Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Health

CDMIQA: A Cross-Domain Perceptual Method and Benchmark Dataset for Medical Image Quality Assessment

Leilei Huang[1], Yue Sun[1], Mingxiang Wu[2], Wei Ke[1], Siyi Xun[1], Muzhen He[3] *, Tao Tan[1] *
[1]Macao Polytechnic University, China
[2]ShenZhen People's Hospital, China
[3]Fujian Provincial Hospital, China
{p2527042, yuesun, wke, p2214963, taotan}@mpu.edu.mo,
wu.mingxiang@szhospital.com, dr_muzhen@fjmu.edu.cn

Abstract

Medical image quality assessment (IQA) serves as a critical safeguard for precise clinical diagnosis and treatment. However, existing methods still face challenges arising from data scarcity and heterogeneity across imaging domains, which confine solutions to domain-specific designs and limit their cross-domain generalization ability. In response, we construct a multi-domain and multi-organ dataset comprising 9,105 2D and 3D medical images across three imaging domains and 18 organs, annotated by radiologists. Building upon this, we propose a cross-domain universal medical IQA method termed CDMIQA, which integrates efficient feature extractors with Hierarchical Perceptual Encoding Modules to capture and refine multi-level perceptual features while mitigating interference from noise and artifacts. Notably, the Adaptive Semantic Perception Module is designed to extract semantic features; it enhances adaptability to domain-specific images by dynamically encoding quality variations across different imaging domains and organs. Furthermore, a Top-down Feature Fusion Module is utilized to progressively aggregate features under the guidance of semantic information, reinforcing the model's feature representation capability. Experimental results on the proposed dataset demonstrate that our method outperforms standard competitors and exhibits superior generalization ability across diverse domains. Our code and dataset are available at https://github.com/Leilei-Huang-work/CDMIQA.

1 Introduction

With the rapid development of medical imaging technologies such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), the dependence on high-quality medical images for clinical diagnosis and treatment continues to increase. Suboptimal image quality not only reduces the confidence of physicians in their diagnoses and raises the risk of clinical misdiagnosis [Chow and Paramesran, 2016; Lévêque *et al.*, 2021; Cavaro-Ménard *et al.*, 2010], but also significantly limits the performance of artificial intelligence (AI)-assisted diagnostic technologies, such as medical image segmentation and lesion detection. Therefore, there is an urgent need to develop efficient medical image quality assessment (IQA) algorithms to support precise clinical diagnosis and treatment.

Objective IQA methods can be categorized based on the information available from reference images into Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) IQA. Since it is often difficult to obtain a perfect reference image for medical images in real-world applications, NR IQA has unique research value in medical imaging [Xun *et al.*, 2025]. Traditional NR IQA methods mainly rely on hand-crafted feature design to address specific distortion types [Saad *et al.*, 2012; Mittal *et al.*, 2012a; Mittal *et al.*, 2012b]. However, these methods face limitations due to their reliance on prior knowledge for manual feature design, making it challenging to represent the complex distortion characteristics of multi-domain medical images.

In recent years, deep learning methods have demonstrated excellent performance in natural IQA. This has motivated numerous studies to adapt these techniques for medical imaging environments [Li *et al.*, 2018; Oszust *et al.*, 2020a]. However, the distinctive characteristics of medical images—such as modality-specific noise and artifact patterns, complex anatomical structures, and the often subtle manifestations of pathological changes—introduce challenges that go beyond those encountered in natural-image assessment. This necessitates the development of specialized deep learning models for different domains, most notably CT-IQA [Gong *et al.*, 2019; Lee *et al.*, 2022; Shi *et al.*, 2024], Retinal-IQA [Zago *et al.*, 2018; Fu *et al.*, 2019; Shi *et al.*, 2022], and MRI-IQA [Qi *et al.*, 2021; Simi *et al.*, 2021; Stepień and Oszust, 2023]. Despite the promising potential of these models, two key factors hinder their performance and generalization in more complex and diverse clinical environments. Primarily, high-quality annotated datasets remain scarce, characterized by limited sample sizes and restricted to single-domain distributions. Furthermore, the development and benchmarking of most methods are still constrained by the inherent deficiencies of these existing datasets, which primarily focus on single-task eval-

*Corresponding authors: Muzhen He, Tao Tan

uation, severely limiting their broader applicability in clinical applications. To address the aforementioned challenges, we first construct a new large-scale cross-domain and cross-organ medical IQA benchmark dataset. Building upon this dataset, we propose a cross-domain perceptual IQA that can dynamically adapt to different medical imaging domains. In summary, the primary contributions are as follows:

- We propose CDMIQA, a cross-domain perceptual image quality assessment method specifically designed to address the challenges posed by heterogeneity across diverse medical imaging domains. By dynamically adapting to domain-specific characteristics, CDMIQA demonstrates exceptional cross-domain adaptability and robust generalization performance.
- We construct a large-scale medical IQA benchmark dataset with 9,105 2D and 3D images from MRI, CT, and Retinex domains, covering 18 organs with radiologist-quality annotations.
- Since medical imaging often involves various complex artifacts and noise, Hierarchical Perceptual Encoding Modules are introduced to capture low, medium, and semantic features to mitigate the interference of artifacts and noise and enhance the robustness of the quality assessment task.
- To enable dynamic adaptation across multi-domain and multi-organ medical IQA tasks, we propose an Adaptive Semantic Perception module to extract semantic features that capture domain and organ-specific characteristics for improved generalization. Additionally, we introduce a Top-down Feature Fusion Module to progressively aggregate features under the guidance of semantic information, reinforcing the model's feature representation capability.

2 Related Work

2.1 Medical IQA Datasets

Despite the availability of domain-specific Medical IQA datasets for CT [Lee *et al.*, 2025], MRI [Oszust *et al.*, 2020b; Stpień *et al.*, 2021], and Retinex imaging [Zhou *et al.*, 2018], critical limitations persist. CT datasets, such as those in [Lee *et al.*, 2025], typically lack 3D volumetric scans and multi-window configurations (e.g., lung and soft tissue windows). While MRI datasets like DB1 [Oszust *et al.*, 2020b] and DB2 [Stpień *et al.*, 2021] share T2-weighted images across multiple organs, there remains a notable absence of T1-weighted multi-organ Medical IQA datasets. Furthermore, existing datasets are often developed in isolation, constrained to single domains or anatomical regions. To bridge these gaps, we present a large-scale MIQA benchmark comprising 9,105 2D and 3D images across three domains—MRI, CT, and Retinex—covering 18 organs and including both lung and soft tissue window settings. All images are annotated with expert-level quality labels.

