Label-free Medical Image Quality Evaluation by Semantics-aware Contrastive Learning in IoMT

Dewei Yi[1] *Member, IEEE*, Yining Hua[1,*], Peter Murchie and Pradip Kumar Sharma, *Senior Member, IEEE*

***Abstract*— With the rapid development of the Internet-of-Medical-Things (IoMT) in recent years, it has emerged as a promising solution to alleviate the workload of medical staff, particularly in the field of Medical Image Quality Assessment (MIQA). By deploying MIQA based on IoMT, it proves to be highly valuable in assisting the diagnosis and treatment of various types of medical images, such as fundus images, ultrasound images, and dermoscopic images. However, traditional MIQA models necessitate a substantial number of labeled medical images to be effective, which poses a challenge in acquiring a sufficient training dataset. To address this issue, we present a label-free MIQA model developed through a zero-shot learning approach. This paper introduces a Semantics-Aware Contrastive Learning (SCL) model that can effectively generalise quality assessment to diverse medical image types. The proposed method integrates features extracted from zero-shot learning, the spatial domain, and the frequency domain. Zero-shot learning is achieved through a tailored Contrastive Language-Image Pre-training (CLIP) model. Natural Scene Statistics (NSS) and patch-based features are extracted in the spatial domain, while frequency features are hierarchically extracted from both local and global levels. All of this information is utilised to derive a final quality score for a medical image. To ensure a comprehensive evaluation, we not only utilise two existing datasets, EyeQ and LiverQ, but also create a dataset specifically for skin image quality assessment. As a result, our SCL method undergoes extensive evaluation using all three medical image quality datasets, demonstrating its superiority over advanced models.**

***Index Terms*— IoMT, Medical imaging, image quality assessment, zero-shot learning.**

I. Introduction

The increasing volume of healthcare requirements places significant pressure and challenges on medical staff [1]. To address this reality, there is a realistic demand for the integration of intelligent and automated technologies in the healthcare industry, giving rise to the concept of the Internet of Medical Things (IoMT). The IoMT aims to establish an intelligent service platform for medical health by seamlessly integrating various complex medical images with global business data [1]. As image processing technologies continue to advance, AI-assisted medical image processing within the IoMT has emerged as a crucial component of smart medicine. By leveraging robust scheduling IoMT platforms and effective image processing algorithms, it holds the promise of alleviating the workload of medical staff through the full utilisation of machine intelligence [2]. The accuracy of biometric measurements heavily relies on the quality of medical images. However, due to the sensitive nature of patient data, strict data privacy standards must be upheld, making it impractical to annotate these images directly. Therefore, it is necessary to explore alternative approaches that ensure data protection within healthcare systems while maintaining the integrity of image quality assessment.

Advanced image processing technologies aid medical professionals in diagnosing patients by assisting in the analysis of vast amounts of medical data [3, 4]. These technologies alleviate the burden on doctors, allowing them to focus on accurate diagnoses and improving healthcare outcomes for patients. The past decade has witnessed remarkable strides in the domains of deep learning and intelligent computing [5]. These cutting-edge technologies have revolutionised the landscape, propelling the boundaries of what is achievable in the analysis and interpretation of medical images. Unlike traditional image recognition tasks, the importance of quality assessment is significantly amplified in the context of medical images. This critical evaluation directly impacts disease diagnosis and grading, making it a crucial component of effective healthcare practices [6, 7]. Image Quality Assessment (IQA) plays a vital role in understanding human perception which can be full-reference IQA or blind IQA [8]. There has been a notable focus on BIQA models, as they offer the ability to evaluate image quality without relying on reference information. Compared to natural IQA, medical IQA is a task that traditionally relies on the expertise of healthcare specialists such as ophthalmologists, radiologists, and other medical professionals. In diabetic retinopathy diagnosis, timely detection of retinal lesions is vital for effective treatment, but it is a process that can be time-consuming and highly dependent on the expertise and experience of healthcare professionals. To facilitate accurate and automated diagnoses, it is necessary to gather large collections of retinal images along with patient information. [9].

Nevertheless, the collection and storage of sensitive data pose legitimate concerns regarding data cybersecurity and

This work was supported by Cancer Research UK (CRUK) under Grant EDDPJT-May23/100001. (∗Corresponding author: Yining Hua)

Dewei Yi, Yining Hua and Pradip Kumar Sharma, are with the Department of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, UK (e-mail: dewei.yi@abdn.ac.uk, yining.hua@abdn.ac.uk, pradip.sharma@abdn.ac.uk).

Peter Murchie is with Centre for Academic Primary Care, Insitute of Applied Health Sciences, University of Aberdeen, Aberdeen, AB25 2ZD, UK (email: p.murchie@abdn.ac.uk).

[1] Dewei Yi and Yining Hua are joint first authors contributing to this work equally.

patient privacy. To solve these issues, one streamline is to safeguard healthcare systems and uphold patient confidentiality by developing robuest security protocols such as intrusion detection, firewall protection, digital forensics, antivirus software, access control, and encryption techniques [10]. Another streamline is to develop domain generalisation methods, where no sensitive data is needed to train a MIQA model while can generalise to assess the quality of medical images. Domain generalisation revolves around two distinct domains: the source domain and the target domain. The primary aim is to train a neural network utilising data from the source domain, enabling it to demonstrate proficiency when confronted with data from the target domain, even without access to labels for the latter. Drawing inspiration from this concept, we are actively investigating the adoption of domain generalisation techniques within the healthcare domain, particularly in the realm of assessing the quality of medical data. This approach holds substantial practical significance as it operates without labels, prioritising the preservation of privacy and security within healthcare systems.

This paper proposes a semantics-aware contrastive learning (SCL) model for medical image quality assessment, which deploys in IoMT for medical data quality assessment. The proposed model is built in medical image label-free manner, where Contrastive Language-Image Pre-training (CLIP) model is introduced to leverage natural language as a flexible prediction space to enable zero-shot learning. To enable generalisation, semantics-aware attributions are extracted from both spatial and frequency domains in a hierarchical manner. More specifically, NSS and local features of patches are extracted spatial domain. Hierarchical features are extracted by steerable Wavelet Decomposition in both local and global levels in frequency domain. Finally, our method integrates the knowledge of zero-shot learning model, NSS, local patches, hierarchical frequency features to derive the final prediction of medical image quality. The main contributions of this paper are summarised as follows.

