figure 1. **Images in the first column:** the sample images in the KonIQ dataset. **Images in the medium column:** DEIQT (Qin et al., 2023) suffers from feature aliasing that focuses excessively on semantic-level noise that is less relevant to quality perception (e.g., the human in the yellow box). **Images in the last column:** Our QFM-IQAM is not sensitive to semantic noise and can clearly distinguish between image pairs with similar semantic content but different quality scores. As a result, our model's predictions are more aligned with the *Mean Opinion Scores* (MOS).

1. Introduction

No-Reference Image Quality Assessment (NR-IQA) is a fundamental research area (Ding et al., 2020; Gu et al., 2020; Zhang et al., 2015b; Hu et al., 2022; Liu et al., 2024), which simulates the human subjective system to estimate distortion of the given image. Current state-of-the-art methods (Zhang et al., 2023; Song et al., 2023) generally leverage pre-trained upstream backbone to extract semantic features, and then, finetuning on NR-IQA datasets. This pipeline provides an effective and efficient training procedure while reducing the requirement of data volume. Since the semantic features obtained from the pre-trained upstream backbone have no direct correlation with the image quality representation (Yang et al., 2020), many researchers focus on the optimization of semantic features in IQA research. For instance, (Talebi & Milanfar, 2018) proposed a model based on Deep Convolutional Neural Networks pre-trained on the ImageNet (Deng et al., 2009) and end-to-end training to optimize semantic features. (Su et al., 2020) integrated NR-IQA tasks with se-

mantic recognition networks, enabling the network to assess image quality after identifying content. (Qin et al., 2023) refined the abstract semantic information obtained from the pre-trained model by introducing a transformer decoder.

However, we empirically find that many image pairs with small feature spatial distances have vastly different quality scores (Sec. 4.6). We attribute this result to feature aliasing, where semantic noise unrelated to quality confuses the quality-aware features. Specifically, due to the sensitivity of upstream backbone networks to semantic features (Zhao et al., 2023; Zhang et al., 2023), they tend to extract similar features for semantically similar image pairs, despite these pairs having distinctly different quality scores. This may mislead the network to overly focus on semantic details rather than true indicators of image quality, resulting in poor quality prediction. We provide an example in Fig.1(b) to illustrate this issue. For a pair of images with obvious quality differences, the baseline overly focuses on similar semantic information in the distorted image (e.g., the person in the yellow box), while ignoring some distorted areas in the red box (e.g., the racket's movement and the flower's overexposure), causing the model to be misled into making a close quality predictions. These findings motivate us to explore a critical problem: how to distinguish images with similar semantic information but different quality scores.

We believe the key to addressing this challenge is to reduce the model's sensitivity to semantic noise. Inspired by adversarial learning (Goodfellow et al., 2014; Madry et al., 2017), which aims to learn adversarial noise that maximizes prediction loss, and reduces the model's sensitivity to this noise by minimizing the model's loss when facing noisy inputs. In this paper, we treat samples with similar quality scores but significantly different semantic features compared to the input image as adversarial noise. This approach is based on our hypothesis that the pre-trained upstream backbone is highly sensitive to semantics. Therefore, introducing appropriate semantic perturbation can significantly impact its quality predictions, while similar quality scores ensure that the misprediction is mainly influenced by the semantic perturbation rather than the quality perturbation.

To achieve this, we propose a novel approach called Quality-Aware Feature Matching Image Quality Metric (QFM-IQM), as shown in Fig. 1(c). Concretely, our QFM-IQM comprises three primary modules. Firstly, the Semantic Noise Feature Matching (SNM) Module is developed to pair each distorted sample with noise samples that have similar quality scores but significantly different semantic features. The Quality Consistency Constraint (QCC) Module uses noise samples as the adversarial feature perturbation to maintain quality information while disrupting semantics and ensures consistent quality predictions for distorted samples before and after semantic changes. The core objective is to enhance the model's robustness to semantic noise by minimizing the loss of quality prediction when faced with semantic noise perturbation. Lastly, the Distilled Label Expansion (DLE) Module utilizes knowledge distillation to provide pseudo-labels for unlabeled samples, enriching the dataset for QCC's adversarial learning. Our contributions include:

- We address a common challenge in IQA: how to distinguish between images that have similar semantic information but different quality scores. To solve this problem, we propose a novel model that can effectively capture the subtle differences in quality among similar images and make accurate judgments based on them.
- We introduce a novel feature adversarial learning mechanism to isolate the quality-related attributes from the semantic content of an image. This way, the model can focus on the most relevant attributes for image quality and avoid being distracted by irrelevant ones.
- Our model employs distillation learning to augment the dataset with authentic images that have varying quality levels. This not only improves the model's generalization ability but also reduces the need for human annotation efforts, which are time-consuming.

2. Related Works

Image Quality Assessment (IQA) (Hu et al., 2021) is broadly classified into Full-Reference (FR) (Cao et al., 2022), Reduced-Reference (RR) (Tao et al., 2009), and No-Reference (NR) methods. NR-IQA assesses quality independently, offering significant application potential.

2.1. NR-IQA with Vision Transformer

Vision Transformer (ViT) (Dosovitskiy et al., 2021) showed promising results on several downstream vision tasks. There were mainly two types of ViT-based NR-IQA methods, including hybrid Transformer (Golestaneh et al., 2022) and pure ViT-based Transformer (Ke et al., 2021b; Yang et al., 2022; You & Korhonen, 2021). The hybrid architecture generally combined the CNNs with the Transformer, which was responsible for the local and long-range feature characterization, respectively. For instance, (Golestaneh et al., 2022) proposed to use the multi-scale features extracted from ResNet-50, which were fed to the transformer encoder to produce a non-local representation of the image. (Ke et al., 2021b) designed a multi-scale ViT-based IQA model to handle the arbitrary size of input images. Such a feature was initially designed to describe the image content, and thus the preserved features were mainly related to the higher-level visual abstractions which were not adequate in characterizing the quality-aware features(Qin et al., 2023).

2.2. NR-IQA with Contrastive Learning

A new learning paradigm, contrastive learning, has emerged for learning discriminative representations among sam-