2.2 Medical IQA Methods

Recent state-of-the-art (SOTA) deep learning-based medical IQA methods have been extensively explored in three domains: Retinex, MRI, and CT. We review representative works from each domain. In Retinex-IQA, [Zago *et al.*, 2018] proposed a method where a combination of unsupervised features from saliency maps and supervised deep features from CNNs is utilized to predict the quality level of retinal images. Subsequently, [Shi *et al.*, 2022] proposed an automatic retinal image analysis framework that combines a VGG16-based transfer learning model with automatic feature generation to assess image quality and distinguish between abnormality-related and artifact-induced degradations in color fundus images. To evaluate MRI image quality, [Qi *et al.*, 2021] proposed a 3D content-adaptive hyperparameter network that captures global spatial information in MRI images and improves performance by enhancing the network's adaptive capacity via parameter reinforcement. Furthermore, [Stepień and Oszust, 2023] proposed a method that fuses two complementary deep architectures to extract MRI-specific perceptual quality features. It leverages multi-level representations via a joint network and enhances performance using features from a retrained single model. Notably, in the realm of CT-IQA, [Shi *et al.*, 2024] simulated the active inference process of an internal generative mechanism using Denoising Diffusion Probabilistic Models (DDPM) to predict primary content. Dissimilarity maps were computed from deviations between distorted and predicted images, and both were fed into a transformer-based IQA model. However, the reliance on DDPM introduces significant computational overhead, limiting its applicability to real-time medical diagnosis. Although these methods have demonstrated promising performance, most deep learning-based IQA models are task-specific, lacking generalizability and versatility in complex clinical applications.

3 Method

3.1 Dataset Construction

As illustrated in Fig. 1, we constructed a cross-domain, diverse medical IQA dataset tailored for heterogeneous quality evaluation needs. This dataset comprises 9,105 images across both 2D and 3D domains, encompassing MRI, fundus, and CT images. It covers 18 distinct organ types, including the lungs, brain, and spine. The dataset aggregates data from five sources: **Source A (CT 3D):** We collected 2,427 routine 3D chest CT scans from the in-house dataset of Shenzhen People's Hospital. Ten radiologists evaluated each series on a 1–5 scale assessing clarity, noise, and contrast, and the final Mean Opinion Scores (MOS) were obtained by averaging these ratings. **Source B (MRI T1):** We collected 300 MRI T1-weighted images from the in-house dataset of Fujian Provincial Hospital. These images were acquired during clinically indicated routine diagnostic examinations of the shoulders, knees, brain, and lumbar spine. Three radiologists independently evaluated the quality of each image on a scale of 1 to 5, with the average MOS serving as the final label. **Source C (MRI T2):** DB1 [Oszust *et al.*, 2020b] and DB2 [Stpień *et al.*, 2021] datasets. **Source D (CT 2D):** LDCTIQA2023 [Lee *et al.*, 2025] dataset. **Source E (Fundus):** Kaggle DR Image Quality dataset [Zhou *et al.*, 2018]. All domain-specific images were annotated by radiologists or trained medical pro-

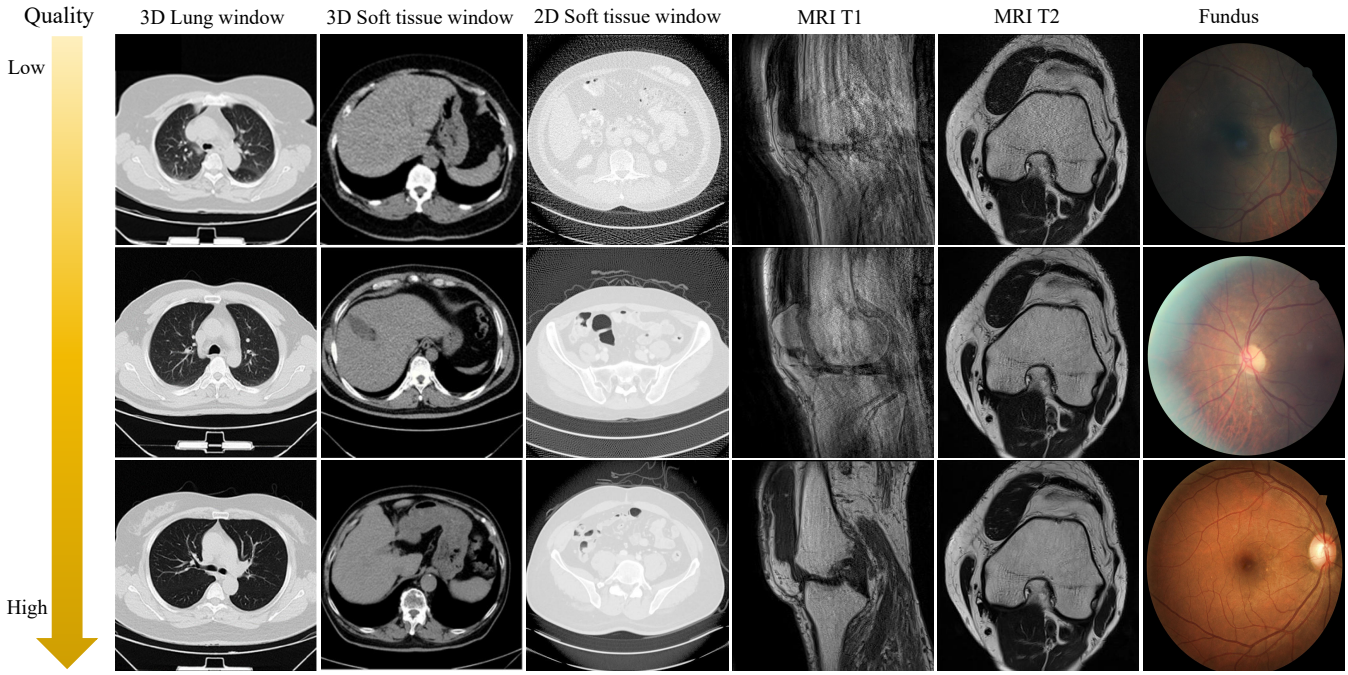

Figure 1: Examples of the proposed dataset. 3D lung window and soft tissue window images are included in Chest-3D; 2D soft tissue window images are included in LDCTIQA2023 [Lee *et al.*, 2025].

fessionals. This large-scale benchmark enables cross-domain training and standardized evaluation of medical IQA methods in diverse clinical scenarios.