- We propose a novel Semantics-aware Contrastive Learning (SCL) model for domain-shifted quality assessment of medical images, which is a zero-learning framework. That is, no training and labels are required from medical data.
- To build a model without using medical images and labels, we focus on designing a blind and domain-shifted image quality assessment. More specifically, we transcend traditional paradigms by leveraging the comprehensive visual language encoded in the CLIP model.
- To achieve better generalisation ability, semantics attributions of both spatial and frequency domains are extracted hierarchically from local to global level. In addition, we synthesise low-illumination and blurred skin images based on good quality skin images from ISIC dataset to create SkinQ dataset.
- To carry out a comprehensive evaluation, we compare our method with other state-of-the-art (SOTA) methods on three distinct medical image datasets, including EyeQ [11], LiverQ [12], and SkinQ [13] datasets, to show its effectiveness in domain generalisation.

II. Related Work

A. Zero-shot Learning

Zero-shot learning is to generalise to unseen object categories when no data is available to train a model [14]. One streamline of zero-shot learning is to built models by using pre-trained contrastive models. Contrastive models focus on learning input representations where similar items are positioned closely together and dissimilar items are placed farther apart in the latent space. This approach has demonstrated its effectiveness not only in self-supervised learning methods but also in facilitating zero-shot transfer learning tasks [14, 15]. Zero-shot transfer learning tackles the challenge of performing a task without the need of accessing to dedicated training sets specifically designed for that task [14]. To exemplify this concept, imagine a scenario where an individual has never encountered a zebra previously. Suppose we offer a comprehensive explanation for a zebra, highlighting its horse-like appearance adorned with distinctive black-and-white stripes. In such a scenario, the individual would be able to recognise a zebra when encountering one. CLIP [14] is a recently proposed pre-trained contrastive model, which is able to learn visual representations from natural language supervision. CLIP is trained by an extensive dataset of 400 million image-text pairs obtained, which provides abundant language supervisions. This unique characteristic enables CLIP to perform various image classification tasks without the need for task-specific optimisation.

Given strict patients data protection regulations and high cost of collecting medical images, zero-shot learning is recognised as a promising solution to solve medical image quality assessment. Therefore, this paper focuses on developing a zero-shot transfer model, which can generalize its learned knowledge to a new task. In our work, zero-shot transfer is used to conduct image quality assessment for various types of medical images (e.g., fundus images, ultrasound images, and dermoscopic images).

B. Blind Image Quality Evaluator (IQE) and Medical IQE

In the absence of ground-truth images, the evaluation of image quality can be performed using no-reference IQA methods [16]. These methods operate on the premise that natural scene images exhibit specific statistical characteristics and suffer from alternations under distortions. The degree of this alteration can be quantified to assess the quality of the image accurately. Instead of relying solely on statistical properties, features extracted from supplementary datasets are utilised to quantify the degradation observed in natural scene images. These features serve as alternatives to statistical properties for assessing image quality. In most of blind IQA methods, a model is trained on degraded images to facilitate the training process [17, 18]. As a result, the state-of-the-art no-reference IQA methods are less effective accounting for the artifacts such as incorrect high-frequency details. On the other hand, given that medical images often exhibit blur and ringing artifacts [19], it is crucial for medical IQA algorithms to share

similarities with existing metrics to measure blur and sharpness. High-quality medical imaging is essential for enabling accurate interpretation, precise diagnosis, informed surgical planning, and effective treatment delivery. The quality of medical images plays a critical role in ensuring the success and reliability of these crucial healthcare processes. Image quality assessment methods, such as NIQA [20] and BRISQUE [21], have proven to be effective in quality assurance and clinical diagnosis across diverse imaging modalities. These methods have found successful applications in fields such as MR imaging [22, 23] and fundus imaging [24, 25].

In medical community, it is very difficult to access full-reference medical image dataset which contains the scores from low-to-high in various quality versions of identical medical images. Hence, blind IQA methods offer a promising approach for medical IQA as they can assess the quality of diverse images without the requirement of having the exact same images. Instead, these methods rely on scoring variations across different images to evaluate their quality. In this work, we attempt to fill two gaps: high cost of labelling medical image quality and the challenging of acquiring sufficient medical images. To tackle with these two issues, a novel blind and zero-shot model is proposed in this work to assess the quality of medical images through learning from perceptual features of spatial and frequency domains in hierarchical levels and then therefore generalising quality assessment on various types of medical images.

C. Natural Scene Statistics

Natural scenes encompass a wide range of images and videos captured using high-quality devices operating in the visual spectrum, representing the visual environment. This distinction sets them apart from other forms of media such as text, computer-generated graphics, cartoons, animations, paintings, drawings, random noise, or images and videos captured from non-visual stimuli like radar, sonar, X-rays, ultrasounds, and so on. The realm of natural scenes constitutes a small subset within the vast expanse of all possible scenes [26, 27]. To comprehend the intricacies of this subspace, numerous researchers have delved into studying the statistical properties and crafting statistical models for natural images [26]. By employing localized models like principal components analysis (PCA) and independent components analysis (ICA), researchers have uncovered intriguing connections between the statistical characteristics of natural scenes and the intricacies of the human visual system (HVS) [27]. These models, which incorporate local statistics that capture human attention. In general, approaches that model the statistics of natural scenes are called natural scene statistics (NSS) models. By leveraging NSS, it enables supervised learning models to predict the quality of images. One noteworthy model in this context is NIQE [20], which is widely recognised as one of the pioneering unified BIQA models. NIQE aims to effectively capture a wide range of distortions. Nevertheless, only employing NSS features in NIQE does not exhibit sufficient sensitivity to the introduction of "unnaturalness" in images caused by real-world distortions [20]. To address this limitation, [28] improved NIQE by incorporating a more robust set of NSS features for the quality predictions of local regions. This enhancement allows for improved detection and evaluation of quality issues in images affected by various types of distortions. NSS have been explicitly incorporated into a number of image processing applications, including image compression, image denoising, and image segmentation, etc. [26].

Although NSS can perform well in processing simplistic distortions in a blind manner, it is insufficient for medical image quality assessment due to more complicated distortions contained in medical images. Taking this into account, our method combine zero-shot learning based model with NSS to achieve better generalisation ability for various types of medical images.

III. Semantics-aware Contrast Learning (SCL) Model

A. The Architecture of SCL Model

An overview of our SCL model is shown in Fig. 1. The original size of an medical image is processed in three branches. For the first branch, input image is used to extract spatial and frequency domain features. More specially, discrete cosine transform (DCT) is conducted to extract local frequency features. Steerable Wavelet Decomposition (SWD) is conducted on input image to generate neighbouring wavelet coefficients and then passed to Gaussian Scale Mixture (GSM) to extract global frequency features. Spatial features are derived from patches instead of individual pixels, enabling them to possess enhanced discriminative strength. Additionally, principal component analysis (PCA) is applied to the patches, and the results of singular values are utilised to capture and describe spatial discontinuities in the image. This approach aids in effectively characterising the structural properties and local variations within the image. With using these three branches, features from both spatial and frequency domains under local and global levels are extracted. For the second branch, a prompt pair {"Good photo." And "Bad photo."} is utilised to exploit CLIP for perception assessment so that ambiguity can be reduced. The features of a paired prompts (normal prompt and its antonym) are defined by t_1 and t_2. Then, the cosine similarity of them are measured to derive final score. For the third branch, the original size is partitioned into fixed-size patches to extract and integrate local quality-aware NSS features. The pristine multivariate Gaussian (PMVG) model is fitted to the feature vector of each patch and its local quality score is computed accordingly. The quality score of full-size medical image is obtained by integrating local quality scores of all patches with using a pooling operation.