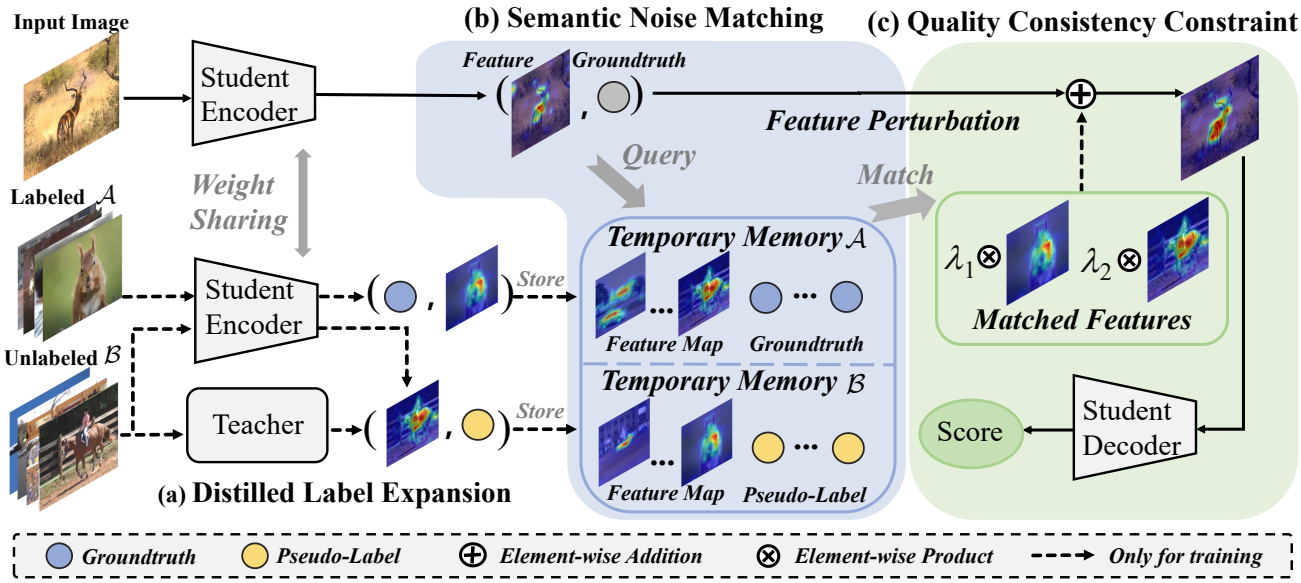

Figure 2. The overview of our QFM-IQM.We begin by extracting image semantic-related features from the student network. Simultaneously, we indirectly augment our training set by introducing a distillation framework and storing the image features and labels in pairs in temporary memory (Sec. 3.4). We use Semantic Noise Matching to select features in the temporary memory with similar quality scores but different semantic features as noise disturbance (Sec. 3.5) and add them to our original features as decoder input to predict image quality, forcing our model to learn to focus on quality-aware features (Sec. 3.6). During inference, scores are directly predicted using the encoder and decoder, while the above three modules are discarded to avoid any additional computational overhead.

ples (Caron et al., 2020; He et al., 2020). This method aims to learn an embedding space where similar samples are attracted, and dissimilar ones are repelled (Jaiswal et al., 2020; Zhao et al., 2023). Recently, contrastive learning in IQA has primarily focused on unsupervised approaches, as seen in key studies like QPT (Zhao et al., 2023), which introduced a self-supervised method with quality-aware contrastive learning for BIQA. Re-IQA (Saha et al., 2023) employed a mixture of experts to train separate encoders for content and quality features, and CONTRIQUE (Madhusudana et al., 2022) used a deep CNN with a contrastive pairwise objective for IQA. These methods have achieved promising results by leveraging self-supervised training, but the cost of pre-training models was often high. In contrast to contrastive learning aimed at constructing discriminative embedding spaces, we leverage adversarial learning (Goodfellow et al., 2014; Madry et al., 2017) to reduce the model's sensitivity to semantic noise. This enhances the model's ability to distinguish between images that are semantically similar but have different quality scores.

3. Methodology

3.1. Preliminaries

In the context of No-Reference Image Quality Assessment (NR-IQA), we define some common notations. We use bold formatting to denote vectors (e.g., $\boldsymbol{x}$, $\boldsymbol{y}$) and matrices (e.g., $\boldsymbol{X}$, $\boldsymbol{Y}$). Specifically, $\boldsymbol{F}$ denotes the feature map of the network output, which includes the VIT's class (CLS) token and patch embedding, and $\boldsymbol{y}$ denotes the quality score.

3.2. Overview

In this paper, we present Quality-Aware Feature Matching Image Quality Metric (QFM-IQM), a novel framework that can effectively distinguish between images with similar semantic features but different quality scores. As depicted in Fig. 2, QFM-IQM seamlessly integrates three main components: Semantic Noise Matching (SNM), Quality Consistency Constraint (QCC) and Distillation Label Extension (DLE). Initially, the input distorted image is fed into a transformer encoder to obtain feature $\boldsymbol{F}_o \in \mathbb{R}^{(M+1)\times D}$ from the final layer. Subsequently, the SNM module matches these features with those that have similar quality scores but different semantics (Sec. 3.5). Next, the matched semantic noise $\hat{\boldsymbol{F}}_o$ is introduced as feature perturbation into feature $\boldsymbol{F}_o$ to obtain feature $\mathcal{F}_o$, and the QCC module constraints between the quality prediction of the $\boldsymbol{F}_o$ and $\mathcal{F}_o$ to remain consistent, forcing the model to be robust to quality-unrelated semantic noise during adversarial learning (Sec. 3.6). Additionally, the DLE module uses knowledge distillation to generate pseudo-labels for unlabeled samples, thereby enriching the adversarial learning dataset of QCC (Sec. 3.4). These components collectively contribute to a more effective

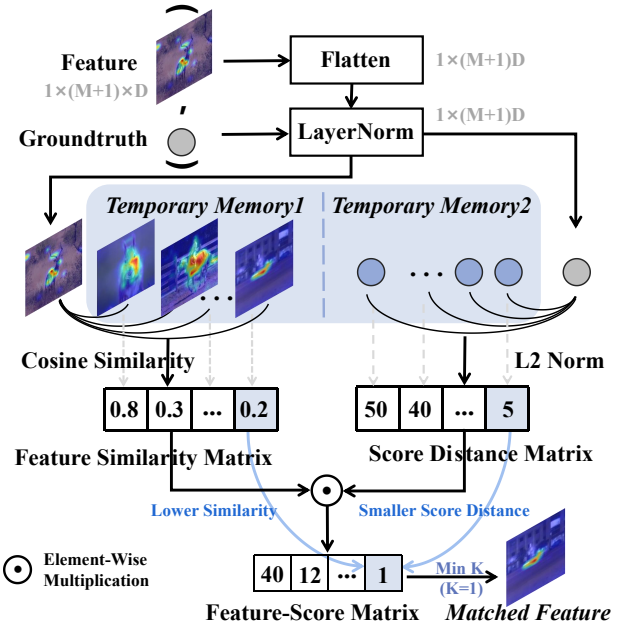

Figure 3. The Semantic Noise Matching (SNM) pipeline takes in pairs of features and labels as input and returns matching features from the dataset as output. Firstly, we flatten the input features and then normalize both the features and labels using layer norm. We then calculate the Feature Similarity Matrix and the Score Distance Matrix using the cosine similarity function and the $\mathcal{L}_2$ norm. Next, we obtain the final Feature-Score Matrix by element-wise multiplication of the computed matrices. We select the features that correspond to the index of the minimum K values of the Feature-Score Matrix in our memory as matched features.

and accurate NR-IQA process. It's worth noting that during the inference process, scores are predicted directly using the encoder and decoder, discarding the above three modules, and avoiding any additional computational overhead.