3.2 Proposed IQA method

Overall Architecture

As illustrated in Fig. 2, initially, the input 2D and 3D images are processed by the preprocessing module. For 3D volumetric data, a Key Slice Selection Module is employed to extract six representative slices, prioritizing areas with high diagnostic value. These selected slices, along with the 2D images, are processed by a pre-trained Vision Transformer (ViT) classifier [Dosovitskiy *et al.*, 2020] (which achieves a classification accuracy of 99.7% on the proposed dataset) to generate a prompt set $\mathcal{P}_m$. This set encompasses spatial structure, imaging modality, anatomical region, and scan type, facilitating dynamic adaptation across diverse image domains. For feature extraction, the framework utilizes a hybrid backbone: VGG16 [Simonyan and Zisserman, 2014] extracts low-level features $\boldsymbol{F}_L$ and mid-level features $\boldsymbol{F}_M$ from its second and third blocks, respectively, while a Vision Transformer recovers semantic information from $\boldsymbol{F}_M$. Considering the impact of various noise types and artifacts (e.g., Gaussian noise, motion artifacts, and metal artifacts) on the quality degradation of medical imaging [Zhang and Metaxas, 2024], we introduce Hierarchical Perceptual Encoding Modules to refine features at multiple levels, effectively mitigating interference in IQA tasks. Specifically, the SPAM leverages spatial and pixel attention mechanisms to encode $\boldsymbol{F}_L$, enhancing the representation of salient local regions to yield $\hat{\boldsymbol{F}}_L$. In parallel, the MSPM utilizes a multi-branch parallel convolutional architecture to encode $\boldsymbol{F}_M$, capturing multi-scale structural information $\hat{\boldsymbol{F}}_M$. Furthermore, the Adaptive Semantic Perception Module incorporates two key components: (1) an Adaptive Attention Allocation mechanism, which uses the multi-scale spatial attention map from MSPM to prioritize informative regions while suppressing noisy backgrounds; and (2) a Domain-Aware Prompt strategy, which encodes the prompt set $\mathcal{P}_m$ into a unified context-aware weight matrix $\boldsymbol{W}$. This matrix acts as a multi-domain prior injected into the self-attention mechanism, guiding the Transformer to generate the semantic feature map $\boldsymbol{F}_S$, thereby enabling dynamic adaptation across multi-domain and multi-organ IQA tasks. Finally, a Top-down Feature Fusion module integrates $\boldsymbol{F}_S$, $\hat{\boldsymbol{F}}_M$, and $\hat{\boldsymbol{F}}_L$ sequentially to enhance feature representation. The fused feature map $\hat{\boldsymbol{F}}_2^F$ is then fed into an MLP to regress the final predicted quality score $\boldsymbol{S}$.

Hierarchical Perceptual Encoding Modules

SPAM. To effectively capture low-level visual features in medical images, we propose a Feature Enhancement Module (SPAM) incorporating both Spatial and Pixel Attention mechanisms. SPAM aims to enhance the representation of salient local regions while suppressing noise and artifact interference. The module employs two levels of attention: spatial attention, which accurately identifies regions affected by local distortions, and pixel attention, which emphasizes key details and mitigates noise and artifacts. Specifically, a given low-level feature map $\boldsymbol{F}_L \in R^{c_1 \times h_1 \times w_1}$ is first normalized through Layer Normalization (LN), then processed through spatial and pixel attention mechanisms individually. The outputs of these attention operations are combined via element-wise summation to produce the final enhanced feature map