B. Zero-shot Transfer

In our work, an effective prompt pairing strategy is introduced to mitigate the ambiguity problem of CLIP. This strategy entails utilising pairs of antonym prompts, such as "Good photo." and "Bad photo.", to facilitate each prediction. Suppose that p_1 and p_2 are features extracted from the two prompts that convey opposite meanings. Initially, the cosine

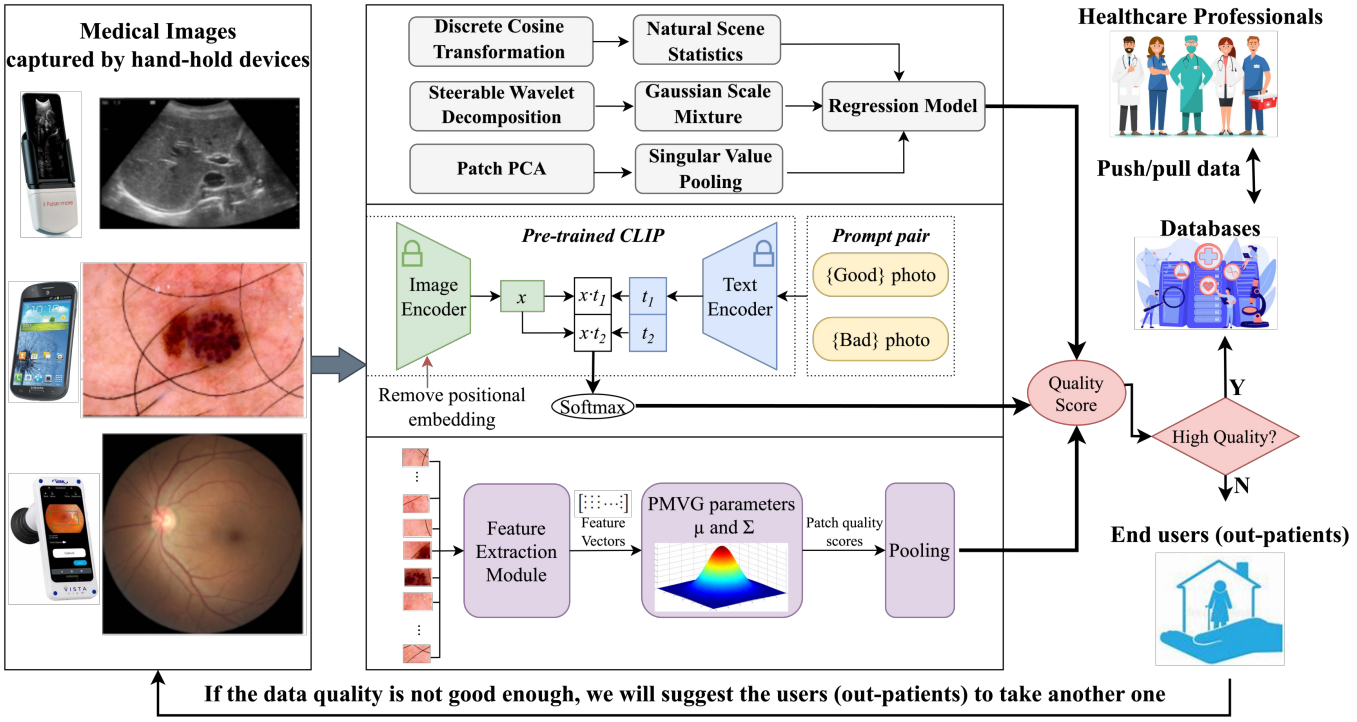


Fig. 1: The overall architecture of the proposed semantics-aware contrastive learning (SCL) model for assessing the quality of medical images. The data flow is also provided starting from IoMT devices (e.g., smart phones) which collecting medical images and then captured medical images are passed to our SCL model deployed in edge devices (smart phones) for quality assessment. If the captured medical images are high quality, these medical images will be passed data centre and healthcare professionals. Otherwise, we will suggest the out-patients to retake the medical images.

similarity between the image feature x and each prompt feature is computed. The final similarity of $SIM^* \in [0,1]$ is calculated using Softmax as follows:

$$SIM_i = \frac{i \odot p_i}{||i|| \cdot ||p_i||}, i \in \{1,2\} \tag{1}$$

$$SIM^* = \frac{e^{SIM_1}}{e^{SIM_1} + e^{SIM_2}} \tag{2}$$

where $i \in \mathbb{R}^D$ and $p \in \mathbb{R}^D$ denote the feature vectors extracted from the image and the prompt, respectively. The operator $\odot$ denotes the dot product between vectors, and the notation $||\cdot||$ represents the ℓ_2 norm of a vector.

Using antonyms (contrasting adjectives) in prompts effectively eliminates prompt ambiguity and significantly enhances performance aligning predictions better with human perception. Following guidance from [29], we use the "[text]" photo for prompt simplicity and common-sense for our domain-shifted problem. We also analyzed the impact of different adjectives using the same template, observing performance variations. For overall image quality assessment, "Good/Bad" prompts correlate more strongly with human perception than "High quality/Low quality" or "High definition/Low definition" prompts, indicating that uncommon adjectives may yield weaker results. Dealing with synonyms presents a challenge, underscoring the importance of meticulous prompt design for accuracy and reliability.

1) Removal of Positional Embedding: The Naive CLIP model's fixed-sized inputs aren't ideal for perception assessment, as resizing and cropping can introduce distortions affecting the score. To address this, we follow [29] by removing positional embeddings and adopting ResNet. Unlike convolutional models, ResNet embeds positional information deep within its architecture, enhancing performance even without explicit positional embeddings. This contrasts with Transformer models, which are more sensitive to positional embeddings' removal and may lead to a drop of performance.

2) Quality Perception: In No-Reference IQA for overall quality perception assessment, we use common antonym pairs like "Good photo" and "Bad photo" and fine-tune them with CLIP-IQA+ [29] without modifying the network weights, benefiting from extensive language-vision training for enhanced generalizability. To assess fine-grained quality aspects like brightness, noisiness, and sharpness, we adapt the same zero-shot model [29] by replacing "good" and "bad" with the attribute and its antonym, e.g., "Bright photo" and "Dark photo" for brightness evaluation. Unlike most learning-based approaches, this zero-shot model doesn't rely on predefined labels making it versatile for assessing various attributes.