3.3. Architecture Design

Transformer Encoder. Our model uses a Transformer encoder to process input image patches. These patches are first transformed into D-dimensional embeddings, including a class token and position embeddings for spatial information. Three linear projection layers transform patch embedding into matrices $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{(M+1)\times D}$ for query, key, and value. This operation, combined with Multi-Head Self-Attention (MHSA), layer normalization, and a Multi-Layer Perceptron (MLP), produces the output feature $\boldsymbol{F}_o$.

Transformer Decoder. Previous works (Qin et al., 2023) found that CLS tokens cannot build an optimal representation for image quality. we utilize a quality-aware decoder that further interprets the CLS token. The decoder uses Multi-Head Cross-Attention (MHCA) with a set of queries Q_d, interacting with encoder outputs to calculate the final quality score $\boldsymbol{y}$. This method improves the learning and generalization capabilities of our NR-IQA model.

3.4. Distillation Label Extension

Considering the challenges transformer-based models face in learning sufficient downstream task knowledge with insufficient training data, leading to overfitting and performance degradation, this paper proposes a new distillation strategy. This strategy employs an extra unlabeled dataset $\mathcal{B}$ for feature adversarial learning during training, effectively using information from samples in the expanded dataset with similar quality scores but different features. Specifically, we first use a teacher model generating pseudo labels for each image in $\mathcal{B}$ and then store in a temporary memory $\mathcal{B}$. Then, labeled datasets $\mathcal{A}$ and $\mathcal{B}$ are input into the student model together, identifying features with similar quality scores but different semantics (Sec. 3.5). These features, treated as noise, are added to the image features, enabling the model to learn from both labeled and unlabeled data and improve its generalization ability for images with different quality.

3.5. Semantic Noise Matching

The Semantic Noise Matching (SNM) Module, comprising Temporary Memory and Semantic Noise Feature Matching components, functions by aligning input image features with Temporary Memory samples in each training batch, efficiently matching noise samples.

Temporary Memory. The objective of the Temporary Memory is to store features and labels of each mini-batch of data, serving as a queryable resource for the SNM module. Specifically, the student feature extractor conducts forward propagation to extract features of distorted images from labeled dataset $\mathcal{A}$ and stores these features alongside the ground truth of the distorted images in Temporary Memory $\mathcal{A}$. Concurrently, the student extracts features of distorted images from unlabelled dataset $\mathcal{B}$, storing them with the pseudo labels of these images in Temporary Memory $\mathcal{B}$. It is important to note that during the training process, the original memory data is overwritten in each iteration of every mini-batch to reset the Temporary Memory.

Semantic Noise Feature Matching. This process involves selecting images with similar quality but different semantics to use as semantic noise features. Then, these noise features are employed as feature perturbations to disrupt semantics while largely preserving quality information. Finally, under the constraint of quality consistency, this effectively reduces sensitivity to semantic noise (referenced in Sec. 3.6). Fig. 3 overviews our Semantic Noise Matching (SNM) module. For a mini-batch of distorted input images, Temporary Memory $\mathcal{A}$ and $\mathcal{B}$ store collections of features and labels, then QFM-IQM leverages the SNM module to select noise samples that have similar quality scores but different semantic features from both labeled and unlabeled datasets. Specifically, the process begins by flattening the feature $\boldsymbol{F}_o$ of the input image, obtained from the encoder, into a fea-

ture vector $\boldsymbol{f}_o$. Subsequently, another feature vector $\hat{\boldsymbol{f}}_o$ is matched from the Temporary Memory $\mathcal{A}$ or $\mathcal{B}$ based on the matching strategy. The evaluation metric for the matching strategy is formulated as the product of feature similarity and score distance. Feature similarity is calculated using cosine distance, while score distance is determined using the $\mathcal{L}_2$ norm. The mathematical expression is as follows:

$$\mathcal{S}((\boldsymbol{f}_o, \mathrm{y_o}), (\hat{\boldsymbol{f}}_o, \hat{\mathrm{y}}_o)) = \frac{\boldsymbol{f}_o \cdot \hat{\boldsymbol{f}}_o}{\|\boldsymbol{f}_o\| \cdot \|\hat{\boldsymbol{f}}_o\|} \cdot \|\mathrm{y_o} - \hat{\mathrm{y}}_o\|. \quad (1)$$

Here, $\mathrm{y_o}$ and $\hat{\mathrm{y}}_o$ represent the quality score of the sample, $\cdot$ represents the dot product between the vectors, and $\|\cdot\|$ represents their $\mathcal{L}_2$ norms. This score is then used as a criterion for the final selection strategy, choosing the top K minimum score features selected as matching samples for the following QCC module's input.

3.6. Quality Consistency Constraint

Existing studies have not addressed the challenge of differentiating between images with similar semantic features but different quality scores and eliminating quality-irrelevant features from high-level pre-trained features. To address this, we introduce the Quality Consistency Constraint (QCC) module to isolate quality-related attributes from an image's semantic content. In contrast to previous contrastive learning approaches, this paper introduces a straightforward yet effective adversarial method to enhance a model's robustness to features unrelated to image quality. This is accomplished by strengthening the consistency of the model's predictions for image quality before and after introducing noise to distorted images. This method reduces the model's sensitivity to semantic noise and compels it to exhibit robustness to subtle variations, ultimately increasing the separation between feature representations of noise samples. Concretely, we first extract the K features with the smallest evaluation metric calculated in the SNM module and incorporate them as feature perturbations into the original features. Since we perform feature selection twice in labeled dataset $\mathcal{A}$ and unlabelled dataset $\mathcal{B}$, the expression of this process is:

$$\mathcal{F}_o = (1-\lambda_1-\lambda_2)\boldsymbol{F}_o + \frac{\lambda_1}{K}\sum_{j=1}^{K}\hat{\boldsymbol{F}}^{\mathcal{A}}_{o,j} + \frac{\lambda_2}{K}\sum_{j=1}^{K}\hat{\boldsymbol{F}}^{\mathcal{B}}_{o,j}, \quad (2)$$

where the weight λ_1 and λ_2 controls the strength of the perturbation. $\hat{\boldsymbol{F}}^{\mathcal{A}}_{o,j}$ and $\hat{\boldsymbol{F}}^{\mathcal{B}}_{o,j}$ represent the jth smallest matching feature from the datasets $\mathcal{A}$ and datasets $\mathcal{B}$. After that, we feed the feature $\mathcal{F}_o$ which is obtained by the Eq. 2 into the decoder to estimate the quality score $\boldsymbol{y}$. Notably, even after adding noise, the predicted quality is still supervised using the ground truth of the original image. By minimizing the prediction loss on samples with semantic noise, the model's robustness to the semantic noise is improved.

