

# Image Quality Assessment for UAV-based Infrastructure Inspection: An AI Pre-selection Strategy

Ya-Li Lin<sup>1\*</sup>, Guan-Chin Su<sup>1</sup>, Lai-Han Zou<sup>1</sup>, Chao-Hung Lin<sup>2</sup>, Jiann-Yeou Rau<sup>2</sup>,  
Wei-Shen Lai<sup>3</sup>, Chih-Chao Hu<sup>3</sup>

<sup>1</sup>PhD student, Department of Geomatics, National Cheng-Kung University, Taiwan

<sup>2</sup>Professor, Department of Geomatics, National Cheng-Kung University, Taiwan

<sup>3</sup>Researcher, Transportation Engineering Division, Institute of Transportation, Taiwan  
[\\*alecfree2@gmail.com](mailto:alecfree2@gmail.com)

**Abstract:** Image quality assessment (IQA) is essential for photogrammetry and computer vision tasks such as defect detection, 3D reconstruction, and infrastructure monitoring. In UAV-based inspection, however, image quality is often degraded by motion blur, unstable illumination, and platform vibration, which undermines the accuracy of downstream analysis. Current quality control still relies on manual inspection, which is labor-intensive and inconsistent. Although reference-free IQA has gained increasing attention, existing approaches largely depend on subjectively labeled datasets (e.g., mean opinion scores), limiting their objectivity and reproducibility in engineering applications. To address these challenges, an objective UAV-oriented IQA framework is proposed, which eliminates the reliance on subjective labeling. Specifically, a CLIP-based Structural Similarity Index (CSSIM) is introduced, which replaces SSIM's structure term with deep feature similarity, providing robustness against misalignment while enabling automated label generation from artificially degraded images. Based on these objective labels, a dual-model strategy is designed which consists of a probability-weighted Swin Transformer model for pixel-wise quality prediction and a Swin-Unet model for efficient full-image inference. This framework provides both high-precision quality maps and scalable pre-selection capability, supporting reliable UAV image screening for large-scale inspection tasks. Experimental results demonstrate that CSSIM consistently outperforms SSIM and its variants in scenarios involving scaling and translation, highlighting its robustness to common UAV distortions. For instance, under large-scale shifts, CSSIM maintained scores above 0.80, whereas SSIM, MS-SSIM, and CW-SSIM dropped to 0.23–0.59. In scaling tests, CSSIM preserved values above 0.93, while other indices fell below 0.62. The probability-weighted model achieved the lowest RMSE (0.043) in pixel-level predictions, whereas Swin-Unet reduced inference time by more than 70% with only marginal accuracy loss, striking a balance between precision and efficiency. Moreover, the proposed nonlinear weighting scheme improved the separability of image-level score distributions, enabling low-quality images to be effectively filtered in large-scale datasets. These findings confirm that the proposed AI pre-selection strategy is both effective and practical, offering an automated and interpretable solution for UAV-based infrastructure inspection.

**Keywords:** UAV, image quality assessment, deep learning, infrastructure monitoring

## 1. Introduction

From autonomous driving and medical imaging to UAV-based infrastructure inspection, modern engineering applications increasingly rely on visual data. In all these domains, image quality directly affects safety, accuracy, and reliability. A blurred scan may obscure diagnosis, just as a distorted UAV image may conceal a crack—underscoring a central engineering truth: reliable image quality assessment (IQA) is indispensable. IQA methods are typically grouped into three categories. Full-reference (FR) approaches compare a test

image against a pristine reference, while reduced-reference (RR) methods use only partial features. By contrast, no-reference (NR) methods assess a single image without reference. These categories trade accuracy for practicality: FR is precise but infeasible in real-world settings; RR relaxes requirements yet still needs side information. NR methods, free from reference images, are thus most applicable in in-the-wild scenarios such as UAV inspections and have become a research focal point.

Most NR-IQA benchmarks rely on subjective mean opinion scores (MOS). While MOS reflects perception, it suffers from variability, bias, and limited reproducibility—misaligned with engineering demands for objective, consistent, and auditable indicators. This creates a crucial gap: objective NR-IQA for high-stakes engineering. In UAV inspections, poor-quality images directly undermine defect detection, yet quality control is still largely manual, slow, and inconsistent. To address this, we propose an engineering-oriented NR-IQA framework tailored to UAV imagery. The framework comprises two stages: (i) introducing CSSIM (CLIP-based Structural Similarity Index) as an objective, translation-robust, size-insensitive indicator; and (ii) designing a dual-model pipeline balancing accuracy and efficiency while producing pixel-wise quality maps. Our contributions are: (i) CSSIM, alleviating dependence on subjective labels; (ii) an automated dataset-generation pipeline; (iii) a classification-based, probability-weighted model that improves prediction accuracy while preserving interpretability; (iv) pixel-wise prediction for regional visualization; and (v) demonstration of transforming FR datasets into NR-IQA models for UAV inspection.

Together, these elements yield an objective, fast, and engineering-ready NR-IQA solution for UAV imagery. The remainder of this paper is organized as follows: Section 2 reviews IQA research with emphasis on NR-IQA; Section 3 details methodology; Section 4 presents experiments; Section 5 concludes and discusses future work.

## 2. Literature Review

Since the advent of digital imaging, IQA has remained central to image processing. Its goal is to detect degradation and guide tasks such as compression, reconstruction, and denoising. Based on reference availability, IQA divides into FR, RR, and NR approaches (Zhai & Min, 2020). Early FR-IQA methods such as mean squared error (MSE) and peak signal-to-noise ratio (PSNR) were efficient but poorly matched perceptual quality. To address this, Wang and Bovik et al. (2004) proposed the Structural Similarity Index (SSIM), evaluating luminance, contrast, and structure. Variants followed, including MS-SSIM (Wang et al., 2003), CW-SSIM (Sampat et al., 2009), FSIM (Zhang et al., 2011), and GMSD

(Xue et al., 2014), incorporating multi-scale and gradient cues. Yet FR methods demand pristine references, limiting real-world deployment. RR-IQA mitigates this by transmitting partial features—e.g., color statistics, frequency coefficients, or entropy descriptors—for reconstruction and quality estimation (Redi et al., 2010; Li & Wang, 2009; Soundararajan & Bovik, 2012). Later works included structural similarity and saliency features (Rehman & Wang, 2012; Min et al., 2018). Despite better bandwidth-accuracy balance, RR still requires original features, restricting applicability.