C. Semantics-aware Perception

1) Local Quality-aware NSS extraction and Integration: A multivariate Gaussian (MVG) model is constructed to represent NSS features from natural images, which serves as a

"reference" for evaluating medical image patch quality. To make NSS features more meaningful for quality prediction, we selectively use patches with significant contrast to improve the effectiveness of image quality assessment. The contrast of a pixel is given as follows.

$$\bar{I}(x,y) = \frac{I(x,y) - \mu(x,y)}{\sigma(x,y) + 1} \tag{3}$$

where x and y are spatial coordinates. $\mu(x,y)$ and $\sigma(x,y)$ are the mean and contrast of the local image. For smoothing, the additional one is added in the denominator.

Then, we calculate patch contrast by combining contrasts within each patch to provide a quantitative assessment of overall contrast for comprehensive image quality analysis. To enhance quality prediction, we compute NSS features at different scales for capturing multi-scale attributes and improving assessment robustness by considering various levels of detail. These selected patches create feature vectors of dimensionality d by combining NSS and gradient magnitude features. Given independent samples $x_i, i \in 1, ..., n$ from an m-dimensional MVG distribution, we can learn the MVG distribution from x_i by using maximum likelihood estimation as follows.

$$f(x) = \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{4}$$

where $x \in R^{m\times 1}$ is the vector variable, and μ and Σ are the mean vector and covariance matrix of x. Note that the MVG model is fully described by the pair (μ, Σ).

After obtaining MVG model (μ, Σ) followed by [28], it can be employed to assess the quality of patches within a given medical image. During training, a medical image is partitioned into k patches. For i-th patch, an NSS feature vector y_i of dimension d is extracted. To reduce the dimensionality of y_i, a pre-learned projection matrix Φ is used below.

$$y_i' = \Phi^T f(x_i), y_i' \in R \tag{5}$$

With the feature set $y'ii = 1^k$ obtained from a test medical image, we can proceed to predict its quality score while different local regions in an image can contribute differently to overall perception of image quality [28]. To address this, we fit each patch i with an MVG model (μ_i, Σ_i) to predict its local quality score. The overall quality score for the test medical image is then calculated by averaging the local quality scores of all patches. To simplify computation, all patches share the same covariance matrix Σ'. Thus, the MVG model for each patch i is represented by (y_i', Σ'). To measure the distortion level of patch i, we employ the following formula:

$$q_i = \sqrt{(\mu - y_i')^T(\frac{\Sigma + \Sigma'}{2})^{-1}(\mu - y_i')} \tag{6}$$

where q_i quantifies the statistical distortion of patch i from the reference statistics derived from high-quality natural images.

2) Spatial and frequency statistical features: For local frequency features, discrete cosine transform (DCT) is introduced to extract the statistics of coefficients. Then, DCT coefficients are fitted with the flexible Generalized Gaussian Distribution (GGD) to capture their underlying statistical properties [16] as follows.

$$\begin{aligned} f(x|\mu,\beta) &= \frac{1}{2\Gamma(1+\beta^{-1})}e^{-|x-\mu|^\beta} \\ \Gamma(z) &= \int_0^\infty t^{z-1}e^{-t}dt \end{aligned} \tag{7}$$

where μ is the mean of x, β is a parameter to control the shape of distribution, and $\Gamma(\cdot)$ is Gamma function. β is more discriminative than μ in describing DCT coefficient statistics, so we use it as a feature for characterizing medical images. In a DCT block, σ is the standard deviation. To quantify block perturbation, we use the ratio $\bar{\sigma} = \frac{\sigma}{\mu}$, where μ is the mean value. We divide the coefficients into three sets in a DCT block and compute their normalised deviations $\bar{\sigma_i}$, where $i = 1, 2, 3$. We also calculate the variation Σ of these normalised deviations as additional features. To mitigate bias of statistics from concatenated blocks, we aggregate them by computing mean values to represent for each medical image.

For global frequency features, the overall wavelet coefficient distribution in medical images may not fit a standard Gaussian distribution well. Instead, we use the Gaussian scale mixture (GSM) model to effectively capture marginal and joint statistics of images following [16]. Considering a group of adjacent wavelet bands denoted by a vector Y ($Y \equiv z \cdot U$), it is classified as a GSM, where $\equiv$ denotes equality in probability distribution, U is a zero-mean Gaussian random vector, and z is a scalar random variable. Q is the covariance of U. The density of Y can be expressed as an integral below.

$$P_Y(y) = \int_0^\infty \frac{1}{(2\pi)^{N/2}|z^2Q|^{1/2}} e^{\left(-\frac{Y^TQ^{-1}Y}{2z^2}\right)} p_z(z)dz \tag{8}$$

where $P_Y(y)$ is the probability of the mixing variable z and N is the number of filters in the neighborhoods. To create neighboring wavelet coefficients, we utilize steerable pyramid decomposition on a medical image in both real and imaginary domains for improved discriminatory capabilities.

For spatial features, we leverage the strong connection between pixel intensity variation and perceptual scores in medical image studies. To enhance discrimination, we extract features from patches, not individual pixels. We employ Principal Component Analysis (PCA) on the image patches, using their associated singular values to represent spatial discontinuity. In smoother medical images, singular values diminish more rapidly toward zero, indicating reduced significance of corresponding eigenvectors in capturing important features.

For spatial features, by recognising the close relationship between the spatial discontinuity of pixel intensity and perceptual scores in subject studies for medical images, we enhance the discriminative strength by extracting features from patches instead of individual pixels. To capture the spatial discontinuity in a more effective manner, the patches of an medical image are applied by principal component analysis (PCA) and then we leverage the associated singular values to represent the spatial discontinuity. The singular values of medical images containing smooth contents tend to diminish more quickly, approaching zero, compared to images with sharp contents. This behavior reflects the reduced importance of corresponding eigenvectors in capturing significant features.

D. Semantics and zero-shot information fusion

The decision-level fusion method of our work is a meta-method classifier that fuses learning-based and non-learning-based image quality assessment models to recognise high-quality and low-quality medical images through a soft voting technique [30]. The soft voting technique predicts HQ and LQ medical images based on the predicted quality scores of each medical image assessment models as follows.

$$\hat{y} = \arg\max_{c} \sum_{k=1}^{m} w_k p_{c,k} \tag{9}$$

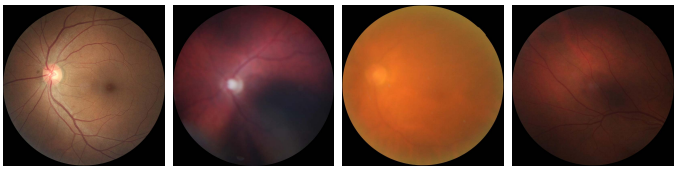
where w_j is a weight that can be given to determine the contribution of each IQA mode and $p_{c,k}$ represents the predicted probability of the class label c and the classifier k.