4. Experiments

4.1. Benchmark Datasets and Evaluation Protocols

We evaluate the QFM-IQM on 8 standard NR-IQA datasets. These include four synthetic datasets: LIVE (Sheikh et al., 2006), CSIQ (Larson & Chandler, 2010), TID2013 (Ponomarenko et al., 2015), and KADID (Lin et al., 2019); and four authentic datasets: LIVEC (Ghadiyaram & Bovik, 2015), KonIQ (Hosu et al., 2020), LIVEFB (Ying et al., 2020), and SPAQ (Fang et al., 2020). Synthetic datasets are generated by applying distortions like JPEG compression and Gaussian blur to original images. LIVE and CSIQ contain 779 and 866 images with five and six distortion types, respectively, while TID2013 and KADID have 3000 and 10,125 images with 24 and 25 distortion types. The authentic datasets contain images captured by different photographers using various mobile devices, with LIVEC comprising 1162 images, SPAQ 11,125 images, and KonIQ 10,073 images sourced from public multimedia resources. LIVEFB is the largest authentic dataset with 39,810 images. The performance of the QFM-IQM model, in terms of prediction accuracy and monotonicity, is assessed using Spearman's Order Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC). Both SRCC and PLCC values range from -1 to 1, with values near 1 denoting higher performance in both metrics.

4.2. Implementation Details

Pre-trained Teacher. Our teacher network, comprising a VIT-S and a decoder, is pre-trained on the largest authentic dataset LIVEFB for offline knowledge distillation on students trained on the IQA dataset. The pre-trained strategy follows the (Qin et al., 2023). Notably, when the student model is trained on the LIVEFB dataset, the teacher model is pre-trained on KonIQ to ensure fairness.

Student Training. To train the student network, we follow the standard approach of cropping input images into ten patches of 224 $\times$ 224 resolution, reshaped into smaller 16 $\times$ 16 patches with a 384-dimensional input token. Using a Transformer encoder based on ViT-S from DeiT III (Touvron et al., 2022), our model has 12 layers and 6 heads, coupled with a single-layer decoder. Training is conducted over 9 epochs with a learning rate of 2×10^{-4}, reducing by a factor of 10 every 3 epochs, using the Adamw optimizer (Loshchilov & Hutter, 2019). The batch size depends on the size of the dataset, which ranges from 16 to 128. For each dataset, 80% of the images are used for training and the remaining 20% of the images are used for testing. This process is repeated ten times to minimize bias. For synthetic distortion datasets, training and testing sets are divided by reference images for content independence. The model's performance is quantified by the average SRCC and PLCC, measuring prediction accuracy and monotonicity.

Table 1. Performance comparison of average SRCC and PLCC, with bold indicating the best results and underlines for the second-best.

	LIVE		CSIQ		TID2013		KADID		LIVEC		KonIQ		LIVEFB		SPAQ	
Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
DIIVINE (Saad et al., 2012)	0.908	0.892	0.776	0.804	0.567	0.643	0.435	0.413	0.591	0.588	0.558	0.546	0.187	0.092	0.600	0.599
BRISQUE (Mittal et al., 2012)	0.944	0.929	0.748	0.812	0.571	0.626	0.567	0.528	0.629	0.629	0.685	0.681	0.341	0.303	0.817	0.809
ILNIQE (Zhang et al., 2015a)	0.906	0.902	0.865	0.822	0.648	0.521	0.558	0.534	0.508	0.508	0.537	0.523	0.332	0.294	0.712	0.713
BIECON (Kim & Lee, 2016)	0.961	0.958	0.823	0.815	0.762	0.717	0.648	0.623	0.613	0.613	0.654	0.651	0.428	0.407	-	-
MEON (Ma et al., 2017)	0.955	0.951	0.864	0.852	0.824	0.808	0.691	0.604	0.710	0.697	0.628	0.611	0.394	0.365	-	-
WaDIQaM (Bosse et al., 2017)	0.955	0.960	0.844	0.852	0.855	0.835	0.752	0.739	0.671	0.682	0.807	0.804	0.467	0.455	-	-
DBCNN (Zhang et al., 2018)	0.971	0.968	0.959	0.946	0.865	0.816	0.856	0.851	0.869	0.851	0.884	0.875	0.551	0.545	0.915	0.911
MetaIQA (Zhu et al., 2020)	0.959	0.960	0.908	0.899	0.868	0.856	0.775	0.762	0.802	0.835	0.887	0.850	0.507	0.54	-	-
P2P-BM (Ying et al., 2020)	0.958	0.959	0.902	0.899	0.856	0.862	0.849	0.84	0.842	0.844	0.885	0.872	0.598	0.526	-	-
HyperIQA (Su et al., 2020)	0.966	0.962	0.942	0.923	0.858	0.840	0.845	0.852	0.882	0.859	0.917	0.906	0.602	0.544	0.915	0.911
TReS (Golestaneh et al., 2022)	0.968	0.969	0.942	0.922	0.883	0.863	0.858	0.859	0.877	0.846	0.928	0.915	0.625	0.554	-	-
MUSIQ (Ke et al., 2021a)	0.911	0.940	0.893	0.871	0.815	0.773	0.872	0.875	0.746	0.702	0.928	0.916	0.661	0.566	0.921	0.918
DACNN (Pan et al., 2022)	0.980	0.978	0.957	0.943	0.889	0.871	0.905	0.905	0.884	0.866	0.912	0.901	-	-	0.921	0.915
DEIQT (Qin et al., 2023)	0.982	0.980	0.963	0.946	0.908	0.892	0.887	0.889	0.894	0.875	0.934	0.921	0.663	0.571	0.923	0.919
Re-IQA (Saha et al., 2023)	0.971	0.970	0.960	0.947	0.861	0.804	0.885	0.872	0.854	0.840	0.923	0.914	**0.733**	**0.645**	**0.925**	0.918
QFM-IQM (Ours)	**0.983**	**0.981**	**0.965**	**0.954**	**0.932**	**0.916**	**0.906**	**0.906**	**0.913**	**0.891**	**0.936**	**0.922**	0.667	0.567	0.924	**0.920**

Table 2. SRCC on the cross datasets validation. The best performances are highlighted in boldface.

Training	LIVEFB		LIVEC	KonIQ	LIVE	CSIQ
Testing	KonIQ	LIVEC	KonIQ	LIVEC	CSIQ	LIVE
DBCNN	0.716	0.724	0.754	0.755	0.758	0.877
P2P-BM	0.755	0.738	0.74	0.77	0.712	-
HyperIQA	0.758	0.735	0.772	0.785	0.744	0.926
TReS	0.713	0.74	0.733	0.786	0.761	-
Re-IQA	-	-	0.769	0.791	0.808	0.929
DEIQT	0.733	0.781	0.744	0.794	0.781	0.932
QFM-IQM	**0.768**	**0.791**	**0.775**	**0.796**	**0.820**	**0.941**

4.3. Overall Prediction Performance Comparison

The results of the comparison between QFM-IQM and 15 classical or state-of-the-art (SOTA) NR-IQA methods, which include hand-crafted feature-based NR-IQA methods like ILNIQE (Zhang et al., 2015a) and BRISQUE (Mittal et al., 2012), as well as deep learning-based methods such as MUSIQ (Ke et al., 2021a) and DEIQT (Qin et al., 2023), are presented in Table 1. It can be observed on six of the eight datasets that QFM-IQM outperforms all other methods in terms of performance. Achieving competitive performance on all of these datasets is a challenging task due to the wide range of image content and distortion types. Therefore, these observations confirm the effectiveness of QFM-IQM in accurately characterizing image quality.