Given these limitations, NR-IQA has become a central research direction. NR approaches predict distorted image quality without references, suiting scenarios like network transmission, surveillance, and compression. Early studies used handcrafted descriptors for blur, noise, or compression, often grounded in natural scene statistics (NSS). Representative works include BIQI (Moorthy & Bovik, 2010), BRISQUE (Mittal et al., 2012), and NIQE (Mittal et al., 2013). These were interpretable and efficient but weak under complex or cross-dataset distortions. Learning-based NR-IQA then gained prominence, from shallow models such as CORNIA (Ye et al., 2012) and QAC (Kang et al., 2014) to deep CNN and Transformer architectures. Models like BIECON (Kim et al., 2017), MEON (Ma et al., 2017), and DeepBIQ (Bosse et al., 2017) unified feature extraction and quality prediction, boosting accuracy with large-scale training. However, they relied heavily on subjective MOS, which is costly and inconsistent. To reduce reliance on MOS, researchers explored self-supervised or pseudo-supervised strategies. These typically employ FR metrics (e.g., SSIM, PSNR) or degradation models to auto-score unlabeled data, generating pseudo labels. Zhang et al. (2018) introduced RankIQA, creating distortion-level pairs and training with ranking loss, improving transferability but requiring known distortion levels. Wu et al. (2020) generated pseudo-references to approximate pristine images, enhancing accuracy though introducing bias. Liu et al. (2022) proposed pre-training with distortion-sensitive tasks, then fine-tuning with limited labeled data, cutting annotation needs. You et al. (2021) combined contrastive learning with degradation-aware pretext tasks, achieving strong predictions with few labels, though alignment with perceptual quality remains open.

Overall, these approaches reduce MOS dependence and improve practical NR-IQA applicability, offering promising directions for objective and scalable quality assessment.

### 3. Methodology

The methodology is organized into two stages, as shown in Figure 1. In the first stage, we construct an image quality dataset with objective labels. High-quality UAV images from

field inspections are preprocessed and degraded to generate images with varying quality. The proposed CSSIM (CLIP-based Structural Similarity Index) then assigns objective quality scores, producing a labeled dataset for training. In the second stage, this dataset trains two IQA models: (1) a Transformer-based high-precision model with pixel-wise inference, and (2) a Swin-UNET-based accelerated model for full-image processing. The framework outputs precise quality maps efficiently, supporting tasks such as rapid screening of low-quality UAV images and guiding occlusion filling in 3D reconstruction.

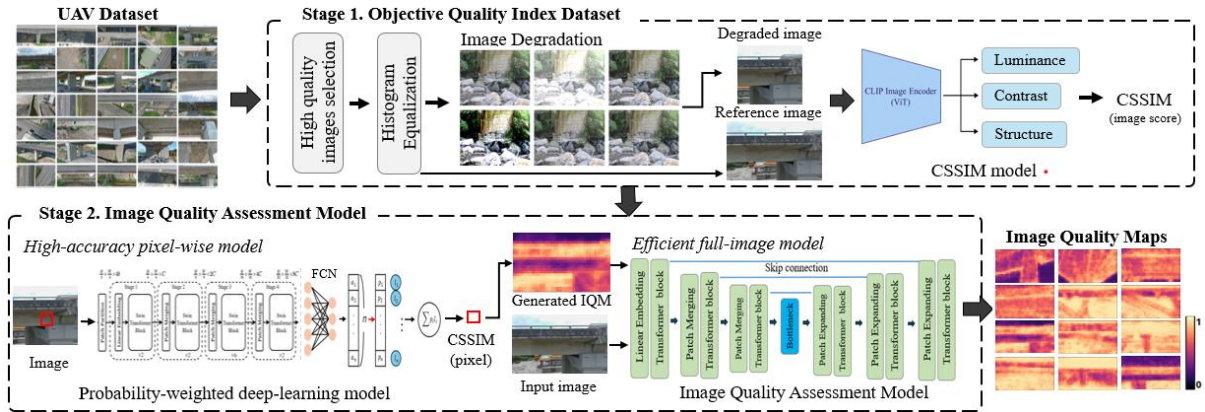


Figure 1. An AI Pre-selection Strategy workflow

### 3.1 Dataset Preparation

To construct a dataset for objective IQA training, both high-quality references and their degraded counterparts are required. Public IQA datasets generally contain natural degradations, which resemble real-world conditions but often lack references and rely on subjective scores, and artificial degradations, generated from pristine images with controllable type and severity. This study adopts artificial degradations to ensure balanced and reproducible samples. To guarantee reliable labels, reference images were first standardized using histogram equalization on the value channel in HSV space, reducing bias from brightness and contrast variation. With consistent references established, degraded images were generated through five common UAV impairments: brightness and contrast changes, Gaussian blur, and horizontal/vertical motion blur. These degradations simulate illumination shifts, weather or parameter effects, misfocus, and UAV motion artifacts, covering the main quality challenges in UAV imaging and enhancing model generalization. Artificially degraded images are shown in Figure 2.

Brightness degradation was implemented by applying a linear transformation to the brightness values of each pixel, making the image appear either brighter or darker. The adjustment is defined as Eq. (1):

$$b(x, y) = \alpha_b \cdot input(x, y) + \gamma_b, \quad (1)$$

where  $input(x, y)$  represents the pixel value of the original image at position  $(x, y)$  and  $b(x, y)$  denotes the adjusted pixel value. Here,  $\alpha_b$  and  $\gamma_b$  are the scaling factor and offset, respectively, both controlled by brightness coefficient ranging from -255 to 255.

Contrast degradation was similarly achieved through a linear transformation applied to pixel values, formulated as Eq. (2):

$$c(x, y) = \alpha_c \cdot input(x, y) + \gamma_c, \quad (2)$$

where  $c(x, y)$  is the adjusted pixel value.  $\alpha_c$  and  $\gamma_c$  represent the scaling factor and offset for contrast adjustment. Gaussian blur degradation was generated using following function, which applies a Gaussian kernel to each pixel of the image. The kernel values are computed based on the Gaussian function, shown in Eq. (3):

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \quad (3)$$

where  $G(x, y)$  is the Gaussian value at position  $(x, y)$ , and  $\sigma$  is the standard deviation controlling the kernel width. The kernel size determines the degree of blurring. After computing the kernel, convolution is applied to each pixel of the image. Horizontal Motion Blur and Vertical Motion Blur were implemented using convolution operations with specifically designed kernels. To provide a clearer and more intuitive illustration of the aforementioned degradation strategies and their effects, the original image and its artificially degraded counterparts are presented in Figure 2.