IV. Experiments Results

A. Dataset

There are three medical image datasets, including EyeQ [11], LiverQ [12], and SkinQ [13], to be used to evaluate the performance of our proposed methods. EyeQ is a fundus image dataset. LiverQ is an ultrasound image dataset. SkinQ is a dermoscopic photograph dataset.

EyeQ dataset is a re-annotated retinal image quality dataset from the EyePACS dataset, which is a diverse retinal image dataset captured using various camera models and types, encompassing a wide range of imaging conditions. In this paper, we utilise 'Good' grade and 'Reject' grade with considering blurring, uneven illumination, low-contrast, and artifacts as common quality indicators. "Good" retinal image exhibits no low-quality factors, and all retinopathy characteristics are clearly visible, as depicted in Fig. 2a. "Reject" retinal image suffers from significant quality issues, rendering it unsuitable for providing a comprehensive and reliable diagnosis, even by ophthalmologists, as illustrated in Figs. 2b to 2d. Moreover, a fundus image that has an invisible disc or macula region is also classified as a "Reject" grade. There are two experts involving in grading the quality of images in the EyePASC dataset for the purpose of re-annotating the EyeQ dataset. Subsequently, the images with ambiguous labels were excluded, resulting in 16818 Good quality and 5540 Reject quality retinal images. A summary of Good quality and Reject quality and lesion levels are provided in Table I and some examples of "Good" and "Reject" quality retinal images are shown in Fig. 2.



(a) Good (b) Blur (c) Low-contrast (d) Illumination

Fig. 2: Examples of "Good" and "Reject" quality image in EyeQ dataset: (a) Good; (b) Reject: Blur; (c) Reject: Low-contrast; (d) Reject: Illumination.

LiverQ dataset comprises clinical ultrasound (US) images of the abdominal liver obtained from a retrospective database

TABLE I: The statistics of the EyeQ dataset, where L-i identifies the level of retinopathy.

	L0	L1	L2	L3	L3	Total
Good	12,308	1,585	2,454	366	104	16,694
Reject	3,739	262	995	191	353	5,540
Total	16,047	1,847	3,449	557	457	22.234

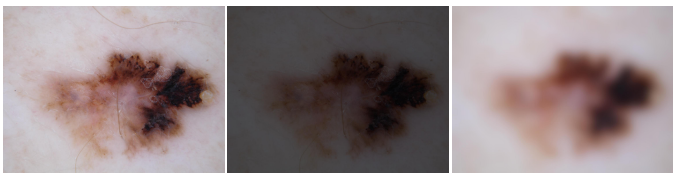
of the University Hospital of Angers in France. This dataset consists of 72 images of varying sizes (1080×810, 1024×768) captured by SuperSonic Aixplorer and Siemes Acuson S2000 systems, exhibiting granular, smooth, cirrhotic, and non-cirrhotic liver textures. The images were anonymised, and ethical approval was obtained from the University Hospital of Angers for their use. The perceived quality of the images was evaluated based on four criteria: image contrast, diagnostic ability, texture conspicuity, and edge sharpness by three radiologists from the affiliated Hospital of Nanjing Medical University in China. These radiologists had different levels of experience and were not familiar with the test images. The assessment was conducted following the European guidelines on quality criteria for diagnostic radiographic images [12], which recommend involving at least two observers to independently assess each image with the given criteria. Fig. 6 illustrates examples of high and low-quality ultrasound images from this dataset.

(a) HQ: 83.33 (b) LQ: 20.00

Fig. 3: Examples of High and Low Quality Ultrasound Images in LiverQ dataset, where (a) represent High Quality (HQ) ultrasound images; (b) represent Low Quality (LQ) ultrasound images; values shown under images are the averaged quality scores of three human experts.

SkinQ dataset is a synthetic image quality dataset derived from the ISIC dataset [13], which consists of skin images captured using various camera models under diverse imaging conditions and guided by a dermatologist. In this paper, we utilise "Good" grade and "Reject" grade with considering blurring and illumination. "Good" grade refers to skin images that exhibit no noticeable quality issues and where all skin characteristics are clearly visible, as depicted in Fig. 4-(a). "Reject" grade is assigned to skin images that exhibit significant quality issues and cannot be relied upon for a comprehensive and accurate diagnosis, even by dermatologists, as illustrated in Fig. 4-(b) and (c). The skin images are extracted from ISIC2018. Good images are image directly from the dataset. Rejected images are synthesised by good images. Then, there are 2005 Good quality and 4010 Reject quality skin images included in the dataset and there are 6015

skin images in SkinQ dataset.



(a) Good (b) Reject: Dark (c) Reject: Blur

Fig. 4: Examples of "Good" and "Reject" quality image in SkinQ dataset: (a) Good; (b) Reject: Dark; (c) Reject: Blur.

B. Implementation and Metrics

The implementation of our method is based on PyTorch, which is a deep learning framework. For the backbone network of zero-shot learning, the ResNet-50 is chosen due to its competent performance [29]. In our experiment, all three medical image datasets, including EyeQ, LiverQ, and SkinQ, are used for evaluating performance. For evaluation metrics, there are five threshold-dependent measures used including accuracy (Acc), precision (P), recall (R), specificity (S), and F_β. The definitions of these five metrics are given below:

$$P = \frac{t_p}{t_p + f_p}, \quad R = \frac{t_p}{t_p + f_n}, \quad S = \frac{t_n}{t_n + f_p}$$
$$Acc = \frac{t_p + t_n}{t_p + f_p + t_n + f_n} \tag{10}$$

where f_p is the number of false positives, t_p is the number of true positives, f_n is the number of false negatives, and t_n are the true negatives.

Moreover, F_β is introduced which considers precision and recall simultaneously, where β is the parameter for adjusting the importance of precision and recall. When β exceeds one, it means that precision is more important. Otherwise, recall is more important. In this paper, precision and recall are both significant for medical image quality assessment so we use F_1 score by setting $\beta = 1$ which applies the same weight to precision and recall [31] because correctly detecting good and missing recognising poor quality are both important for medical images.

$$F_\beta = (1 + \beta^2) \times \frac{P \times R}{(\beta^2 \times P) + R} \tag{11}$$

C. Quantitative Evaluation

We use above-mentioned accuracy, precision, recall, specificity, and F_1 as metrics to benchmark MIQA performance, which are evaluated on all EyeQ, LiverQ, and SkinQ medical image datasets. For performance on these three datasets, we have several interesting observations from Table II, Table III, and Table IV.