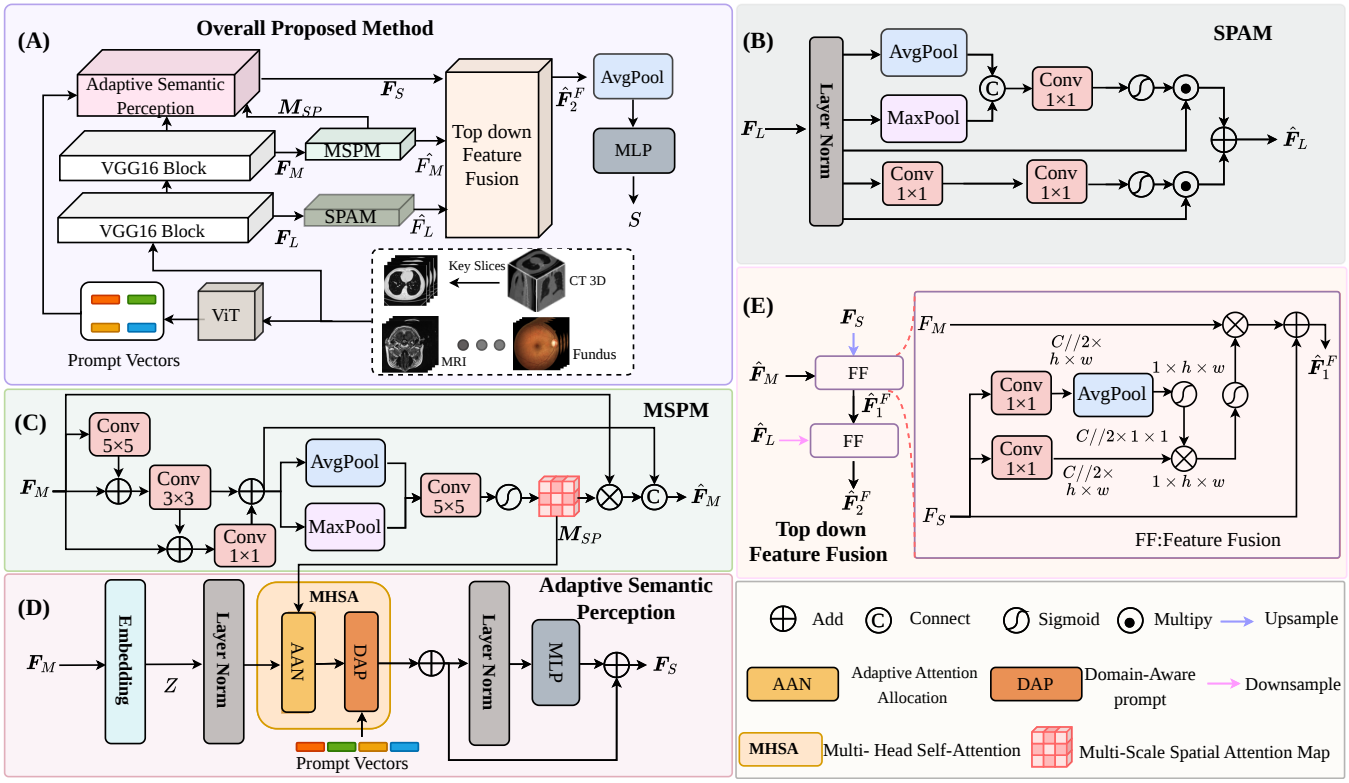

Figure 2: The framework of our proposed method. (A) Overall architecture. Hierarchical Perceptual Encoding Modules: ((B) SPAM, (C) MSPM, (D) Adaptive Semantic Perception Module) and (E) Top-down Feature Fusion. Feature F_L and F_M are extracted from the 2nd and 3rd blocks of VGG16[Simonyan and Zisserman, 2014].

$\hat{\boldsymbol{F}}_L$. Formally, this process can be expressed as follows:

$$\hat{\boldsymbol{F}}_L = \mathrm{SA}(\mathrm{LN}(\boldsymbol{F}_L)) \oplus \mathrm{PA}(\mathrm{LN}(\boldsymbol{F}_L)), \tag{1}$$

where $\mathrm{LN}(\cdot)$ denotes Layer Normalization, and $\mathrm{SA}(\cdot)$ and $\mathrm{PA}(\cdot)$ represent spatial attention and pixel attention.

MSPM. In medical imaging, key features such as organ structures and tissue boundaries exhibit clear multi-scale characteristics. To effectively capture these structural variations across different scales, we propose the Multi-Scale Structural Perception Module (MSPM), which utilizes a multi-branch parallel convolutional architecture. The central idea of MSPM is to employ multiple convolutional layers with different kernel sizes—namely 1×1, 3×3, and 5×5—in parallel, enabling the extraction of multi-scale features simultaneously. Each convolutional branch focuses on distinct spatial details, facilitating comprehensive encoding of structural information across scales. Specifically, a given intermediate feature map $\boldsymbol{F}_M \in R^{c_2 \times h_2 \times w_2}$ is processed through these parallel convolutional layers to generate multi-scale feature representations $(\boldsymbol{F}_1^M, \boldsymbol{F}_3^M, \boldsymbol{F}_5^M)$. These features are concatenated along the channel dimension to form a rich multi-scale feature map. To enhance the spatial focus, the combined features undergo spatial pooling using both max pooling and average pooling, followed by a 5×5 convolution. The resulting feature is then modulated via element-wise multiplication with a sigmoid-activated multi-scale spatial attention map to emphasize salient regions. The formulation of this process is as follows:

$$\begin{cases} \boldsymbol{F}_5^M = \mathrm{C}_5(\boldsymbol{F}_M) \\ \boldsymbol{F}_3^M = \mathrm{C}_3(\boldsymbol{F}_M \oplus \boldsymbol{F}_5^M) \\ \boldsymbol{F}_1^M = \mathrm{C}_1(\boldsymbol{F}_M \oplus \boldsymbol{F}_3^M) \end{cases} \tag{2}$$

$$\begin{cases} \boldsymbol{M}_{SP} = \sigma\left(\mathrm{C}_5\left(\mathrm{Pool}\left(\mathrm{Concat}\left(\boldsymbol{F}_1^M, \boldsymbol{F}_3^M, \boldsymbol{F}_5^M\right)\right)\right)\right) \\ \hat{\boldsymbol{F}}_M = \mathrm{Concat}\left(\boldsymbol{F}_M \odot \boldsymbol{M}_{SP}, \boldsymbol{F}_1^M, \boldsymbol{F}_3^M, \boldsymbol{F}_5^M\right) \end{cases} \tag{3}$$

where $\boldsymbol{M}_{SP}$ represents the multi-scale spatial attention map, $\mathrm{Concat}(\cdot)$ represents the concatenation operation, $\mathrm{Pool}(\cdot)$ represents the avg-pooling and max-pooling operations. C_n represents the convolutional layer with kernel size $n \in \{1, 3, 5\}$, and $\sigma(\cdot)$ represents the Sigmoid activation function.

Adaptive Semantic Perception Module. Accurate extraction of semantic information is pivotal for medical image quality assessment. However, the inherent heterogeneity of medical imaging data (e.g., varying modalities, anatomical regions, and scanning protocols) challenges the generalization capability of standard CNNs. Furthermore, non-diagnostic noise and artifacts tend to accumulate with increasing network depth, corrupting high-level semantic features. To address these issues, we replace the final convolutional stage of VGG16 with a Vision Transformer-based module. Lever-

aging the global receptivity of Transformers and scalability, we introduce two novel components: an *Adaptive Attention Allocation* mechanism to suppress noise and artifacts, and a *Domain-Aware Prompt* strategy to incorporate structured prior knowledge.

1) Adaptive Attention Allocation. This allocation block is designed to mitigate accumulation of noise and artifacts that typically arises with increasing network depth. This mechanism is conceptually inspired by clinical diagnostic practices, wherein radiologists dynamically modulate their attention across image regions—prioritizing informative areas while suppressing irrelevant or noisy background information. First, we sort the tokens based on the multi-scale spatial attention map $\boldsymbol{M}_{SP}$ in ascending order of saliency. We then evenly divide them into n sub-regions denoted as $\mathcal{G}_i$, where $i \in \{1, 2, \ldots, n\}$, thereby forming an allocation attribution map $\mathcal{G}$. The intermediate features $\boldsymbol{F}_M$, after being projected and encoded with positional encoding into $\boldsymbol{Z}$, are partitioned into subsets $\boldsymbol{Z}_i$ corresponding to $\mathcal{G}_i$. Subsequently, these feature subsets are dynamically aggregated based on group-specific aggregation rates $\boldsymbol{\theta}_i$. Crucially, regions with higher importance are assigned a smaller aggregation rate (implying fewer tokens are merged) to preserve fine-grained details, whereas background regions undergo higher aggregation to reduce redundancy. The adaptive attention allocation mechanism is formulated as follows:

