ScoreNet: Netting Lightweight Quality Scores for Better Visual Assessment with Large Multi-Modality Models

Supplementary Material

1. Details on ScoreNet Architecture

Previous works, such as Coop [6] and CoCoOp [5], demonstrated that soft prompting can improve classification accuracy in both zero-shot and few-shot scenarios by eliciting richer information from LMMs. These approaches freeze all model parameters except for the prompt, optimizing it via backpropagation based on image features. ScoreNet is motivated by this idea but introduces a novel strategy by leveraging contextual information from both the input image and lightweight IQA metrics. While these IQA scores alone are not state-of-the-art and may not perfectly correlate with MOS, they provide a valuable signal for refining the prompt, enhancing feature extraction for IQA tasks.

Figure 1 illustrates how integrating IQA scores improves accuracy by ensembling fast, existing metrics. Although insufficient individually, these scores provide conditional information that soft prompting can exploit to extract deeper insights from the foundation model.

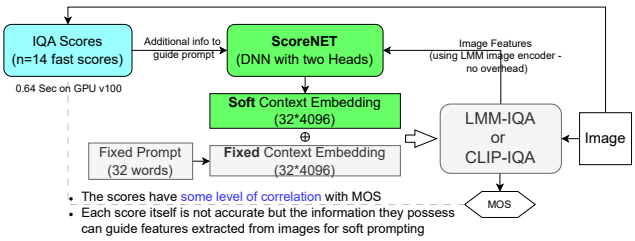

Figure 1. Illustration of ScoreNet's design and the information flow between its Metric Context Generator and Image Context Generator.

Unlike CoCoOp, ScoreNet is specifically designed for IQA rather than zero-/few-shot classification. The CoCoOp paper demonstrated that a fixed trainable prompt embedding struggles to generalize to unseen images. Consequently, we exclude a detailed comparison with CoOp and instead focus on comparing ScoreNet with CoCoOp's soft prompting technique. Both methods conditionally optimize context embeddings using image features. However, ScoreNet introduces three key distinctions:

1. As illustrated in Fig. 2, CoCoOp learns a context embedding matrix of size $(k \times l)$, which serves as input to the LMM's text encoder. In contrast, ScoreNet learns a context embedding $M \in \mathcal{R}^{k \times l}$ and add it to the base embedding that represents the default prompt of the chosen LMM (e.g. CLIP-IQA or Q-Align).
2. ScoreNet incorporates an additional head that extracts perceptual quality context from IQA metric scores, significantly enhancing LMM-IQA performance.
3. Unlike CoCoOp's approach, which generates a vector of length l and adds it uniformly to each row of the context embedding matrix, ScoreNet employs a more flexible and efficient mechanism. For each image, the prompt learner generates a $(k + l)$-dimensional vector, which is then expanded into a $(k \times l)$ conditional embedding matrix and combined element-wise with the base context embedding. This design increases adaptability without incurring substantial computational overhead.

An alternative design choice for the *Context Embedding Generator* was to allow ScoreNet to independently learn all $k \times l$ elements of the matrix M. However, this approach would lead to overfitting, reducing generalizability, and imposing substantial GPU memory overhead during both training and inference. Instead, ScoreNet learns only $k + l$ weights rather than $k \times l$, significantly improving efficiency while maintaining adaptability.

2. DXOMark Chart (DMC) Dataset

To generate the DMC dataset, the DXOMark Chart (DMC) [3] is used under D65 light conditions. For more information about similar scenarios conducted by DXOMark, see [3]. we designed an experiment where a single observer evaluates 100 images to derive Mean Opinion Scores (MOS). Each image is compared in pairs with nine other images, and the number of times an image is selected as "better" determines its MOS, ranging from 0 to 9. If the observer is unable to decide between two images, both receive a score of 0.5. The experiment involves presenting two images at a time, with the option to include a reference image for comparison.

2.1. Image Generation

The DMC images were created using a newly released smartphone camera, under ISO 1600 and normal video capturing conditions employing a readback program that enabled precise control over the Image Signal Processor (ISP) modules responsible for factors such as sharpness, spatial noise, and contrast. Over 100 parameters were systematically adjusted in this perturbation experiment to generate images with varying levels of quality. for our dataset, we utilized middle frame 50 of the captured raw video through smartphone.