First, our proposed SCL method achieves the best label-free performance of fundus images in terms of accuracy and F_1 in all EyeQ, LiverQ, and SkinQ datasets. Accuracy and F_1 are two of the most metrics to evaluate the holistic performance. In EyeQ dataset, our proposed method can reach 95.00% on accuracy and 89.48% on F_1. In LiverQ dataset, our proposed method can reach 84.72% on accuracy and 74.42% on F_1. In SkinQ dataset, our proposed method can reach 83.16% on accuracy and 87.12% on F_1. With regard to accuracy and F_1, our proposed method significantly outperforms other SOTA methods.

TABLE II: Performance Comparison of Zero-shot Image Quality Assessment on EyeQ (Fundus) Dataset (Unit %)

Method	Acc.	Precision	Recall	Specificity	F_1
CLIPIQA+ [29]	91.00	98.17	64.87	99.60	78.12
CNNIQA [32]	68.46	40.57	58.72	71.67	47.99
DBCNN [33]	52.38	33.83	**96.37**	37.89	50.08
ILNIQE [28]	76.88	51.84	94.13	71.19	66.86
MANIQA [34]	65.42	39.27	72.44	63.11	50.93
MUSIQ [35]	88.64	94.27	57.64	98.85	71.54
NIMA [36]	82.28	63.40	67.42	87.18	65.35
NIQE [20]	46.49	31.21	96.26	30.10	47.13
NRQM [16]	77.04	88.59	8.41	99.64	15.36
PAQ2PIQ [37]	75.75	**99.16**	2.13	**99.99**	4.17
PI [38]	88.09	75.96	76.01	92.07	75.98
TReS [39]	73.63	48.29	90.51	68.07	62.97
SCL (Ours)	**95.00**	93.58	93.58	98.06	**89.48**

Second, DBCNN outperforms other methods with regard to the recall in EyeQ and LiverQ datasets. Although DBCNN has great performance on recall, its precision and specificity are 33.83% and 37.89% which fail half of our method on EyeQ dataset. Similar conclusion can be drawn on LiverQ dataset. In LiverQ dataset, DBCNN has the best performance on recall while its specificity is only 8.00% and the specificity of our method can reach 90.00%. Different from EyeQ and LiverQ datasets, NIMA achieves the best performance with regard to recall in SkinQ dataset. Although NIMA outperforms others on recall, its specificity is only 0.70% where the specificity of our method can reach 65.49% in SkinQ dataset.

TABLE III: Performance Comparison of Zero-shot Image Quality Assessment on LiverQ (Ultrasound) Dataset (Unit %)

Method	Acc.	Precision	Recall	Specificity	F_1
CLIPIQA+ [29]	73.61	57.90	50.00	84.00	53.66
CNNIQA [32]	45.83	30.23	59.09	40.00	40.00
DBCNN [33]	36.11	32.35	**100.0**	8.00	48.89
ILNIQE [28]	69.44	50.00	13.64	94.00	21.43
MANIQA [34]	44.44	33.93	86.36	26.00	48.72
MUSIQ [35]	54.17	38.78	86.36	40.00	53.52
NIMA [36]	51.39	36.74	81.82	38.00	50.70
NIQE [20]	83.33	72.73	72.73	88.00	72.73
NRQM [16]	70.83	**100.0**	4.546	**100.0**	8.696
PAQ2PIQ [37]	69.44	50.00	4.546	98.00	8.333
PI [38]	72.22	75.00	13.64	98.00	23.08
TReS [39]	36.11	31.25	90.91	12.00	46.51
SCL (Ours)	**84.72**	76.19	72.73	90.00	**74.42**

Third, for the precision and specificity on EyeQ and SkinQ datasets, the best performances are achieved by PAQ2PIQ method, which are 99.16% and 99.99% on EyeQ dataset and 95.59% and 98.63% on SkinQ dataset. However, when coming to its F_1, 4.17%. In contrast, our method can achieve 89.48% of F_1 on EyeQ dataset. Simiar to above observation, the recall and F_1 of NRQM in SkinQ dataset is only 19.85% and 32.87%, respectively. In contrast, our method can achieve 94.93% of recall and 87.12% of F_1 on SkinQ dataset. For the precision and specificity, there is a divergence between EyeQ

and SkinQ datasets with LiverQ dataset. The best performance of Precision and specificity achieve by PAQ2PIQ method on EyeQ dataset and SkinQ. For LiverQ dataset, NRQM obtains the best performance in terms of precision and specificity. However, similar to PAQ2PIQ in EyeQ dataset, the F_1 of NRQM in LiverQ dataset is only 8.70%. In contrast, our method can achieve 61.22% of F_1 on LiverQ dataset.

TABLE IV: Performance Comparison of Zero-shot Image Quality Assessment on SkinQ Dataset (Unit %)

Method	Acc.	Precision	Recall	Specificity	F_1
CLIPIQA+ [29]	78.85	76.95	92.46	58.43	83.99
CNNIQA [32]	68.61	65.99	98.40	23.90	79.00
DBCNN [33]	68.68	65.95	98.82	23.45	79.11
ILNIQE [28]	76.83	76.41	88.80	58.85	82.14
MANIQA [34]	61.29	66.11	72.83	43.97	69.31
MUSIQ [35]	64.80	63.38	97.92	15.11	76.95
NIMA [36]	60.28	60.17	**99.98**	00.70	75.13
NIQE [20]	75.10	70.95	99.05	39.15	82.68
NRQM [16]	80.58	85.80	81.05	79.88	83.36
PAQ2PIQ [37]	51.35	**95.59**	19.85	**98.63**	32.87
PI [38]	82.53	79.59	95.32	63.32	86.75
TReS [39]	74.36	78.30	79.24	67.04	78.76
SCL (Ours)	**83.16**	80.50	94.93	65.49	**87.12**

D. Qualitative Evaluation

Qualitative evaluation is conducted on all EyeQ, LiverQ, and SkinQ medical image datasets. In EyeQ medical image dataset, all images are collected from real world. Therefore, for each image, only one type of quality is given, i.e., "Good", "Reject: Blur", "Reject: Low". Some examples on prediction of "Good" or "Reject" retinal images are provided in Fig. 5. Given that EyeQ dataset only identifies the categories of retinal images as "Good" or "Reject" without the corresponding quality scores, the ground truth of categories are provided along with the predicted quality scores of our SCL method. The range of predicted scores is from 0 to 100. "0" means lowest quality and "100" means the highest quality and therefore the threshold is set to "50" to determine "Good" or "Reject" retinal images. There is an interesting finding that our zero-shot learning method does not perform very well on case of Fig. 5-(d). After having a further analysis, we are not surprised for it. Although this retinal is rejected by illumination, only above part of this image is low-illumination and the texture of this image is clear along with good contrast.