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{Z}\boldsymbol{W}^Q \\ \boldsymbol{K} = \mathcal{A}(\boldsymbol{Z}, \mathcal{G}, \boldsymbol{\theta})\boldsymbol{W}^K \\ \boldsymbol{V} = \mathcal{A}(\boldsymbol{Z}, \mathcal{G}, \boldsymbol{\theta})\boldsymbol{W}^V \\ \mathcal{A}(\boldsymbol{Z}, \mathcal{G}, \boldsymbol{\theta}) = \mathrm{Cat}_{i=1}^{n}\left(\mathrm{FC}\left(\boldsymbol{Z}_i, \boldsymbol{\theta}_i\right)\right) \end{cases} \tag{4}$$

where $\mathcal{A}(\cdot)$ represents the adaptive aggregation module, which performs importance-based aggregation via fully connected layers. $\mathrm{Cat}(\cdot)$ denotes the concatenation operator. The input features $\boldsymbol{Z}$ are projected into Query $\boldsymbol{Q}$, Key $\boldsymbol{K}$, and Value $\boldsymbol{V}$. Note that only $\boldsymbol{K}$ and $\boldsymbol{V}$ are aggregated, while $\boldsymbol{Q}$ maintains full resolution during attention calculation, significantly reducing computational complexity while filtering out background noise.

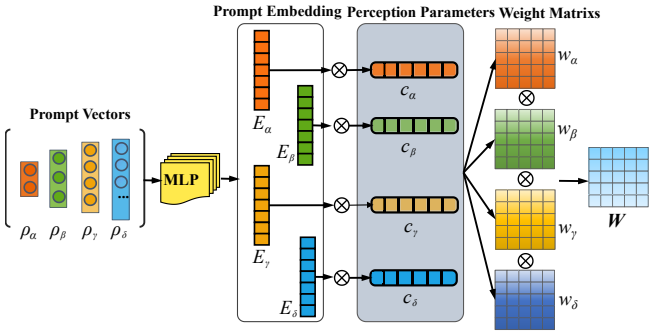

Figure 3: The Overview of the Domain-Aware Prompt Strategy.

2) Domain-Aware Prompt. To enhance generalization across diverse medical domains, as illustrated in Fig. 3, we subsequently propose a Domain-Aware Prompt strategy. This module constructs structured prior knowledge using a set of four orthogonal prompt vectors $\mathcal{P}_m$, where $m \in \{\alpha, \beta, \gamma, \delta\}$. Specifically, the spatial prompt $\mathcal{P}_\alpha$ encodes spatial structures (2D slices, 3D volumes), the modality prompt $\mathcal{P}_\beta$ characterizes imaging physics (CT, MRI, Fundus), the organ region prompt $\mathcal{P}_\gamma$ identifies anatomical regions (e.g., brain, abdomen, Knee), and the type prompt $\mathcal{P}_\delta$ describes imaging parameters (lung window, soft tissue window, T1/T2 sequences). These vectors are initialized as one-hot vectors indicating distinct prompt options. First, each vector is independently mapped through a separate MLP to obtain a prompt embedding. Subsequently, perception parameters $\boldsymbol{c}_m$ are applied via matrix multiplication to generate specific weight matrices $\boldsymbol{w}_m$. These matrices are then fused into a unified weight matrix $\boldsymbol{W}$ via element-wise multiplication to capture the joint distribution of domain attributes. This process is expressed as follows:

$$\begin{cases} \mathcal{P}_m \in R^{\kappa_m \times 1} \\ \boldsymbol{E}_m = \mathrm{MLP}_m(\mathcal{P}_m), \quad \boldsymbol{E}_m \in R^{d \times 1} \\ \boldsymbol{w}_m = \boldsymbol{E}_m \boldsymbol{c}_m, \quad \boldsymbol{c}_m \in R^{1 \times d} \\ \boldsymbol{W} = \boldsymbol{w}_\alpha \odot \boldsymbol{w}_\beta \odot \boldsymbol{w}_\gamma \odot \boldsymbol{w}_\delta, \quad \boldsymbol{W} \in R^{d \times d} \end{cases} \tag{5}$$

where κ_m denotes the number of options for prompt type m, and d denotes the feature dimension of the transformer block. The unified weight matrix $\boldsymbol{W}$ is integrated into the attention mechanism through a parametrized bilinear weighting scheme, formalized as:

$$\boldsymbol{Z}_S = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}(\boldsymbol{W}\boldsymbol{K}^T)}{\sqrt{d}}\right)\boldsymbol{V}. \tag{6}$$

Subsequently, the feature map undergoes r layers of Multi-Head Self-Attention (MHSA) equipped with these mechanisms, followed by a MLP module. The residual connection is defined as:

$$\begin{cases} \boldsymbol{Z}_S^r = \mathrm{MHSA}\left(\mathrm{LN}(\boldsymbol{Z}_S^{r-1})\right) + \boldsymbol{Z}_S^{r-1} \\ \hat{\boldsymbol{Z}}_S = \mathrm{MLP}\left(\mathrm{LN}(\boldsymbol{Z}_S^r)\right) + \boldsymbol{Z}_S^r \end{cases} \tag{7}$$

Finally, $\hat{\boldsymbol{Z}}_S$ is reshaped to form semantic feature map, denoted as $\boldsymbol{F}_S \in R^{c_3 \times h_3 \times w_3}$.

Top-down Feature Fusion

We designed a Top-down Feature Fusion strategy, which sequentially integrates features from semantic, mid-level, and low-level representations. Guided by semantic features, this approach enhances the perception of key regions and facilitates cross-layer information complementarity. Specifically, we first perform the fusion of high-level semantics with mid-level features. We generate an enhanced representation $\hat{\boldsymbol{F}}_S$ by applying convolution and upsampling to the semantic feature $\boldsymbol{F}_S$. This generates a gating mask $\boldsymbol{\Phi}_S$:

$$\boldsymbol{\Phi}_S = \sigma\left(\mathrm{C}_1\left(\sigma\left(\mathrm{C}_1\left(\mathrm{AvgPool}\left(\hat{\boldsymbol{F}}_S\right)\right)\right)\right) \odot \mathrm{C}_1\left(\hat{\boldsymbol{F}}_S\right)\right) \tag{8}$$

Subsequently, $\boldsymbol{F}_M$ is modulated by the weights $\boldsymbol{\Phi}_S$, and the fused features are obtained as:

$$\hat{\boldsymbol{F}}_1^F = \hat{\boldsymbol{F}}_M \odot \boldsymbol{\Phi}_S + \hat{\boldsymbol{F}}_S \tag{9}$$