4.4. Generalization Capability Validation

To assess the generalization capabilities of QFM-IQM, we conduct a series of cross-dataset validation experiments. In these tests, QFM-IQM trains on one dataset and then evaluates it on different datasets without adjustments to its parameters. The results, presented in Table 2, demonstrate QFM-IQM's superior performance over SOTA models across six datasets. This includes notable gains on the LIVEC dataset and strong results on the KonIQ dataset. These results, benefiting from the noise-resistant training provided by the SNM and QCC, as well as the extensive semantic noise of various distortions provided by the DLE, underscore QFM-IQM's exceptional ability to generalize.

4.5. Ablation Study

The core idea of our QFM-IQM is to improve the model's ability to distinguish subtle quality differences by enhancing its robustness to semantic noise. Therefore, we explore the impact of Quality Consistency Constraint (QCC) operation on the model when compared with different datasets in Table 3. Let $\mathcal{A}^{QCC}$ denotes the QCC operation only compare with the labeled dataset $\mathcal{A}$, and $\mathcal{B}^{QCC}$ denotes the QCC operation only compare with the unlabeled dataset $\mathcal{B}$. In the ablation experiments on QCC, we use Semantic Noise Matching (SNM) by default.

Effective of QCC on labeled dataset $\mathcal{A}$. We explore the effectiveness of the QCC module on the original training dataset. After using QCC for training, performance on three datasets improves to varying degrees. Particularly on the LIVEC authentic dataset, there is a significant improvement of 2.1 points, and the variance is significantly reduced. However, we find that on the TID2013 synthetic dataset, although there is a significant improvement of 2 points, the stability of the model decreases. We analyze this as due to the synthetic dataset having less rich semantic information than the authentic dataset, causing our QCC module to force the model to ignore the quality-related semantic information, thereby weakening its stability.

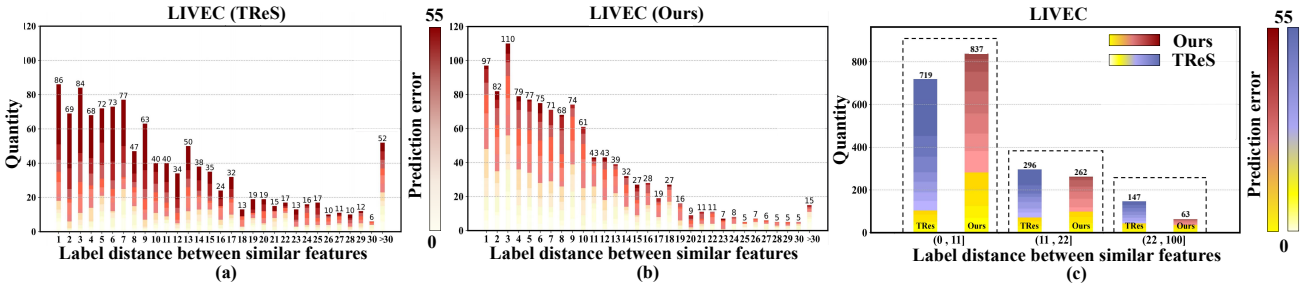

Figure 4. Qualitative analysis on image pairs have similar semantic features but different quality scores. The bar charts in (a) and (b) display the number of image pairs with the **most similar** features within a specific label distance range. For TReS, there are many image pairs with similar features but large differences in quality scores (even with label distance greater than 30, TReS considers 52 samples as feature-similar). Our model greatly alleviates this issue and achieves a lead in accuracy (a large number of light yellow areas represent smaller prediction errors). In bar char (c), we integrate the statistical results of (a) and (b) to make the above conclusions clearer.

Table 3. Ablation experiments on TID2013, LIVEC and KonIQ datasets. Here, $\mathcal{A}^{QCC}$ and $\mathcal{B}^{QCC}$ refer to the Quality Consistency Constraint without and with the Distillation Label Extension module, where bold entries indicate the best result.

Index	$\mathcal{A}^{QCC}$	$\mathcal{B}^{QCC}$	TID2013		LIVEC		KonIQ	
			PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
a)			0.899 (±0.017)	0.891 (±0.028)	0.883 (±0.012)	0.865 (±0.021)	0.926 (±0.003)	0.913 (±0.003)
b)	✔		0.919 (±0.022)	0.901 (±0.026)	0.904 (±0.009)	0.885 (±0.011)	0.934 (±0.001)	0.922 (±0.002)
c)		✔	0.925 (±0.016)	0.910 (±0.022)	0.907 (±0.006)	0.882 (±0.010)	0.935 (±0.002)	0.921 (±0.002)
d)	✔	✔	**0.932 (±0.015)**	**0.916 (±0.017)**	**0.913 (±0.007)**	**0.891 (±0.009)**	**0.936 (±0.002)**	**0.922 (±0.003)**

Effective of QCC on unlabeled dataset $\mathcal{B}$. We further explore the effectiveness of the QCC module on the expanded unlabeled training dataset. Compared to $\mathcal{A}^{QCC}$, $\mathcal{B}^{QCC}$ achieves more significant performance improvements. Additionally, the issue of decreased stability on the TID2013 dataset is rectified, ensuring stability across all datasets. This indicates that using additional pseudo-labels to expand our dataset allows our QFM-IQM to learn from images with different quality levels and feature distributions. Consequently, our model can distinguish quality-related features through more extensive data comparisons. The combination of $\mathcal{A}^{QCC}$ and $\mathcal{B}^{QCC}$ leads to a superior model for IQA.

Effective on SNM and QCC. We broaden our investigation to include ablation studies on SNM and QCC mechanisms. Specifically, we undertake two sets of SNM ablation experiments: the first set, SNM^F, matches pairs of samples dissimilar in features, while the second set, SNM^Q, matches pairs similar in quality. Additionally, an extra ablation study without using the QCC involves matching noise samples $(\boldsymbol{x}_1, \boldsymbol{y}_1)$ and $(\boldsymbol{x}_2, \boldsymbol{y}_2)$ to a given pair $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{y}$ is the quality label, adjusting the quality labels by a specific ratio ($\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = 0.1$) to incorporate noise sample labels, formulated as $\boldsymbol{y} = (1 - \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\boldsymbol{y} + \boldsymbol{\beta}_1\boldsymbol{y}_1 + \boldsymbol{\beta}_2\boldsymbol{y}_2$.

Our findings underscore the critical role of QCC's constraints because even slight changes in mixing ratios may significantly alter the perceived quality category, potentially misleading the model during the supervision process. Moreover, the SNM mechanism emphasizes the importance of matching samples with similar quality but significantly different features because matching such noise samples can lead to the model making the largest errors in quality prediction. Therefore, considering only one aspect in the matching process may result in suboptimal solutions and fail to effectively improve the model's robustness to noise samples.

Table 4. Ablation experiments on LIVEC and TID2013 datasets. Here, SNM^Q and SNM^F refer to the SNM module only matching pairs with similar quality and dissimilar features, respectively.