In LiverQ ultrasound image dataset, all ultrasound images are scored by three human experts. Some examples on prediction of "High-quality" or "Low-quality" ultrasound images are provided in Fig. 6, where the left scores in the bracket are predicted quality scores of our SCL method and the right scores in the bracket are averaged quality scores of three human experts which is treated as ground truth. All quality scores are normalised between 0 to 100 and higher score value means better image quality. The threshold is also set to "50" to determine "HQ" or "LQ" ultrasound images.

In SkinQ image dataset, there are three types of skin images, including "Good" images, "Reject: Blur" images, "Reject: Dark". "Good" images are realistic dermoscopic images extracted from ISIC dataset. "Reject: Blur" and "Reject: Dark" images are synthesised based on real-world ISIC skin images. To evaluate the performance in a more comprehensive way, the SkinQ dataset includes both matched image groups and unmatched image groups. If a "Good" skin image have both blurred and darked versions of it, it belongs to matched image groups. Otherwise, it belongs to unmatched image groups. Some examples on the predictions of "Good" or "Reject" dermoscopic images are provided in Fig 7. The scores under images are predicted quality scores of our SCL method. All quality scores are normalised between 0 to 100 and higher score value means better image quality. The threshold is also set to "50" to determine "Good" or "Reject" skin images.

With using matched synthetic skin images from SkinQ dataset, low-quality skin images obtain lower quality scores from our SCL model. In Fig. 7, the score under each image is its corresponding predicted quality score from our SCL model. The first column is raw skin images which are high-quality images. The second column is low-illumination skin images. The third column is blurred skin images. We can see that blurred images have the lowest scores followed by this. Low-illumination of images have the second lowest scores.

E. Ablation Study

Comprehensive ablation studies are performed to assess the contributions of each module in our proposed method, including natural scene statistics (NSS), Local Integration (LI), Local and Global frequency (LGF), as well as to study the influence of contrastive learning (CL) in Table I. When only using NSS features, the accuracy and F_1 can achieve 46.49% and 47.13% on EyeQ. Then, local features are integrated. The accuracy and F_1 can be improved to 76.88% and 66.86%, respectively. As argued in [16], both local and global feature of frequency domain can also provide a significant contribution to extract the insights for quality assessment. In light of this, we adapt the local and global frequency domain features to improve the accuracy and F_1 to 92.25% and 83.62%.

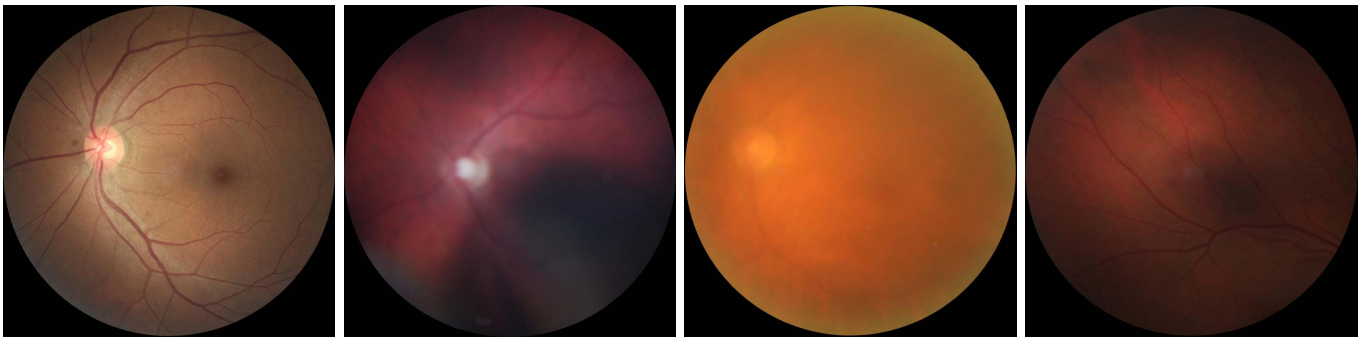
Next, we introduce the contrastive learning to improve the generalisation ability. Contrastive learning is conducted in a zero-shot learning manner to achieve better generalisation without the requirement of labelled medical images. After introducing the contrastive learning, the accuracy of F_1 are further improved to 95.00% and 89.48%, respectively.

TABLE V: Ablation Study on EyeQ Dataset, where NSS: natural scene statistics, LI: Local Integration, LGF: Local and Global frequency, CL: contrastive learning (Unit %).

Baseline (NSS)	LI	LGF	CL	Acc	F_1
✓				46.49	47.13
✓	✓			76.88	66.86
✓	✓	✓		92.25	83.62
✓	✓	✓	✓	95.00	89.48

V. Conclusion

In this paper, we proposed a semantics-aware contrastive learning (SCL) model to implement zero-shot transfer medical image assessment so as to achieve generalised medical image

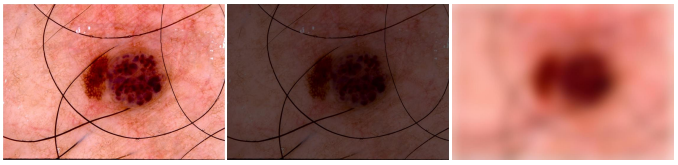


(a) Good (59.23) (b) Reject: Blur (27.25) (c) Reject: Low-contrast (41.18) (d) Reject: Illumination (38.63)

Fig. 5: Visual examples of EyeQ and predicted quality score is shown in the bracket. (a) is image graded as "Good" quality. (b)-(d) are images graded as poor ("Reject") quality. For rejected images, it can be rejected due to blur, low-contrast, and low-illumination.

(a) HQ: (60.50/83.33) (b) LQ: (47.50/26.67) (c) HQ: (58.09/80.00) (d) LQ: (40.21/20.00)

Fig. 6: Examples of High and Low Quality Ultrasound Images in LiverQ dataset, where (a) represent High Quality (HQ) ultrasound images; (b) represent Low Quality (LQ) ultrasound images; values shown under images are (predicted score / ground-truth score), where the ground-truth is the averaged quality scores of three human experts.



(a) Good (71.51) (b) Dark (54.82) (c) Blur (24.94)

Fig. 7: Visual examples of SkinQ and predicted quality score is shown in the bracket. (a) is image graded as "Good" quality. (b) is image graded as "poor" quality due to dark. (c) is image with poor predicted quality due to blur.

quality assessment, where Contrastive Language-Image Pre-training model is introduced to leverage natural language as a flexible prediction space so as to achieve label-free quality assessment for various types of medical images. Moreover, semantics-aware attributions are extracted from both spatial and frequency domains to further enhance the robustness for various types of medical image quality assessment. More specifically, NSS and local patch features are extracted from spatial domain. In frequency domain, both local and global frequency features are extracted with using steerable Wavelet Decomposition. With considering features extracted from CLIP model, spatial domain, and frequency domain, the final quality score of a given medical image can be derived.