Dataset	Metric	BRISQUE	NIQE	CNNIQA	DBCNN	MANIQA	TOPIQ	Q-Instruct	**Proposed**
Chest-3D (CT 3D)	SRCC↑	0.2234	0.1022	0.4641	0.4675	0.7072	0.5223	0.2702	**0.7325**
	PLCC↑	0.3926	0.0872	0.4990	0.4882	0.7455	0.5482	0.4075	**0.7616**
	RMSE↓	0.5032	0.7938	0.2530	0.2752	0.1276	0.2321	–	**0.1251**
LDCTIQA 2023 (CT 2D)	SRCC↑	0.7466	-0.3999	0.8661	0.9582	0.9402	0.9432	0.7694	**0.9732**
	PLCC↑	0.7852	-0.4398	0.8866	0.9535	0.9482	0.8512	0.8583	**0.9801**
	RMSE↓	0.1825	2.5408	0.1036	0.0681	0.0692	0.0882	–	**0.0577**
MRI T1	SRCC↑	0.3259	0.1801	0.5663	0.6532	0.6775	0.5905	0.4253	**0.7195**
	PLCC↑	0.3558	0.2092	0.5886	0.6801	0.6882	0.6238	0.4705	**0.7102**
	RMSE↓	0.8082	0.9055	0.5593	0.4339	0.4210	0.5072	–	**0.3938**
DB1 (MRI T2)	SRCC↑	0.2898	0.2018	0.6231	0.6594	**0.8032**	0.7157	0.3235	0.7572
	PLCC↑	0.5011	0.3072	0.7532	0.7979	**0.8259**	0.7985	0.4179	0.7836
	RMSE↓	0.5658	0.6785	0.4411	0.4282	**0.4021**	0.4283	–	0.4357
DB2 (MRI T2)	SRCC↑	0.8599	0.3504	0.9012	0.9236	0.9229	0.9202	0.6059	**0.9305**
	PLCC↑	0.9013	0.3927	0.9189	0.9352	0.9402	0.9315	0.6592	**0.9433**
	RMSE↓	0.4076	0.8231	0.3754	0.3185	0.2958	0.3287	–	**0.2791**
Kaggle DR (Fundus)	SRCC↑	0.6882	0.3225	0.7842	0.7532	0.8552	0.7329	0.4234	**0.8722**
	PLCC↑	0.6463	0.4126	0.7734	0.7449	**0.9322**	0.6988	0.4019	0.9216
	RMSE↓	0.4215	0.6988	0.3002	0.3629	0.1795	0.2901	–	**0.1702**

Table 1: Performance comparison on the Proposed dataset. Best results are highlighted in **bold**.

The fusion from mid-level to low-level features follows a similar process. First, the mid-level feature $\hat{\boldsymbol{F}}_2^F$ is aligned with the low-level feature $\hat{\boldsymbol{F}}_L$, consistent with the previous step, to compute the weight $\boldsymbol{\Phi}_M$. Subsequently, a weighted sum is performed to obtain the fused feature $\hat{\boldsymbol{F}}_2^F$. Finally, $\hat{\boldsymbol{F}}_2^F$ is passed through a global average pooling layer and MLP to generate the final quality score $\boldsymbol{S}$:

$$\boldsymbol{S} = \mathrm{FC}_1\left(\mathrm{FC}_{128}\left(\mathrm{FC}_{256}\left(\mathrm{AvgPool}\left(\hat{\boldsymbol{F}}_2^F\right)\right)\right)\right) \quad (10)$$

where FC_n denotes a fully connected layer with n output channels. The network is trained using L_2 loss.

4 Experiments

4.1 Implementation Details

All experiments were conducted on a server equipped with an Intel Xeon W-2245 CPU and two NVIDIA RTX A6000 GPUs. The proposed model was implemented using the PyTorch 1.10 with Python 3.9. During the training phase, the batch size was set to 32, and the learning rate was fixed at 2×10^{-4}. The model was optimized using the Adam optimizer with a weight decay of 0.05 for a total of 50 epochs. The dataset was randomly split into training and testing sets with a ratio of 8:2. To evaluate the generalization capability across domains, data from all sub-domains were aggregated to train a unified model. Furthermore, to quantitatively evaluate the performance of the proposed method, three widely used metrics were employed: Spearman Rank Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), and Root Mean Squared Error (RMSE).

4.2 Performance Comparison

To evaluate the performance of the proposed method, we selected seven representative IQA metrics for comparison, including traditional methods (BRISQUE [Mittal *et al.*, 2012a], NIQE [Mittal *et al.*, 2012b]), deep learning methods (CNNIQA [Kang *et al.*, 2014], DBCNN [Zhang *et al.*, 2018a], MANIQA [Yang *et al.*, 2022], TOPIQ [Chen *et al.*, 2024]), and the multi-modality large language model Q-Instruct [Wu *et al.*, 2024]. In the implementation, all comparative IQA methods were executed using the source codes released by the authors. The experimental results on domain-specific datasets, summarized in Table 1, demonstrate that the proposed method consistently outperforms existing IQA metrics. Specifically, the method ranks first on Chest-3D, LDCTIQA2023, DB2, and Kaggle-DR, and second on DB1, indicating a high degree of consistency with human subjective perception. Traditional IQA methods relying on handcrafted features (e.g., BRISQUE, NIQE) exhibit the poorest performance due to their limited representational capacity. In contrast, deep learning-based methods achieve considerably better results, largely attributed to their superior ability in feature extraction and representation learning. Furthermore, although Q-Instruct has achieved remarkable capabilities in low-level visual image quality assessment through large-scale fine-tuning, there remains significant room for improvement in medical image quality assessment. Notably, MANIQA, a transformer-based IQA method renowned for its global feature modeling capability, attains SOTA performance on DB1. However, these methods remain inferior to the proposed method, primarily because they lack cross-domain adaptive evaluation mechanisms.