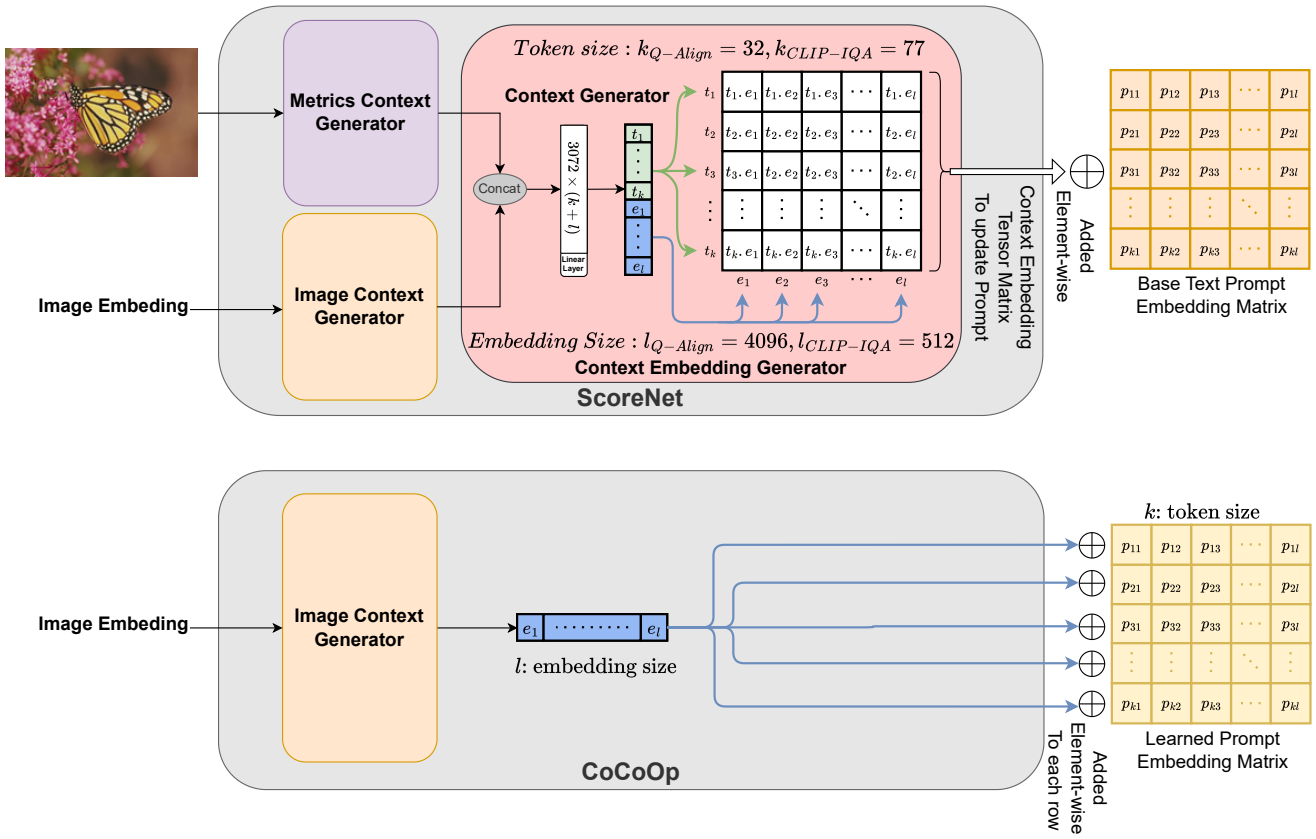

Figure 2. Differences between ScoreNet and CoCoOp.

100 images with different qualities are generated to ensure a semi-uniform distribution of MOS, covering a broad spectrum of visual quality, as illustrated in Fig. 3, which shows sample images with different opinion scores. The aim of this experiment was to create a dataset that can serve as a benchmark for evaluating the effectiveness of IQA metrics in ISP tuning tasks—one of the key applications of IQA scores. By incorporating a wide range of image qualities, this dataset provides a valuable tool for assessing the ability of IQA methods to guide real-world ISP optimization.

2.2. Evaluation Protocol

For each observer, 450 comparisons are made ($100 \times 9/2 = 450$). Observers are instructed to take their time, typically spending up to 10 seconds per comparison. This results in an experimental session lasting up to 75 minutes. The evaluation is repeated for $N = 15$ observers, and the results are averaged to obtain the final MOS for each image.

The dataset generation process is inspired by the TID2013 experiment design [4]. However, in our approach, image pairs are selected randomly in each round, contrasting with the adaptive pairing method used in TID2013, which pairs images based on their scores. Additionally, observers are asked to compare only two images at a time, with the reference image being optional due to the larger size of our images.

Observers follow a step-by-step procedure for each comparison:

1. Open an image pair (e.g., 001_A.png and 001_B.png) using an image viewer tool.
2. Compare the two images and decide which one is of higher quality. The reference image can be included in the comparison if needed.
3. If a decision is made, the image deemed worse is removed from the pair. If no decision is possible, no image is removed.
4. Repeat this process for all image pairs.