In order to have a comprehensive evaluation, not only using existing EyeQ fundus image and LiverQ ultrasound image datasets, we also create a new SkinQ dataset which includes original skin images and two types (blur and low-illumination) synthetic skin images based on ISIC dataset. Therefore, our proposed method was evaluated on three medical image datasets, including EyeQ (fundus images), LiverQ (ultrasound image). Experimental results demonstrate that our proposed method outperforms other SOTA methods on zero-shot medical image quality assessment. We also conduct an ablation study to investigate the contribution of various components in our method.

Acknowledgment

For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References

[1] H. Li, K. Yu, B. Liu, C. Feng, Z. Qin, and G. Srivastava, "An efficient ciphertext-policy weighted attribute-based encryption for the internet of health things," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 1949–1960, 2021.

[2] Z. Guo, K. Yu, A. Jolfaei, F. Ding, and N. Zhang, "Fuz-spam: label smoothing-based fuzzy detection of spammers in internet of things," *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 11, pp. 4543–4554, 2021.

[3] S. Xu, J. Gu, Y. Hua, and Y. Liu, "Dktnet: Dual-key transformer network for small object detection," *Neurocomputing*, vol. 525, pp. 29–41, 2023.

[4] A. P. James and B. V. Dasarathy, "Medical image fusion: A survey of the state of the art," *Information fusion*, vol. 19, pp. 4–19, 2014.

[5] Y. Liu, D. Zhang, Q. Zhang, and J. Han, "Part-object relational visual saliency," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3688–3704, 2021.

[6] D. Yi, P. Baltov, Y. Hua, S. Philip, and P. K. Sharma, "Compound scaling encoder-decoder (cosed) network for diabetic retinopathy related bio-marker detection," *IEEE journal of biomedical and health informatics*, 2023.

[7] H. Xu and J. Ma, "Emfusion: An unsupervised enhanced medical image fusion network," *Information Fusion*, vol. 76, pp. 177–186, 2021.

[8] Y. Huang, L. Li, Y. Yang, Y. Li, and Y. Guo, "Explainable and generalizable blind image quality assessment via semantic attribute reasoning," *IEEE Transactions on Multimedia*, 2022.

[9] J. M. P. Dias, C. M. Oliveira, and L. A. da Silva Cruz, "Retinal image quality assessment using generic image quality indicators," *Information Fusion*, vol. 19, pp. 73–90, 2014.

[10] Y. Li, Y. Zuo, H. Song, and Z. Lv, "Deep learning in security of internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 22, pp. 22 133–22 146, 2021.

[11] Z. Shen, H. Fu, J. Shen, and L. Shao, "Modeling and enhancing low-quality retinal fundus images," *IEEE Trans. Med. Imaging*, vol. 40, no. 3, pp. 996–1006, 2020.

[12] M. Outtas, L. Zhang, O. Deforges, A. Serir, W. Hamidouche, and Y. Chen, "Subjective and objective evaluations of feature selected multi output filter for speckle reduction on ultrasound images," *Physics in Medicine & Biology*, vol. 63, no. 18, p. 185014, 2018.

[13] P. Tschandl, C. Rosendahl, and H. Kittler, "Data descriptor: the ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," 2018.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[15] M. R. Taesiri, F. Macklon, and C.-P. Bezemer, "Clip meets gamephysics: Towards bug identification in gameplay videos using zero-shot transfer learning," in *Proceedings of the 19th International Conference on Mining Software Repositories*, 2022, pp. 270–281.

[16] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017.

[17] P. Ye and D. Doermann, "Active sampling for subjective image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4249–4256.

[18] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, 2011, pp. 305–312.

[19] Y. Liu, J. Wang, S. Cho, A. Finkelstein, and S. Rusinkiewicz, "A no-reference metric for evaluating the quality of motion deblurring," *ACM Transactions on Graphics*, 2013.

[20] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[21] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. on Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.

[22] L. S. Chow, H. Rajagopal, R. Paramesran, A. D. N. Initiative *et al.*, "Correlation between subjective and objective assessment of magnetic resonance (mr) images," *Magnetic resonance imaging*, vol. 34, no. 6, pp. 820–831, 2016.

[23] L. S. Chow and H. Rajagopal, "Modified-brisque as no reference image quality assessment for structural mr images," *Magnetic resonance imaging*, vol. 43, pp. 74–87, 2017.

[24] T. Köhler, A. Budai, M. F. Kraus, J. Odstrčilik, G. Michelson, and J. Hornegger, "Automatic no-reference quality assessment for retinal fundus images using vessel segmentation," in *Proceedings of the 26th IEEE international symposium on computer-based medical systems*. IEEE, 2013, pp. 95–100.

[25] S. Wang, K. Jin, H. Lu, C. Cheng, J. Ye, and D. Qian, "Human visual system-based fundus image quality assessment of portable fundus camera photographs," *IEEE transactions on medical imaging*, vol. 35, no. 4, pp. 1046–1055, 2015.

[26] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: Jpeg2000," *IEEE Trans. on Image Process.*, vol. 14, no. 11, pp. 1918–1927, 2005.

[27] S. V. R. Dendi and S. S. Channappayya, "No-reference video quality assessment using natural spatiotemporal scene statistics," *IEEE Trans. on Image Process.*, vol. 29, pp. 5612–5624, 2020.

[28] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, 2015.

[29] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," *arXiv preprint arXiv:2207.12396*, 2022.

[30] A. Gumaei, W. N. Ismail, M. R. Hassan, M. M. Hassan, E. Mohamed, A. Alelaiwi, and G. Fortino, "A decision-level fusion method for covid-19 patient health prediction," *Big Data Research*, vol. 27, p. 100287, 2022.

[31] D. Yi, J. Su, and W.-H. Chen, "Probabilistic faster r-cnn with stochastic region proposing: Towards object detection and recognition in remote sensing imagery," *Neurocomputing*, vol. 459, pp. 290–301, 2021.

[32] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1733–1740.

[33] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, 2018.

[34] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "Maniqa: Multi-dimension attention network for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1191–1200.

[35] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 5128–5137.

[36] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Trans. on Image Process.*, vol. 27, no. 8, pp. 3998–4011, 2018.

[37] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3575–3585.

[38] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 pirm challenge on perceptual image super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[39] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "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*, 2022, pp. 1220–1230.