4.3 Ablation Study

To investigate the contribution of each individual component, we conducted ablation experiments separately on the five subsets. The variations include: (1) without the SPAM module (w/o SPAM); (2) without the MSPM module (w/o MSPM); (3) without the Adaptive Semantic Perception module (w/o ASPM); and (4) without multi-level features, using only the final level (w/o MultiLevel). Table 2 lists the results in terms

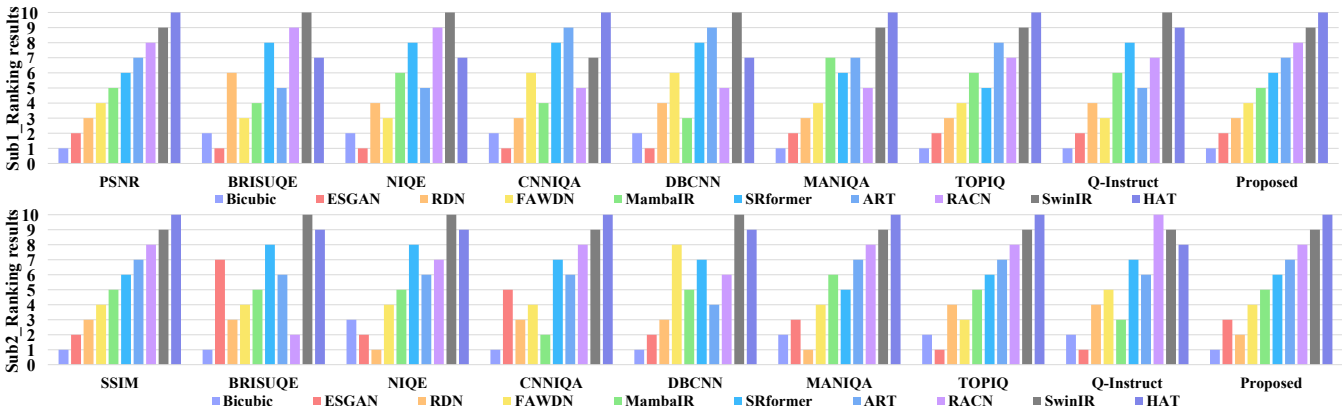

Figure 4: The ranking results of MRI-SR algorithms on two subsets of IXI T1 dataset, according to FRIQA scores and NRIQA scores.

Components	All Dataset		
	SRCC	PLCC	RMSE
w/o SPAM	0.8005	0.8204	0.2806
w/o MSPM	0.7927	0.8133	0.3032
w/o ASPM	0.7801	0.7974	0.2867
w/o MultiLevel	0.6531	0.6731	0.3760
w/ ALL Comp	**0.8308**	**0.8501**	**0.2436**

Table 2: Average ablation results on the proposed dataset.

of PLCC, SRCC, and RMSE for these variations. It can be observed that eliminating any quality component leads to a drop in the final average performance across all datasets, indicating that each component contributes to improving the overall performance. In particular, the w/o MultiLevel method shows a significant performance drop. This emphasizes the importance of multi-level fusion in medical image quality assessment, as it captures a wider range of distortion characteristics and aligns with the cognitive process of multi-level visual information fusion and analysis in human vision.

4.4 Performance Evaluation on MRI Image Super-Resolution Task

An excellent medical IQA method is essential not only for predicting image quality but also for evaluating medical image reconstruction techniques, such as super-resolution (SR). FR-IQA methods, known for their robustness, currently dominate SR performance benchmarking. However, their reliance on pristine reference images limits their applicability in real-world clinical settings, where such references are often unavailable. Thus, we evaluated the effectiveness of the proposed method specifically for assessing MRI SR performance. Our experiments benchmarked ten prominent MRI SR algorithms: Bicubic, ESGAN [Wang *et al.*, 2018], RDN [Zhang *et al.*, 2018c], FAWDN [Chen *et al.*, 2020], MambaIR [Guo *et al.*, 2024], SRformer [Zhou *et al.*, 2023], ART [Zhang *et al.*, 2022], RCAN [Zhang *et al.*, 2018b], SwinIR [Liang *et al.*, 2021], and HAT [Chen *et al.*, 2023]. The gold standards were established using FR-IQA metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). We selected 400 LR–HR image pairs from the IXI T1 dataset, which were further divided into two equal subsets of 200 pairs for separate evaluation using PSNR and SSIM. As shown in Fig. 4, the quality scores produced by our NR-IQA method show stronger consistency with those obtained via PSNR and SSIM, successfully distinguish the performance rankings among the different SR algorithms, mirroring the rankings derived from PSNR and SSIM. These results demonstrate that the proposed method represents a promising alternative as a no-reference benchmark metric for medical image super-resolution tasks, particularly in scenarios where reference images are unavailable.

5 Conclusion

We propose CDMIQA in this paper, a cross-domain perceptual image quality assessment method designed to handle the challenges of heterogeneous quality assessment across diverse medical imaging domains. First, we construct a multi-domain, multi-organ medical IQA dataset annotated by radiologists to enhance generalization ability and serve as a benchmark. Based on this, CDMIQA incorporates backbone networks and Hierarchical Perceptual Encoding Modules to extract and refine multi-level perceptual features while suppressing interference from artifacts and noise. Notably, we propose Adaptive Semantic Perception Module, which integrates an Adaptive Attention Allocation mechanism and a Domain-Aware Prompt strategy to extract semantic features, dynamically distributes attention to focus on critical regions while injecting domain-specific priors into the model, thereby enabling dynamic adaptation across diverse imaging domains and organs. Finally, a Top-Down Feature Fusion Module is introduced to aggregate these multi-scale feature representations under the guidance of semantic information, ensuring precise and robust quality prediction. Experimental results on the constructed dataset demonstrate superior performance and robust generalization of our method. In the future, we expect to extend CDMIQA into a multi-task learning paradigm that produces both global quality scores and localized degradation maps to offer superior guidance for clinical diagnosis and AI-assisted diagnostic technologies like medical image reconstruction and segmentation.

Acknowledgements

This work was supported in part by the Science and Technology Development Fund of Macao under Grant 0014/2025/RIC, and in part by the Macao Polytechnic University under Grant RP/FCA-16/2025.

References

[Cavaro-Ménard *et al.*, 2010] Christine Cavaro-Ménard, Lu Zhang, and Patrick Le Callet. Diagnostic quality assessment of medical images: Challenges and trends. In *2010 2nd European Workshop on Visual Information Processing (EUVIP)*, pages 277–284. IEEE, 2010.

[Chen *et al.*, 2020] Lihui Chen, Xiaomin Yang, Gwanggil Jeon, Marco Anisetti, and Kai Liu. A trusted medical image super-resolution method based on feedback adaptive weighted dense network. *Artificial Intelligence in Medicine*, 106:101857, 2020.