Once all comparisons are completed, the scores are calculated based on how many times each image was preferred. The final dataset is obtained after all observers have completed the task, and the results are averaged to produce a comprehensive set of MOS values. Although some image pairs may be compared more than once during the evaluation process, this occurrence is rare and does not significantly affect the overall results. Figure 4 shows the histogram of MOS values for all 100 images in the dataset,

(a) MOS:8.5 (b) MOS:4.4

(c) MOS:1.8 (d) MOS:0

Figure 3. DXOMark Dataset samples.

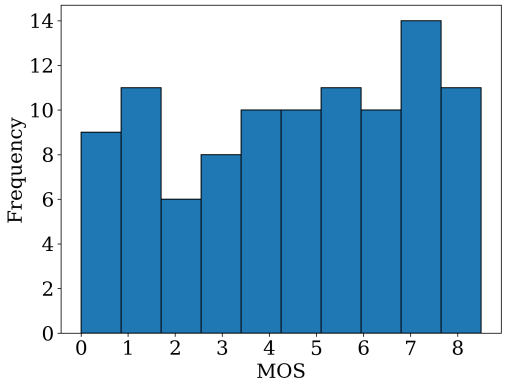

Figure 4. DXOMark Dataset MOS Histogram.

indicating a relatively uniform distribution of MOS values across the sample.

2.3. Key Differences from TID2013

Unlike TID2013, where image pairings are adjusted based on previous comparisons, our methodology simplifies the pairing process by using random image pairs in every round. Additionally, while TID2013 suggests a 2-3 second evaluation time per comparison, we allow observers more time due to the larger image size and the increased complexity of the task as sometimes the difference between the image pairs is subtle and locally induced but the Camera ISP pipeline.

3. End-to-End ISP Tuning Using ScoreNet

To demonstrate a key application of a robust image quality assessment (IQA) metric, we designed a synthetic Image Signal Processing (ISP) pipeline, illustrated in Fig. 5. This synthetic ISP serves as a reverse approximation of an end-to-end real-world digital camera ISP tuning process, allowing controlled distortion simulations for research purposes. Then we define an ISP pipeline Tuning scenario using ScoreNet-Align and Q-Align as objectives and compare their end-to-end performance.

3.1. Synthetic ISP Pipeline

A real ISP pipeline processes raw images/frames from the camera sensor, applying image adjustments through hardware and AI-driven modules. Our synthetic ISP pipeline, illustrated in Fig. 5, uses a reverse-engineering approach. Starting with a pristine, high-quality 4K image as input, it simulates the impact of the ISP pipeline parameters on the overall quality of the image. The synthetic ISP is driven by a set of parameters $X = [x_1, x_2, x_3, x_4, x_5]$, each representing specific processing settings. These parameters are internally mapped to distortion levels through a nonlinear transformation, as depicted in Fig. 5. This nonlinearity