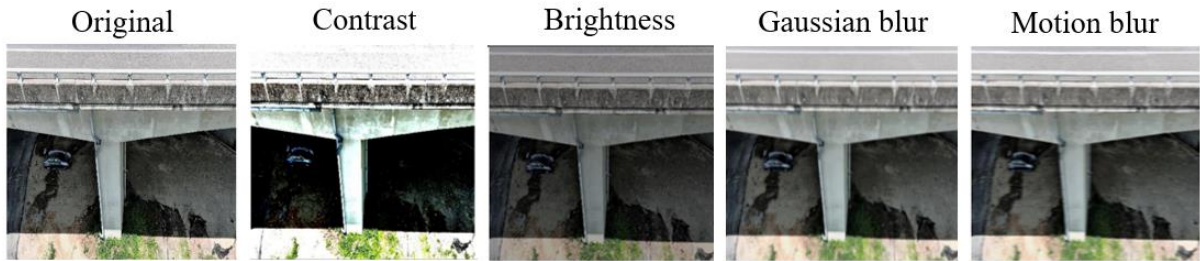


Figure 2. Examples of the original and artificially degraded images.

### 3.2 CSSIM Index Construction and Objective Label Generation

In the field of IQA, the Structural Similarity Index (SSIM) has long been regarded as a fundamental method for measuring differences between two images. The core idea is to evaluate image similarity through a combination of luminance, contrast, and structural components, which are formulated as:

$$SSIM(x, y) = [L(x, y)]^\alpha \cdot [C(x, y)]^\beta \cdot [S(x, y)]^\gamma, \quad (4)$$



where,

$$L(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \quad C(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \quad S(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}.$$

Here,  $L(x, y)$ ,  $C(x, y)$ , and  $S(x, y)$  denote the luminance, contrast, and structural similarity components, with parameters  $\alpha \cdot \beta \cdot \gamma > 0$  controlling their relative importance, and constants  $C_1$ ,  $C_2$ ,  $C_3$  ensuring numerical stability. SSIM values range from  $-1$  to  $1$ , reaching  $1$  for identical images and approaching  $-1$  for completely dissimilar ones. By modeling human sensitivity to structure, luminance, and contrast, SSIM effectively captures degradations such as blur or brightness/contrast shifts. However, its reliance on the cross-covariance term  $\sigma_{xy}$  makes it sensitive to translation, rotation, and scale, where even minor misalignments can significantly reduce scores.

To overcome this, we introduce the CLIP model (Radford, 2021) as a feature extractor. By replacing SSIM's statistical structure term with deep feature similarity, we propose the CLIP-based Structural Similarity Index (CSSIM), shown in Figure 3. CSSIM preserves SSIM's focus on structural consistency while leveraging robust semantic features, improving tolerance to translation and scale variations. Specifically, CSSIM computes cosine similarity between  $L_2$ -normalized CLIP embeddings over multi-scale, overlapping patches, reducing sensitivity to small shifts or rotations while still penalizing true structural inconsistencies.

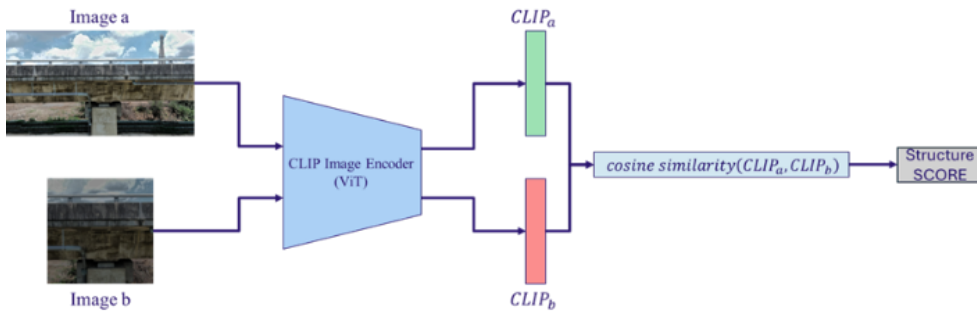


Figure 3. Illustration of the CSSIM Structural Scoring Framework.

In practice, CSSIM first extracts CLIP features from local windows of the compared images, and similarity is computed using cosine similarity, that is,

$$CLIP\_Structure(x, y) = \frac{CLIP_x \cdot CLIP_y}{\|CLIP_x\| \|CLIP_y\|}, \quad (5)$$

where  $CLIP_x$  and  $CLIP_y$  are the feature vectors produced by CLIP for the two images. Based on this, the CSSIM formulation can be rewritten as

$$CSSIM(x, y) = L(x, y) \cdot C(x, y) \cdot CLIP\_Structure(x, y). \quad (6)$$

However, cosine similarity values are typically concentrated in the high-value range (e.g., 0.7–1.0). If directly used as labels, this distribution would make it difficult for a model to distinguish different quality levels. To mitigate this, we design a mapping function  $F$ , shown in Eq. (8) to redistribute CSSIM scores into the range [0,1]:

$$F(CSSIM) = \frac{1}{1 + e^{-20(CSSIM - 0.8)}}. \quad (8)$$

This nonlinear stretching function preserves the monotonicity of the metric while producing a more uniform distribution of labels, thereby enhancing discrimination for medium- and low-quality images. Ultimately, the CSSIM labels generated through this process provide a reliable and objective basis for pixel-wise quality prediction models, effectively supporting UAV-based IQA tasks.

### 3.3 Image Quality Assessment Models

To balance prediction accuracy with the efficiency required for practical use, this study proposes a dual-model framework. The first model employs the Swin Transformer backbone with a probability-weighting strategy for pixel-wise quality prediction. It produces high-precision quality maps, serving both as a tool for theoretical validation and label generation, and as a way to transform reference-dependent metrics into a no-reference approach. The second model, based on the Swin-Unet architecture, directly generates quality maps from entire images, greatly reducing inference time and computational cost. This makes it suitable for UAV scenarios demanding real-time processing and large-scale analysis. Together, the two models complement each other, bridging high accuracy with high efficiency.