Component			LIVEC		TID2013	
SNM^Q	SNM^F	QCC	PLCC	SRCC	PLCC	SRCC
			0.883	0.865	0.899	0.891
✔			0.897	0.880	0.918	0.899
	✔		0.898	0.879	0.916	0.897
✔	✔		0.901	0.882	0.921	0.907
✔		✔	0.906	0.885	0.922	0.902
	✔	✔	0.904	0.884	0.920	0.900
✔	✔	✔	**0.913**	**0.891**	**0.932**	**0.916**

Ablation study of unlabeled dataset selection. We explore using the synthetic dataset KADID (10,125 images) and the authentic datasets KonIQ (10,073 images) and LIVEFB (39,810 images) as unlabeled data sources, as summarized in Table 4. The results indicate more significant improve-

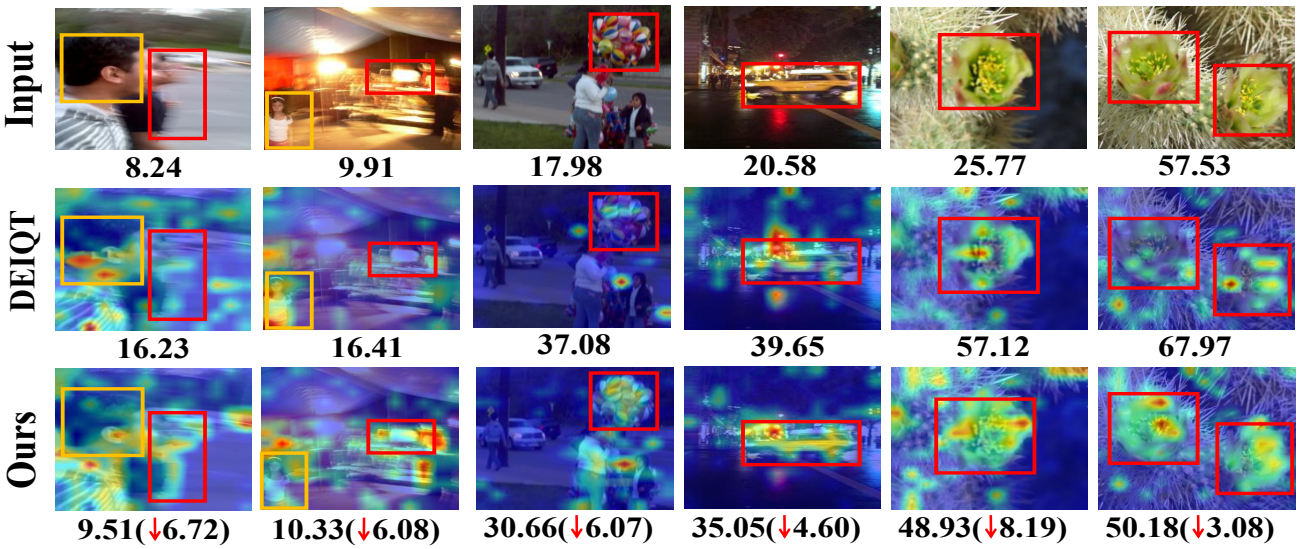

Figure 5. Activation maps of DEIQT (Qin et al., 2023) and QFM-IQM using Grad-CAM (Selvaraju et al., 2017) on authentic dataset LIVEC. The scores below the first row of images represent *Mean Opinion Scores* (MOS). Our model pays more attention to image distortion regions (red boxes) and is less sensitive to semantic noise (yellow boxes), and correspondingly our image quality prediction ability is closer to the true value. Rows 1-3 represent input images, CAMs from DEIQT, and CAMs from QFM-IQM, respectively.

Table 5. Ablation experiments on the LIVEC and TID2013 datasets for unlabeled data selection in the DLE module.

Unlabeled Dataset	LIVEC		TID2013	
	PLCC	SRCC	PLCC	SRCC
-	0.904	0.885	0.919	0.901
KADID	0.904	0.887	0.921	0.899
KonIQ	0.912	0.889	0.928	0.913
LIVEFB	**0.913**	**0.891**	**0.932**	**0.916**

Table 6. Analysis of the K in memory searching on both authentic and synthetic datasets. Bold entries indicate the best performance.

K-NN	LIVE		LIVEC	
	PLCC	SRCC	PLCC	SRCC
K=1	**0.983**	**0.981**	0.910	0.887
K=2	0.981	0.978	0.905	0.887
K=3	0.982	0.979	0.900	0.880
K=4	0.982	0.980	0.904	0.885
K=6	0.981	0.978	**0.913**	**0.891**
K=8	0.978	0.976	0.903	0.877

ments with richer unlabeled datasets (e.g., on the TID2013 dataset, LIVEFB improves by 1.3%, whereas KADID and KonIQ improve by 0.2% and 0.9%, respectively). We draw two conclusions: (1) Increasing the volume of unlabeled data enhances performance. As data becomes more diverse, unlabeled datasets provide a wider variety of adversarial noise samples, compelling the model to improve its robustness against a broader range of noise features. (2) Authentic datasets offer more significant benefits than synthetic datasets. Specifically, synthetic datasets are often generated by applying various degrees of distortion to a single original image. For instance, the KADID dataset, with 10,125 distorted images, contains only 81 unique content images, implying that synthetic datasets offer limited semantic information. Consequently, models struggle to effectively learn robustness against semantic noise from synthetic datasets.

4.6. Qualitative Analysis

Analyze similar features that have different quality. In this section, we illustrate our motivation by counting the number of most similar features extracted by the TReS and QFM-IQM models from image pairs with different quality score differences. As shown in Fig. 4, the vertical axis represents quantity, and the horizontal axis represents the range of label differences. According to Fig. 4(a), we observe that SOTA pre-trained models like TReS struggle to distinguish feature distances between image pairs with different quality scores. Specifically, in Fig. 4(a), TReS incorrectly identifies 52 image pairs with a quality score difference greater than 30 as having very similar features. In this paper, We believe this is due to feature confusion, where the model is misled by similar semantics, causing it to perceive similar features in images with different quality scores. This leads to incorrect quality predictions. Our approach mitigates this issue by reducing the model's sensitivity to semantic noise. Fig. 4(c) summarizes (a) and (b), showing our model's effectiveness. Compared to TReS, the QFM-IQM model reduces the number of image pairs with similar features but different scores by approximately 56%, while increasing pairs with close scores and similar features by about 16%. This

Table 7. Sensitivity study of hyperparameters λ_1 on LIVE, LIVEC, and KonIQ datasets. Bold entries indicate the best performance.

	LIVE		LIVEC		KonIQ	
λ_1	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
E-1	0.978	0.976	0.884	0.857	0.925	0.911
E-2	0.979	0.976	0.877	0.865	0.934	0.919
E-4	0.979	0.979	0.897	0.875	0.931	0.916
E-5	0.980	0.978	0.904	0.884	0.932	0.917
E-7	**0.983**	**0.981**	**0.910**	**0.887**	**0.936**	**0.922**
E-8	0.981	0.979	0.902	0.886	0.933	0.917

Table 8. Sensitivity study of hyperparameters λ_2 on LIVEC and TID2013 datasets. Bold entries indicate the best performance.