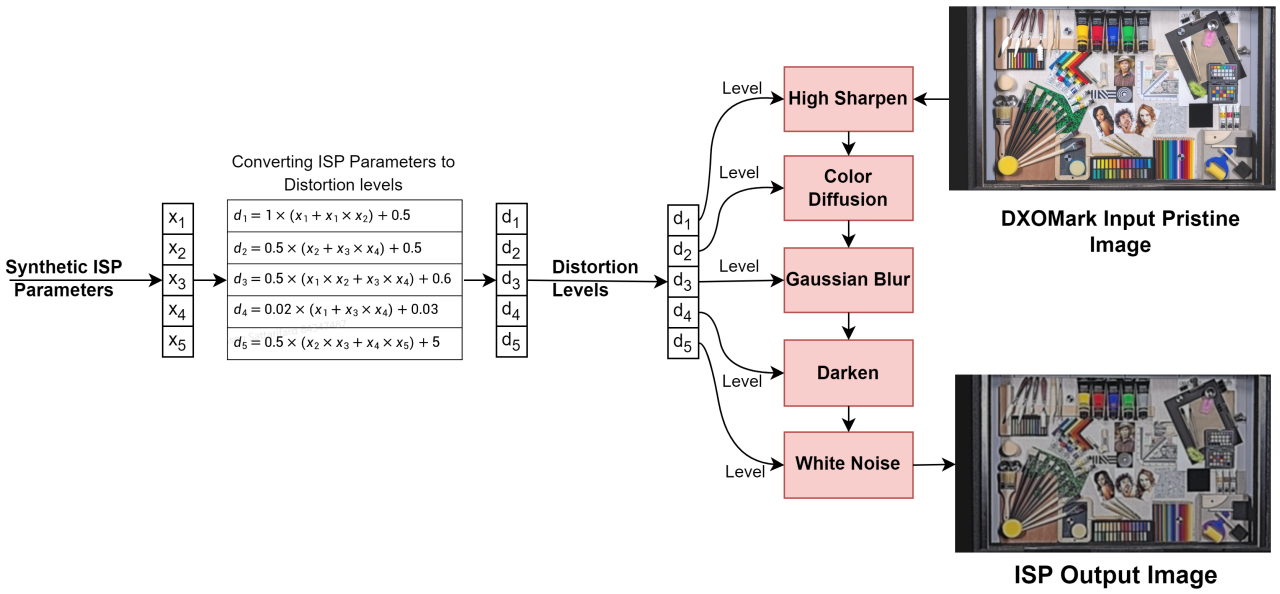

Figure 5. Synthetic Image Signal Processing Pipeline: This pipeline takes input parameters X to distort a pristine image. The objective is to identify the optimal parameter set X_o that minimizes distortion levels, preserving the original image quality. The nonlinear relationship between X and distortion levels d_i is unknown to the ISP optimizer and can vary depending on the specific study.

mimics the complex adjustments made by real ISP systems while enabling flexible, parameterized control over image quality. The synthetic ISP applies five distinct distortion functions to the pristine input image in a sequential manner, replicating a staged processing workflow similar to a reverse real ISP. Each stage introduces specific artifacts and degradations, designed to simulate commonly encountered distortions in digital imaging systems. For each type of distortion, we leverage functional modules inspired by those described in [1], where robust distortion models are proposed for accurate IQA performance. The sequence and type of distortions implemented are deliberately designed to challenge and validate IQA metrics. By controlling the distortion parameters, we can systematically vary the level and type of degradation in output images, which allows for in-depth testing of IQA algorithms under a range of conditions.

To promote reproducibility and further research in IQA, we will make the source code for this synthetic ISP pipeline publicly available.

3.2. ISP Auto-Tuning Using IQA Metrics

With the synthetic ISP pipeline established, we now explore a practical industrial application of reliable IQA metrics in optimizing ISP performance. Specifically, we aim to evaluate the effectiveness of ScoreNet-Align and Q-Align in guiding ISP parameter tuning to achieve high-quality output with minimal distortions.

Objectives	d_i: Distortion levels					l_1 Loss
	d_1	d_2	d_3	d_4	d_5	
Q-Align	0.578	0.644	0.548	0.023	5.74	1.547
ScoreNet-Align	0.571	0.518	0.511	0.021	5.150	1.354

Table 1. Optimization Results for Q-Align and ScoreNet-Align Objectives on Synthetic ISP: Each tuning scenario was executed 10 times, with the best-performing instance shown.

For this purpose, we employ the Evolutionary Strategy (ES) algorithm [2], a well-known single-objective genetic optimization method, to tune the ISP parameters $X = [x_1, x_2, x_3, x_4, x_5]$. We conduct the optimization under two scenarios: in the first, we use ScoreNet-Align as the objective loss function to be maximized, while in the second, we use Q-Align. By setting the IQA metric as the objective, we allow ES to identify ISP parameter configurations X that maximize image quality according to each metric.

The goal of this experiment is to assess which IQA metric can better drive the ISP tuning process, resulting in parameter sets that yield lower distortion levels, denoted as D. After running each optimization scenario for a fixed search budget of 50 population and a population size of 30, we evaluate the resulting distortion levels by computing the l_1 loss of the distortion parameters D obtained from the optimized X. This end-to-end l_1 loss serves as an indicator of the overall quality achieved through each IQA metric's guidance. An accurate IQA metric should effectively steer

the genetic search process, enabling the discovery of ISP parameter sets that preserve image quality by minimizing distortions. A lower L_1 loss in the D parameters signifies a more precise IQA metric, as it demonstrates a more effective optimization of the ISP settings. The results of this comparison are presented in Table 1 highlighting the performance differences between ScoreNet-Align and Q-Align in this auto-tuning context.

References

[1] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 189–198, 2024.

[2] Hans-Georg Beyer and Hans-Paul Schwefel. Evolution strategies–a comprehensive introduction. *Natural computing*, 1:3–52, 2002.

[3] DXOMark. Dxomark mobile scores for smartphone cameras. [Online]. Available: `https://www.dxomark.com/dxomark-mobile-scores-smartphone-cameras/`, 2024. Accessed: 2024-11-13.

[4] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015.

[5] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.