#### 3.3.1 Probability-weighted Model based on Swin Transformer

The Swin Transformer (Liu et al., 2021) was adopted as the feature extraction backbone in this study. It inherits the advantages of the self-attention mechanism introduced in the Transformer architecture (Vaswani et al., 2017), while its window partitioning and shifted-window design effectively reduce computational complexity without sacrificing global contextual modeling capability. Its hierarchical structure progressively compresses the spatial dimensions of images while enhancing feature representation power, enabling the capture of multi-scale semantic information. These properties make the Swin Transformer particularly suitable for IQA of high-resolution UAV imagery.

In terms of output design, this study reformulates the IQA task from a regression problem into a classification problem. Specifically, the CSSIM value range [0,1] is uniformly divided into 21 intervals (step size = 0.05), and the model outputs the corresponding probability distribution. After normalization, the raw outputs are converted into probabilities  $\{p_i\}$ , which are then multiplied by the numerical labels of each interval  $\{l_i\}$  and summed. As shown in Eq. (9):

$$\hat{y} = \sum_{i=1}^{21} p_i \cdot l_i. \quad (9)$$

This probability-weighting mechanism improves stability and accuracy, while mitigating vanishing gradients often encountered in regression. Labels are further smoothed with a Gaussian distribution rather than a single discrete value, allowing the model to better capture score uncertainty and enhance generalization under diverse degradations. Training uses Root Mean Squared Error (RMSE) as the loss function. Through pixel-wise inference, the model produces detailed quality maps reflecting spatial variation of local image quality, with the structure shown in Figure 4. However, its computational cost remains high, limiting direct applicability to large-scale or efficient UAV-based IQA.

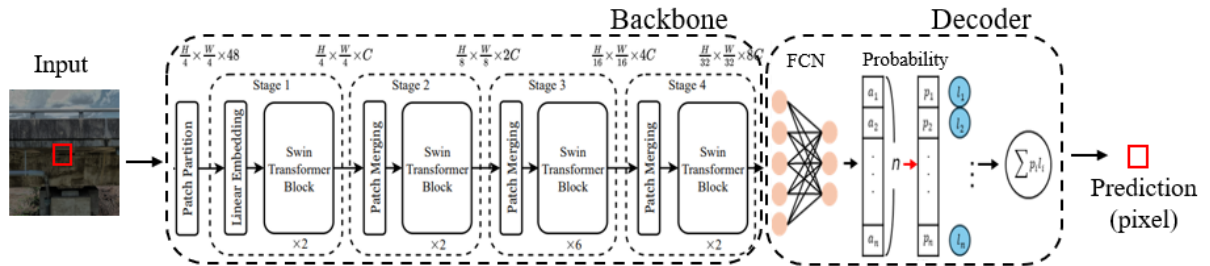


Figure 4. The architecture of Swin Transformer based probability weight model

### 3.3.2 Image Quality Assessment Model based on Swin-Unet:

To overcome the limitations of pixel-wise models in inference efficiency, this study further develops an IQA model based on Swin-Unet (Cao et al., 2021). To overcome the limitations of pixel-wise models in inference efficiency, this study further develops a fast quality map generation model based on Swin-Unet (Cao et al., 2021). Swin-Unet integrates the hierarchical feature extraction capability of the Swin Transformer with the encoder-decoder architecture of U-Net, enabling it to capture both global and local semantic information while preserving detailed features and boundaries through skip connections. Compared with the standalone Swin Transformer, Swin-Unet is more suitable for producing



full-image quality maps, achieving a good balance between computational efficiency and prediction accuracy.

For the training strategy, the quality maps generated by the probability-weighted model are employed as ground truth labels, allowing Swin-Unet to learn quality map generation in a semantic segmentation manner. With this design, Swin-Unet can directly output a corresponding quality map from the full input image, and the overall CSSIM value is then calculated through a nonlinear weighting scheme, shown in Figure 5.

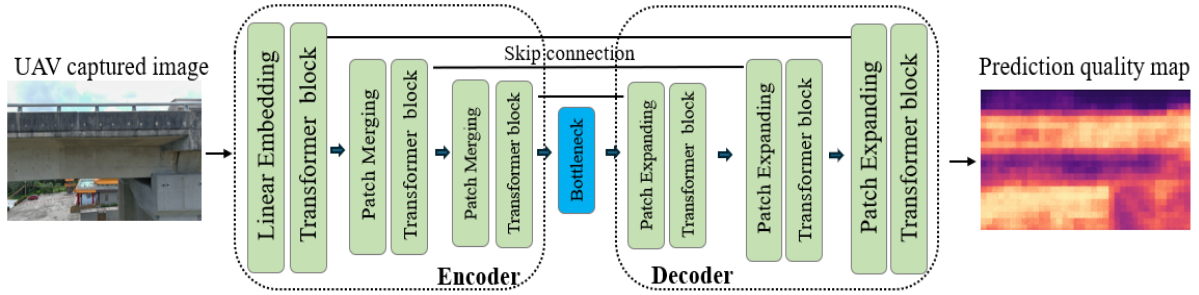


Figure 5. Swin-Unet model for IQM generation

Unlike simple averaging, nonlinear weighting emphasizes the contribution of high-quality regions while penalizing degraded regions, ensuring that the final score better reflects image quality when severe distortions are present, rather than being diluted by a large number of intact pixels, as formulated in Eq. (10). This approach significantly reduces the time required for quality map generation and lowers computational demands, enabling rapid UAV-based IQA that meets the efficiency and scalability requirements of practical engineering applications.

$$Q_{final} = \min \left( 1, \frac{\log(1 + \mathbb{E}[\tilde{i}_\gamma^{1+\gamma}])}{\log(1 + Q_{max})} \cdot (\mathbb{E}[i])^\alpha \right), \quad (10)$$

where,

$$\tilde{i}_\gamma = i\sigma(\gamma(i - 0.5)), \quad \sigma(x) = \frac{1}{1 + e^{-x}},$$

Here,  $i \in [0,1]$  denotes the pixel value of the quality map,  $\gamma > 0$  controls the strength of penalty or reward,  $Q_{max}$  is a constant,  $\mathbb{E}[\cdot]$  represents the average over the entire image, and  $\alpha$  regulates the contribution of the global mean to prevent a few high values from being diluted by a large number of medium and low values.