	LIVEC		TID2013	
λ_2	PLCC	SRCC	PLCC	SRCC
E-1	0.909	0.890	0.914	0.893
E-2	0.902	0.884	0.918	0.906
E-4	0.911	**0.893**	0.927	0.911
E-5	0.912	**0.893**	**0.932**	0.914
E-7	**0.913**	0.891	**0.932**	**0.916**
E-8	0.905	0.885	0.918	0.900

indicates a better positive correlation between features and quality scores, enhancing overall prediction performance. Additionally, QFM-IQM shows fewer high prediction errors (fewer instances in red), further proving its effectiveness. In conclusion, our model emphasizes quality-related features more than TReS, enhancing quality-aware features and significantly reducing quality estimation bias.

Visualization of attention map. We utilize GradCAM (Selvaraju et al., 2017) to visualize the feature attention maps in Fig. 5. The results show that our model accurately and comprehensively focuses on the distorted regions (including both quality-related semantic degradations and low-level artifacts), while DEIQT (Qin et al., 2023) fails to accurately pinpoint the true areas of quality degradation. This indicates that our model excels in learning the subtle differences between semantic structures and quality-aware features. We achieve this by filtering out features irrelevant to quality from the pre-trained model, focusing on learning those quality-aware features. It is worth noting that for overall blurry images, such as those in the third column, our model tends to focus on subject areas that are inherently more prone to becoming blurry, such as floating balloons.

4.7. Analysis of the K in Memory Searching

In this analysis, we examine the effect of the memory searching operation's searching number, K. We conduct experiments on both synthetic and authentic datasets, LIVE and LIVEC, with K values of 1, 2, 3, 4, 6, and 8. The results, as shown in Table 6, demonstrate that the performance on holistic IQA datasets remains stable with various K values, with only a slight fluctuation within 1%. Specifically, in the synthetic dataset LIVE, which comprises only 30 distinct content images, the types of distortions (like Gaussian blur and JPEG) are generally uniform across the whole image. Thus, the model receives relatively less contamination at the semantic level, and setting K to 1 achieves the best results. On the other hand, in the authentic dataset LIVEC, all the training data contain different contents, leading to various semantic level degradations. Therefore, effectively utilizing quality perception features from pre-trained semantic features is particularly important, and we find that the best performance is achieved by setting the K value to 6. However, setting the K value too high can be detrimental. Excessive noise can distort the original features of the image that the model has learned.

4.8. Sensitivity study of hyperparameters

In this paper, we use λ_1 and λ_2 in Eq. 2 to balance the contrastive learning and distillation, respectively. To this end, we conduct a sensitivity study on different feature perturbation weights λ_1 and λ_2 to explore the effect of QCC, maintaining the search number K at 1. As shown in Table 7 and Table 8, our findings reveal that the QCC module has relative sensitivity to the hyperparameters. Specifically, small values of λ weaken our QCC's impact, whereas large values of λ cause significant changes in the feature space and result in performance degradation.

5. Conclusion

In this paper, we introduce QFM-IQM, which tackles the challenge of differentiating images with similar semantics but varying quality scores. This is accomplished through the SNM module, which selects images with similar quality scores but notable feature differences as perturbations. The QCC module reduces the model's sensitivity to these perturbations, enabling the model to isolate quality-related attributes from an image's semantic content. Additionally, QFM-IQM uses distillation learning to enhance the dataset with authentic images, improving the model's generalization across different scenarios. Experiments on eight benchmark IQA datasets demonstrate the effectiveness of this approach.

Acknowledgements

This work was supported by National Science and Technology Major Project (No. 2022ZD0118202), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001).

Impact Statement

This paper presents work whose goal is to advance the field of Image quality assessment. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., and Samek, W. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017.

Cao, Y., Wan, Z., Ren, D., Yan, Z., and Zuo, W. Incorporating semi-supervised and positive-unlabeled learning for boosting full reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5851–5861, 2022.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Fang, Y., Zhu, H., Zeng, Y., Ma, K., and Wang, Z. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3677–3686, 2020.

Ghadiyaram, D. and Bovik, A. C. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015.

Golestaneh, S. A., Dadsetan, S., and Kitani, K. M. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1220–1230, 2022.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gu, S., Bao, J., Chen, D., and Wen, F. Giqa: Generated image quality assessment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 369–385. Springer, 2020.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

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

Hu, R., Yang, R., Liu, Y., and Li, X. Simulation and mitigation of the wrap-around artifact in the mri image. *Frontiers in Computational Neuroscience*, 15:746549, 2021.

Hu, R., Monebhurrun, V., Himeno, R., Yokota, H., and Costen, F. An uncertainty analysis on finite difference time-domain computations with artificial neural networks: improving accuracy while maintaining low computational costs. *IEEE Antennas and Propagation Magazine*, 65(1): 60–70, 2022.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Ke, J., Wang, Q., Wang, Y., Milanfar, P., and Yang, F. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021a.

Ke, J., Wang, Q., Wang, Y., Milanfar, P., and Yang, F. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021b.

Kim, J. and Lee, S. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11 (1):206–220, 2016.

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

Lin, H., Hosu, V., and Saupe, D. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3. IEEE, 2019.

Liu, Y., Zhang, B., Hu, R., Gu, K., Zhai, G., and Dong, J. Underwater image quality assessment: Benchmark database and objective method. *IEEE Transactions on Multimedia*, 2024.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., and Zuo, W. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2017.

Madhusudana, P. C., Birkbeck, N., Wang, Y., Adsumilli, B., and Bovik, A. C. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mittal, A., Moorthy, A. K., and Bovik, A. C. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

Pan, Z., Zhang, H., Lei, J., Fang, Y., Shao, X., Ling, N., and Kwong, S. Dacnn: Blind image quality assessment via a distortion-aware convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7518–7531, 2022.

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 processing: Image communication*, 30:57–77, 2015.

Qin, G., Hu, R., Liu, Y., Zheng, X., Liu, H., Li, X., and Zhang, Y. Data-efficient image quality assessment with attention-panel decoder. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, 2023.

Saad, M. A., Bovik, A. C., and Charrier, C. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012.

Saha, A., Mishra, S., and Bovik, A. C. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5846–5855, 2023.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Sheikh, H. R., Sabir, M. F., and Bovik, A. C. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006.

Song, T., Li, L., Cheng, D., Chen, P., and Wu, J. Active learning-based sample selection for label-efficient blind image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., and Zhang, Y. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3667–3676, 2020.

Talebi, H. and Milanfar, P. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8): 3998–4011, 2018.

Tao, D., Li, X., Lu, W., and Gao, X. Reduced-reference iqa in contourlet domain. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(6):1623–1627, 2009.

Touvron, H., Cord, M., and Jégou, H. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022.

Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., and Yang, Y. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1191–1200, 2022.

Yang, X., Li, F., and Liu, H. Ttl-iqa: Transitive transfer learning based no-reference image quality assessment. *IEEE Transactions on Multimedia*, 23:4326–4340, 2020.

Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., and Bovik, A. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3575–3585, 2020.

You, J. and Korhonen, J. Transformer for image quality assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 1389–1393. IEEE, 2021.

Zhang, L., Zhang, L., and Bovik, A. C. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015a.

Zhang, P., Zhou, W., Wu, L., and Li, H. Som: Semantic obviousness metric for image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2394–2402, 2015b.

Zhang, W., Ma, K., Yan, J., Deng, D., and Wang, Z. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018.

Zhang, W., Zhai, G., Wei, Y., Yang, X., and Ma, K. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14071–14081, 2023.

Zhao, K., Yuan, K., Sun, M., Li, M., and Wen, X. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22302–22313, 2023.

Zhu, H., Li, L., Wu, J., Dong, W., and 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*, pp. 14143–14152, 2020.

A. Appendix Overview

The supplementary material is organized as follows: **Qualitative Analysis** provides more analysis of the effectiveness of the QFM-IQM and is further illustrated in conjunction with our motivation. **Training and Evaluation Details:** shows more training and evaluation details. **Comparison with CONTRIQUE**: provides more analysis of existing methods CONTRIQUE (Madhusudana et al., 2022) based on contrastive learning.

B. Qualitative Analysis

In this section, we provide an additional explanation about the content of Fig. 4 to help readers understand our motivation. Our motivation stems from a phenomenon we discovered through experimentation: existing state-of-the-art (SOTA) models based on pre-training, such as TReS, struggle to adequately distinguish the feature distances between image pairs with different quality scores. This results in inaccurate quality predictions. As shown in Fig. 4(a), TReS mistakenly identifies image pairs with a quality score difference greater than 30 as having very similar features. We believe this may be due to feature confusion, where current models are misled by the similar semantics of the image pairs, leading them to perceive similar features in images with different quality scores, thereby failing to adequately separate their features. This results in erroneous quality predictions. Our method addresses this feature confusion by specifically designing for it. We select noisy samples with different semantic features but similar quality scores, and through noise perturbation, force the model to reduce its sensitivity to semantic noise. This effectively increases the feature distance between image pairs with similar semantics but different quality scores. As illustrated in Fig. 4(b), our QFM-IQM significantly mitigates the aforementioned issue. Specifically, it alleviates the situation where image pairs with a quality score difference greater than 30 still share very similar features, and the correlation between features and quality scores is more positively aligned, meaning similar quality scores have similar quality features. Additionally, the error in our quality predictions is smaller, further demonstrating the effectiveness of our approach.

C. Training and Evaluation Details

Efficacy of subcomponents in the training phase: Our **Distilled Label Expansion** augments the dataset using authentic images of different quality levels on LIVEFB and KonIQ to overcome the lack of sufficient training data in IQA. Our **Semantic Noise Feature Matching** compares the features of input with other features to match similar quality features but different semantics (Fig.3 of the main paper). These matched features are considered noises to be added to the original features in the process of **Quality Consistency Constraint**. Our Quality Consistency Constraint constrains the model to still be supervised by the ground truth for the quality prediction of the noise-added samples. This enables the model to emphasize the subtle differences among distortion images, while concurrently reducing the significance placed on semantic aspects.

Unlabeled data: Our unlabeled data comes from the LIVEFB and KonIQ datasets, but we do not use the label information provided by these datasets during training. During training, we randomly sample a mini-batch of unlabeled images for training. **Implement details of training phrase:**

Table 9. Training preprocessing details of selected BIQA datasets.

Dataset	Resolution	Resize	Batch Size	Label Range
LIVE	768×512	512×384	12	DMOS [0,100]
CSIQ	512×512	512×512	12	DMOS [0,1]
TID2013	512×384	512×384	48	MOS [0,9]
KADID	512×384	512×384	128	MOS [1,5]
LIVEC	$500P \sim 640P$	$500P \sim 640P$	16	MOS [1,100]
KonIQ	$768P$	512×384	128	MOS [0,5]
LIVEFB	$160P \sim 700P$	512×512	128	MOS [0,100]
SPAQ	$1080P \sim 4368P$	512×384	128	MOS [0,100]

To train the student network, we adopt a standard approach of randomly cropping input images into ten patches, each with a 224×224 resolution. Subsequently, we reshape these patches into a sequence of smaller patches with a patch size of p = 16 and an input token dimension of D = 384. Furthermore, we present additional training preprocessing details for various datasets, which are not included in the main paper, in Table 9. For different benchmarks, we employ different training settings for a fair comparison. **Teacher performance:** We ensured the high accuracy of the pre-trained teacher model on the dataset to provide reliable pseudo-labels, as shown in Table 10.

Table 10. Performance comparison measured by averages of SRCC and PLCC, where bold entries indicate the best results.

	KonIQ		LIVEFB	
Method	PLCC	SRCC	PLCC	SRCC
QFM-IQM	**0.936**	**0.922**	**0.667**	**0.567**
Teacher	0.930	0.914	**0.667**	0.566

Inference phrase: During the inference process, we only employ a trained encoder-decoder architecture to directly assess the image's quality which benefits from the QFM-IQM to precisely isolate quality-aware features within the pre-trained semantic feature space, thereby enhancing the accuracy of our quality predictions.

Table 11. Performance comparison measured by averages of SRCC and PLCC, where bold entries indicate the best results.

	LIVE		CSIQ		TID2013		KADID		LIVEC		KonIQ		LIVEFB		SPAQ	
Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
CONTRIQUE	0.961	0.960	0.955	0.942	0.857	0.843	**0.937**	**0.934**	0.857	0.845	0.906	0.894	0.641	0.580	0.919	0.914
QFM-IQM(ours)	**0.983**	**0.981**	**0.965**	**0.954**	**0.932**	**0.916**	0.906	0.906	**0.913**	**0.891**	**0.936**	**0.922**	**0.667**	**0.567**	**0.924**	**0.920**

Table 12. SRCC on the cross datasets validation. The best results are highlighted in bold.

Training	Testing	CONTRIQUE	QFM-IQM
LIVEC	KonIQ	0.676	**0.775**
KonIQ	LIVEC	0.731	**0.796**
LIVE	CSIQ	**0.823**	0.820
CSIQ	LIVE	0.925	**0.941**

D. Comparison with CONTRIQUE

In CONTRIQUE, the approach revolves around treating the IQA representation learning problem as a classification task. In this framework, images are categorized into various classes based on distortion types, and contrastive loss is utilized to learn distinctive features that can discriminate between these classes. However, a limitation of this approach is that each real image is treated as a separate class, which limits the effective mining of information from distorted images with similar quality. In contrast, our QFM-IQM is not confined to a specific type of distortion. Instead, we approach learning quality-aware features from a different perspective, aiming to reduce the sensitivity of the pre-trained network to semantic content through adversarial learning. This technique effectively separates the attributes related to image quality from the semantic features. As shown in Table 12 and Table 11, QFM-IQM exhibits excellent performance in both comparative and cross-dataset validation experiments.