[Chen *et al.*, 2023] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22367–22377, 2023.

[Chen *et al.*, 2024] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024.

[Chow and Paramesran, 2016] Li Sze Chow and Raveendran Paramesran. Review of medical image quality assessment. *Biomedical signal processing and control*, 27:145–154, 2016.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fu *et al.*, 2019] Huazhu Fu, Boyang Wang, Jianbing Shen, Shanshan Cui, Yanwu Xu, Jiang Liu, and Ling Shao. Evaluation of retinal image quality assessment networks in different color-spaces. In *International conference on medical image computing and computer-assisted intervention*, pages 48–56. Springer, 2019.

[Gong *et al.*, 2019] Hao Gong, Lifeng Yu, Shuai Leng, Samantha K Dilger, Liqiang Ren, Wei Zhou, Joel G Fletcher, and Cynthia H McCollough. A deep learning-and partial least square regression-based model observer for a low-contrast lesion detection task in ct. *Medical physics*, 46(5):2052–2063, 2019.

[Guo *et al.*, 2024] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *European conference on computer vision*, pages 222–241. Springer, 2024.

[Kang *et al.*, 2014] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014.

[Lee *et al.*, 2022] Wonkyeong Lee, Eunbyeol Cho, Wonjin Kim, Hyebin Choi, Kyongmin Sarah Beck, Hyun Jung Yoon, Jongduk Baek, and Jang-Hwan Choi. No-reference perceptual ct image quality assessment based on a self-supervised learning framework. *Machine Learning: Science and Technology*, 3(4):045033, 2022.

[Lee *et al.*, 2025] Wonkyeong Lee, Fabian Wagner, Adrian Galdran, Yongyi Shi, Wenjun Xia, Ge Wang, Xuanqin Mou, Md Atik Ahamed, Abdullah Al Zubaer Imran, Ji Eun Oh, et al. Low-dose computed tomography perceptual image quality assessment. *Medical Image Analysis*, 99:103343, 2025.

[Lévêque *et al.*, 2021] Lucie Lévêque, Meriem Outtas, Hantao Liu, and Lu Zhang. Comparative study of the methodologies used for subjective medical image quality assessment. *Physics in Medicine & Biology*, 66(15):15TR02, 2021.

[Li *et al.*, 2018] Ruoyu Li, Guangzhe Dai, Zhaoyang Wang, Shaode Yu, and Yaoqin Xie. Using signal-to-noise ratio to connect the quality assessment of natural and medical images. In *Tenth International Conference on Digital Image Processing (ICDIP 2018)*, volume 10806, pages 1327–1332. SPIE, 2018.

[Liang *et al.*, 2021] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[Mittal *et al.*, 2012a] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

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

[Oszust *et al.*, 2020a] Mariusz Oszust, Adam Piórkowski, and Rafał Obuchowicz. No-reference image quality assessment of magnetic resonance images with high-boost filtering and local features. *Magnetic resonance in medicine*, 84(3):1648–1660, 2020.

[Oszust *et al.*, 2020b] Mariusz Oszust, Adam Piórkowski, and Rafał Obuchowicz. No-reference image quality assessment of magnetic resonance images with high-boost filtering and local features. *Magnetic resonance in medicine*, 84(3):1648–1660, 2020.

[Qi *et al.*, 2021] Kehan Qi, Haoran Li, Chuyu Rong, Yu Gong, Cheng Li, Hairong Zheng, and Shanshan Wang. Blind image quality assessment for mri with a deep three-dimensional content-adaptive hyper-network. *arXiv preprint arXiv:2107.06888*, 2021.

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

[Shi *et al.*, 2022] Chuying Shi, Jack Lee, Gechun Wang, Xinyan Dou, Fei Yuan, and Benny Zee. Assessment of image quality on color fundus retinal images using the automatic retinal image analysis. *Scientific Reports*, 12(1):10455, 2022.

[Shi *et al.*, 2024] Yongyi Shi, Wenjun Xia, Ge Wang, and Xuanqin Mou. Blind ct image quality assessment using ddpm-derived content and transformer-based evaluator. *IEEE Transactions on Medical Imaging*, 43(10):3559–3569, 2024.

[Simi *et al.*, 2021] VR Simi, Damodar Reddy Edla, and Justin Joseph. A no-reference metric to assess quality of denoising for magnetic resonance images. *Biomedical Signal Processing and Control*, 70:102962, 2021.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Stepień and Oszust, 2023] Igor Stepień and Mariusz Oszust. No-reference image quality assessment of magnetic resonance images with multi-level and multi-model representations based on fusion of deep architectures. *Engineering Applications of Artificial Intelligence*, 123:106283, 2023.

[Stpień *et al.*, 2021] I Stpień, R Obuchowicz, A Piórkowski, and M Oszust. Fusion of deep convolutional neural networks for no-reference magnetic resonance image quality assessment. *Sensors*, 21(4):1043, 2021.

[Wang *et al.*, 2018] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[Wu *et al.*, 2024] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25490–25500, 2024.

[Xun *et al.*, 2025] Siyi Xun, Qiaoyu Li, Xiaohong Liu, Pu Huang, Guangtao Zhai, Yue Sun, Mingxiang Wu, Tao Tan, et al. Charting the path forward: Ct image quality assessment-an in-depth review. *Journal of King Saud University Computer and Information Sciences*, 37(5):1–24, 2025.

[Yang *et al.*, 2022] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1191–1200, 2022.

[Zago *et al.*, 2018] Gabriel Tozatto Zago, Rodrigo Varejao Andreao, Bernadette Dorizzi, and Evandro Ottoni Teatini Salles. Retinal image quality assessment using deep learning. *Computers in biology and medicine*, 103:64–70, 2018.

[Zhang and Metaxas, 2024] Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical image analysis*, 91:102996, 2024.

[Zhang *et al.*, 2018a] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018.

[Zhang *et al.*, 2018b] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[Zhang *et al.*, 2018c] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.

[Zhang *et al.*, 2022] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. *arXiv preprint arXiv:2210.01427*, 2022.

[Zhou *et al.*, 2018] Kang Zhou, Zaiwang Gu, Annan Li, Jun Cheng, Shenghua Gao, and Jiang Liu. Fundus image quality-guided diabetic retinopathy grading. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 245–252. Springer, 2018.

[Zhou *et al.*, 2023] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12780–12791, 2023.