### 3.4 Model Training and Supervision

The training framework focuses on two models: the probability-weighted model (Swin Transformer based) and the Swin-Unet model. The primary goal of the probability-

weighted model is to transform full-reference (FR) data into no-reference (NR) predictions. Since CSSIM depends on surrounding structure, a single pixel is insufficient. Thus, for each target pixel, a 500×500 patch centered on it is extracted as input, preserving local features while avoiding the heavy cost of processing full-resolution UAV images. During training, distorted patches and their references are used to compute CSSIM scores, which serve as labels for learning the mapping “from local patch to central pixel quality score.”

After training, full NR quality maps are reconstructed through pixel-wise cropping and inference. By contrast, the Swin-Unet model takes entire images as input and uses probability-weighted quality maps as supervisory labels to predict complete quality maps. This architecture captures both global and local features while removing the cost of pixel-wise inference, thereby improving efficiency with only slight accuracy loss—well-suited for UAV scenarios requiring efficient large-scale processing. Both models adopt the RMSE as the loss function, which is defined as

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^n (\hat{y}_i - y_i)^2}, \quad (11)$$

where  $\hat{y}_i$  denotes the model prediction and  $y_i$  represents either the CSSIM label (for the probability-weighted model) or the quality map label (for the Swin-Unet model).

In terms of experimental environment and settings, all models were trained on a Windows workstation equipped with an NVIDIA GeForce RTX 4060 Ti GPU and 64 GB of memory. The probability-weighted model was trained for 50 epochs using the Adam optimizer with a learning rate of  $1 \times 10^{-5}$ . The Swin-Unet model was trained for 150 epochs using the SGD optimizer, with the same learning rate of  $1 \times 10^{-5}$ . Both models utilized the pretrained weights of *swin\_tiny\_patch4\_window7\_224\_lite*. In addition, the architecture of the Swin-Unet model was modified to support arbitrary input sizes, rather than being constrained to the original 224×224 resolution.

## 4. Experimental Results and Discussion

### 4.1 Performance of CSSIM

CSSIM, proposed in this study, integrates the CLIP image feature extractor to address the inherent limitations of conventional SSIM, which is highly sensitive to image misalignment and scale variations. Previous efforts have also sought to improve SSIM, such as Multi-Scale Structural Similarity (MS-SSIM) and Complex Wavelet Structural Similarity (CW-SSIM). MS-SSIM enhances robustness to blur and contrast changes by

employing multi-scale pyramid decomposition to capture structural information at different resolutions. CW-SSIM, on the other hand, exploits phase information in the complex wavelet domain, enabling greater consistency under small translations and rotations. Based on these considerations, SSIM, MS-SSIM, and CW-SSIM were selected as baseline methods for comparison with CSSIM. Two common distortion scenarios were designed to validate the relative performance: image scaling and image translation.

Table 1: Average score of different indexes under scaling and translation.

Distortion Type	Method	Average Score	Distortion Type	Method	Average Score
X,Y Shift 0~100 (interval=5)	<b>CSSIM</b>	<b>0.9912</b>	Scale 0.8~1.2	<b>CSSIM</b>	<b>0.9428</b>
	SSIM	0.5894		SSIM	0.5578
	MS-SSIM	0.5287		MS-SSIM	0.5476
	CW-SSIM	0.6921		CW-SSIM	0.6241
X,Y Shift 0~2000 (interval=100)	<b>CSSIM</b>	<b>0.8000</b>	Scale 1.0~2.0	<b>CSSIM</b>	<b>0.9298</b>
	SSIM	0.2346		SSIM	0.5875
	MS-SSIM	0.4678		MS-SSIM	0.5141
	CW-SSIM	0.4682		CW-SSIM	0.5788

Table 1 summarizes the average scores of all indexes under different experimental conditions. CSSIM consistently outperformed all baselines across settings: in the scaling test, CSSIM achieved averages of 0.93–0.94, while in the translation test it reached 0.99 (shift = 100) and 0.80 (shift = 2000). These results are substantially higher than SSIM (0.23–0.59), MS-SSIM (0.47–0.52), and CW-SSIM (0.46–0.69).

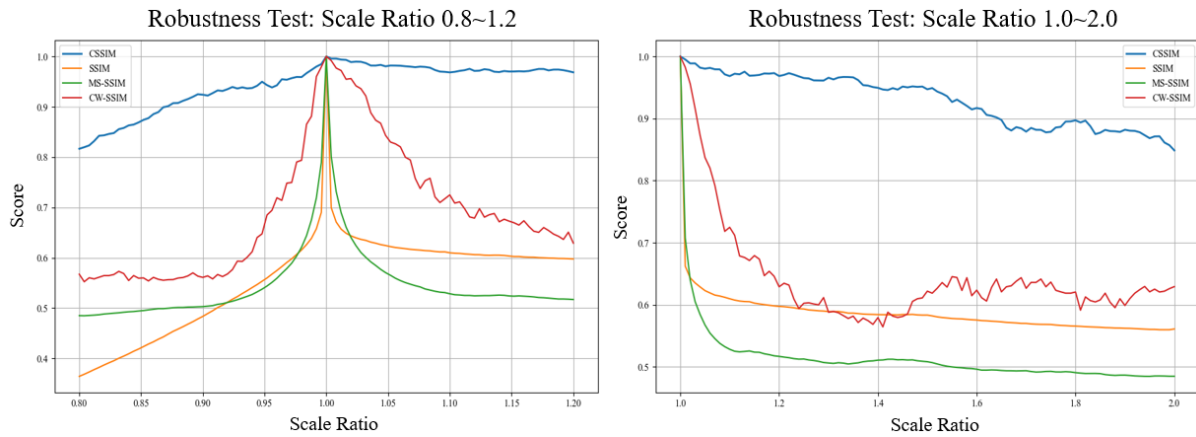


Figure 6. IQA indexes under scaling. Left: scale ratio from 0.8 to 1.2. Right: scale ratio from 1.0 to 2.0.

In the scaling experiment, test images were resized within the ranges of scale factor 0.8–1.2 and 1.0–2.0, and the score curves of each metric were plotted, as shown in Figure 6, both SSIM and MS-SSIM dropped sharply to the range of 0.5–0.6 once the scale factor exceeded 1.05. CW-SSIM exhibited smoother variations but still fell significantly below

CSSIM when larger scaling factors were applied. In contrast, CSSIM consistently maintained values between 0.86 and 0.95 across the entire range, with a smooth and monotonic decreasing trend, clearly demonstrating its robustness to scale-induced distortions. Further, in the translation experiment, images were shifted along the X- and Y-axes by up to 100 pixels (5-pixel intervals) and 2000 pixels (100-pixel intervals). Two-dimensional distribution maps were generated to visualize the variations in index scores. As shown in Figure 7, SSIM and MS-SSIM quickly decreased to 0.5–0.6 with translations of less than 10 pixels. CW-SSIM maintained slightly higher scores under small translations but exhibited noticeable degradation at larger displacements. By comparison, CSSIM remained highly stable with scores above 0.9 even under large translations, indicating near immunity to alignment errors. Collectively, the experiments demonstrate that CSSIM delivers remarkable robustness and consistency under both scaling and translation distortions, effectively overcoming the sensitivity of traditional SSIM to alignment errors. This establishes CSSIM not only as a reliable objective index for IQA but also as a solid foundation for dataset labeling and UAV image screening in practical applications.

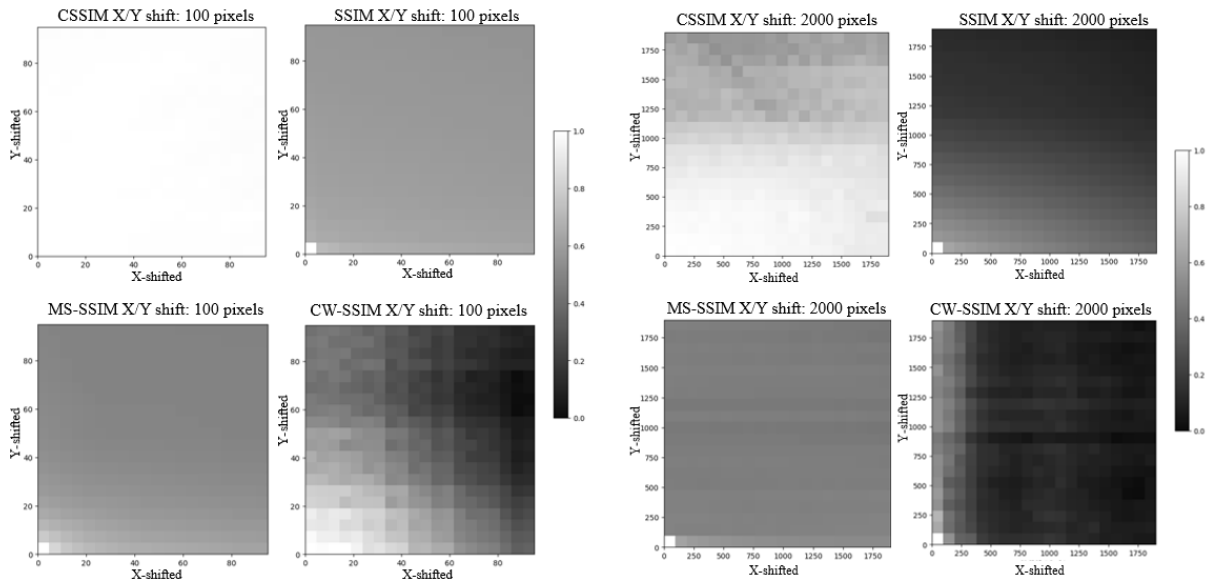


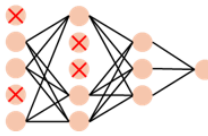
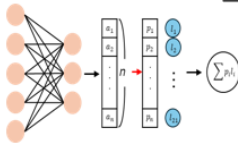
Figure 7. Heatmaps of IQA indexes under X/Y shift. Left: shift step from 0 to 100 pixels; Right: shift step from 0 to 2000 pixels.

## 4.2 Comparative Analysis of Model Architectures and Dual Strategies

To validate the feasibility of the proposed image quality map generation framework, this study first conducted a comparative analysis of different backbone architectures and prediction strategies, aiming to identify the most suitable design. Regarding backbone selection, traditional convolutional neural networks (CNNs) and the ResNet family provide

strong feature extraction capabilities but remain limited in modeling long-range dependencies and global semantic information. With the emergence of self-attention based models, such as the Vision Transformer and its derivatives, superior representational capacity has been demonstrated. Therefore, this study specifically compared convolutional architectures with attention-based architectures to determine which category is more appropriate for image quality map generation. For prediction strategies, two common approaches exist: (i) regression models that directly output a single quality score, and (ii) classification models that assign an input image to the most probable quality interval and use the corresponding value as the score. However, the discrete nature of classification inevitably constrains numerical precision. To address this issue, we propose a classification-based probability-weighted strategy. Specifically, the  $[0,1]$  interval is partitioned into multiple sub-intervals, and each ground-truth score is converted into a Gaussian distribution over these intervals. The model then predicts the probability of each interval, and the final quality score is obtained by the weighted average of interval values. This design preserves the stability of classification while achieving the granularity of regression. Consequently, it is necessary to compare its performance against standard regression methods and to evaluate the influence of the number of intervals.

Table 2: Comparative analysis of different backbone and prediction strategies.

IQM Model		Backbone	Validation RMSE
	Numerical regression model	CNN	0.1648
		ResNet-18	0.1571
		Vision Transformer	0.0751
		<b>Swin Transformer</b>	<b>0.0523</b>
	Probability-weighted classification-11 classes	CNN	0.0855
		ResNet-18	0.0794
		Vision Transformer	0.0444
		<b>Swin Transformer</b>	<b>0.0362</b>
	Probability-weighted classification-21 classes	CNN	0.0704
		ResNet-18	0.0578
		Vision Transformer	0.0337
		<b>Swin Transformer</b>	<b>0.0193</b>
	Probability-weighted classification-41 classes	CNN	0.0695
		ResNet-18	0.0561
		Vision Transformer	0.0314
		<b>Swin Transformer</b>	<b>0.0189</b>

As summarized in Table 2, the regression baseline is consistently outperformed by the proposed probability-weighted approach. Further analysis shows that increasing the number of intervals from 11 to 21 significantly reduces prediction error, whereas extending to 41



intervals yields only marginal improvements while introducing additional computational costs. Thus, 21 intervals are selected as the optimal trade-off. In terms of backbone comparison, the Swin Transformer achieves substantially lower RMSE than CNN, ResNet-18, and Vision Transformer, confirming its advantage in multi-scale feature modeling and global context representation. Based on these results, the Swin Transformer-based 21-interval probability-weighted model was adopted as the primary framework to ensure accurate pixel-level quality prediction.

Nevertheless, the pixel-wise design inherently limits inference speed. To improve efficiency for practical applications, Swin-Unet was introduced as an alternative. By combining an encoder-decoder structure with the representational strength of Transformers, Swin-Unet can directly generate a complete quality map from the entire input image, thereby avoiding the heavy computational burden of pixel-wise inference. Compared with the probability-weighted Swin Transformer model, Swin-Unet exhibits a slight reduction in accuracy but achieves a substantial improvement in inference time (see Table 3), forming a complementary “high-accuracy-high-efficiency” design.

Table 3: Comparison of RMSE and inference time between two models.

IQA model	RMSE	Average inferencing time
Swin-Transformer based probability weight model	0.0019	532.3 s
Swin-Unet model	0.0044	0.3 s

In summary, the probability-weighted model offers high accuracy, making it suitable as a core tool for label generation and theoretical validation, while Swin-Unet provides the efficiency necessary for real-time screening and large-scale UAV image applications. The combination of these two models balances precision and efficiency, demonstrating the practical potential of the proposed framework.

### 4.3 Case Study and Pre-selection Strategy

In this section, we evaluate the proposed AI pre-selection strategy using a UAV image dataset collected from field surveys. The goal is to automatically remove low-quality images from large-scale UAV datasets, thereby improving the efficiency and reliability of inspection and reconstruction.

We first compare two methods for computing overall quality: mean CSSIM and nonlinear weighted CSSIM. As shown in Figure 8, mean CSSIM values cluster in the high-

score region, limiting their ability to distinguish low-quality images and complicating threshold selection. In contrast, nonlinear weighted CSSIM amplifies degraded regions, ensuring severely distorted images receive lower scores even when most pixels appear acceptable. This yields a more balanced distribution and clearer separation between high- and low-quality images, enabling automated threshold determination.

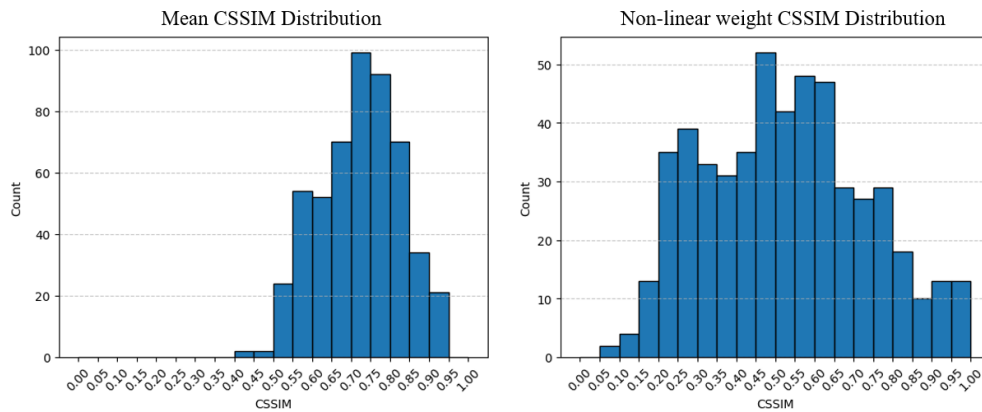


Figure 8. Histogram of Mean CSSIM and Non-linear weight CSSIM distribution.

Next, we present a case study on UAV images to validate the interpretability of CSSIM maps. Figure 9 shows six cases: the top row displays original images, and the bottom row their CSSIM maps with nonlinear weighted scores. The maps highlight degraded regions and align well with observed distortions. For instance, Case 1 suffers from overexposure, reducing target visibility and scoring 0.3985. Case 2, taken under a bridge, shows insufficient brightness and strong shadows, obscuring details (score 0.4646). Case 3 exhibits severe horizontal motion blur, making boundaries indistinct (score 0.3538). In contrast, Cases 4 and 5 remain relatively clear despite minor shadows or backlighting, achieving higher scores of 0.8371 and 0.8934. Based on this dataset, a threshold of 0.4 can exclude severely degraded images, while those between 0.4 and 0.5 (often affected by shadows) may still be retained for enhancement.

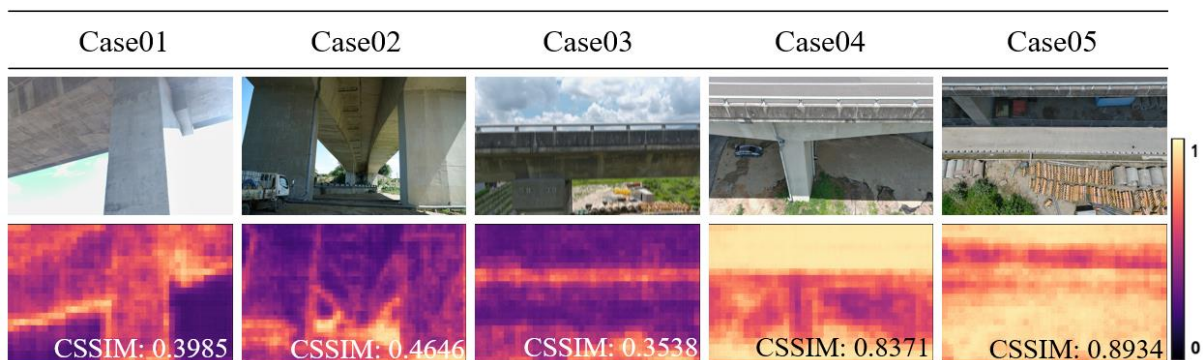


Figure 9. Case Study. Upper: Original image; Lower: IQMs and its CSSIM score.

Overall, both the distribution analysis and the case study confirm the effectiveness of the proposed image screening method. It not only enables automated and interpretable screening of UAV images but also significantly improves the efficiency of large-scale UAV data processing, providing a more reliable foundation for downstream engineering applications such as bridge inspection and 3D reconstruction. Furthermore, the resulting quality-controlled datasets reduce manual triage effort, stabilize training pipelines for downstream models, and help prioritize re-acquisition tasks, when necessary, thereby enhancing operational robustness across long-term monitoring campaigns.

## 5. Conclusion and Recommendation

This study presents an engineering-oriented AI pre-selection strategy for UAV image quality assessment. By introducing the CLIP-based Structural Similarity Index (CSSIM), we overcome the inherent sensitivity of SSIM to scaling and translation while establishing an objective dataset labeling pipeline that no longer relies on subjective human scoring. Through extensive robustness experiments, CSSIM demonstrated superior stability across multiple distortions—including illumination/exposure shifts, sensor noise, complex motion blur, and mild geometric perturbations—validating its reliability as a quality indicator. To operate CSSIM, we developed a dual-model framework. The Swin Transformer based probability weighted model provides accurate pixel-level predictions, making it well-suited for objective label generation and theoretical validation. Complementarily, the Swin-Unet model enables efficient full-image inference, allowing the framework to scale to large UAV datasets and real-time applications. Together, these models form a practical balance between accuracy and efficiency.

Through distribution analysis and case studies, we demonstrated the feasibility of applying nonlinear weighted CSSIM scores for automated screening at scale and in diverse operational conditions across datasets. The results confirmed that severely degraded images can be consistently excluded, while moderately degraded images can be retained for enhancement, ensuring both data quality and quantity for downstream engineering tasks such as bridge inspection and 3D reconstruction. In summary, the proposed AI pre-selection strategy not only advances the objectivization of NR-IQA but also bridges the gap between academic development and practical UAV deployment. Future work will explore integrating the strategy into online UAV inspection workflows, expanding its applicability to diverse infrastructure types, and refining thresholding schemes for adaptive, task-specific quality control.

## References

- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612. <https://doi.org/10.1109/TIP.2003.819861>
- Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multi-scale structural similarity for image quality assessment. In *Proc. 37th Asilomar Conf. on Signals, Systems & Computers* (pp. 1398–1402).
- Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C., & Markey, M. K. (2009). Complex wavelet structural similarity: A new image similarity index. *IEEE Transactions on Image Processing*, 18(11), 2385–2401. <https://doi.org/10.1109/TIP.2009.2025923>
- Zhang, L., Zhang, L., Mou, X., & Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8), 2378–2386. <https://doi.org/10.1109/TIP.2011.2109730>
- Xue, W., Zhang, L., Mou, X., & Bovik, A. C. (2014). Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2), 684–695. <https://doi.org/10.1109/TIP.2013.2293423>
- Li, Q., & Wang, Z. (2009). Reduced-reference image quality assessment using divisive normalization-based image representation. *IEEE Journal of Selected Topics in Signal Processing*, 3(2), 202–211. <https://doi.org/10.1109/JSTSP.2009.2015992>
- Redi, J. A., Gastaldo, P., Heynderickx, I., & Zunino, R. (2010). Color distribution information for the reduced-reference assessment of perceived image quality. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(12), 1757–1769. <https://doi.org/10.1109/TCSVT.2010.2087810>
- Soundararajan, R., & Bovik, A. C. (2012). RRED indices: Reduced reference entropic differencing for image quality assessment. *IEEE Transactions on Image Processing*, 21(2), 517–526. <https://doi.org/10.1109/TIP.2011.2169979>
- Rehman, A., & Wang, Z. (2012). Reduced-reference image quality assessment by structural similarity estimation. *IEEE Transactions on Image Processing*, 21(8), 3378–3389. <https://doi.org/10.1109/TIP.2012.2197011>
- Min, X., Gu, K., Zhai, G., Liu, J., Yang, X., & Zhang, W. (2018). Saliency-induced reduced-reference quality index for natural scene and screen content images. *Signal Processing*, 145, 127–136. <https://doi.org/10.1016/j.sigpro.2017.11.004>
- Moorthy, A. K., & Bovik, A. C. (2010). A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, 17(5), 513–516. <https://doi.org/10.1109/LSP.2010.2044910>
- Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12), 4695–4708. <https://doi.org/10.1109/TIP.2012.2214050>
- Mittal, A., Soundararajan, R., & Bovik, A. C. (2013). Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3), 209–212. <https://doi.org/10.1109/LSP.2012.2227726>
- Ye, P., Kumar, J., Kang, L., & Doermann, D. (2012). Unsupervised feature learning framework for no-reference image quality assessment. In *Proceedings of the IEEE*

- Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1098–1105). IEEE.
- Kang, L., Ye, P., Li, Y., & Doermann, D. (2014). Convolutional neural networks for no-reference image quality assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1733-1740). IEEE.
- Kim, J., & Lee, S. (2017). BIECON: Blind image evaluator with convolutional neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 28(5), 1195-1206.
- Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., Yong, H., Li, Y., & Wang, Z. (2017). End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3), 1202-1213.
- Bosse, S., Maniry, D., Müller, K. R., Wiegand, T., & Samek, W. (2017). Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1), 206–219.
- Zhang, W., Ma, K., Yan, J., Deng, D., & Wang, Z. (2018). RankIQA: Learning from rankings for no-reference image quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).
- Wu, H., Li, L., Liu, H., & Lei, Y. (2020). Pseudo-reference image quality assessment with deep learning. *Signal Processing: Image Communication*, 85, 115875.
- Liu, X., Yang, K., Zhang, L., & Wang, Z. (2022). Pseudo-supervised no-reference image quality assessment via quality-aware pretext tasks. *IEEE Transactions on Image Processing*, 31, 1396-1409.
- You, Q., Yan, X., Yan, Z., & Zhang, L. (2021). Self-supervised learning for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2210-2219).
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 10012–10022).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Neural Information Processing Systems (NeurIPS)* (Vol. 30, pp. 5998–6008).
- Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, X. (2021). Swin-Unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*.