# PSIM: Perceptual Similarity Index Measure

Supplementary Material

## 7. Implementation Details

### 7.1. Training and Testing Details

We have used VGG-16 pretrained on ImageNet as the feature backbone for our proposed model. No data augmentation is applied during training. Following common practice, we use the AdamW optimizer with a weight decay of $1 \times 10^{-5}$ and an initial learning rate of $3 \times 10^{-5}$. A cosine annealing learning-rate schedule is applied with $T_{\max} = 20$, $\eta_{\min} = 0$, and $\eta_{\max} = \text{lr}$, where lr denotes the initial learning rate of the optimizer. All models are trained for 100 epochs, and early stopping based on validation performance is used to reduce training time.

Following standard practice, we use $384 \times 384$ training patches for KADID-10k and BAPPS, and $224 \times 224$ patches for PieAPP and PIPAL. All patches are obtained by random cropping during training. All experiments are implemented in PyTorch and trained on an NVIDIA RTX 6000 GPU. Hyperparameters and optimizer settings remain consistent across datasets, and we use the official training, validation, and testing splits provided for each benchmark.

For testing, we use the full-resolution image as a single input and do not apply multi-patch or multi-crop testing.

### 7.2. Loss Functions

Different datasets provide different forms of human judgments and therefore require different loss formulations. We describe the losses used for each label type below.

#### 7.2.1. MOS-labeled datasets

For datasets that provide mean opinion scores (MOS), we first normalize all MOS values to the range $[0, 1]$. The model is then trained using the standard mean squared error (MSE) loss between the predicted score, $\hat{y}$ and the normalized ground-truth MOS, $y$.

#### 7.2.2. 2AFC-labeled datasets

Datasets such as PieAPP and BAPPS provide pairwise preference judgments obtained through a two-alternative forced-choice (2AFC) protocol. Each training sample consists of a reference image and two distorted versions $(I_r, I_A, I_B)$, along with the human preference probability $p_{AB}$ indicating how often $I_A$ is preferred over $I_B$.

Following prior work [41], we first compute the perceptual error for each distorted image independently:

$$\hat{y}_A = \text{PSIM}(I_r, I_A) \tag{15}$$

$$\hat{y}_B = \text{PSIM}(I_r, I_B) \tag{16}$$

Then we convert those scores into a preference probability using the Bradley–Terry (BT) model:

$$\hat{p}_{AB} = \frac{1}{1 + \exp\left(\hat{y}_A - \hat{y}_B\right)} \tag{17}$$

The training objective minimizes the MSE between the predicted and ground-truth preference probabilities:

$$\mathcal{L}_{\text{2AFC}} = \frac{1}{N} \sum_{i=1}^{N} \left( \hat{p}_{AB}^{(i)} - p_{AB}^{(i)} \right)^2 \tag{18}$$

where $N$ is the number of samples.

During evaluation, consistency between model predictions and human judgments is measured using the standard 2AFC accuracy metric [61]. Given predicted scores $(\hat{y}_A, \hat{y}_B)$ and human preference scores $(p_A, p_B)$, the 2AFC accuracy is computed as:

$$\begin{aligned} \text{2AFC accuracy} = {} & \mathbb{1}(\hat{y}_A < \hat{y}_B)\, \mathbb{1}(p_A < p_B) \\ & + \mathbb{1}(\hat{y}_A > \hat{y}_B)\, \mathbb{1}(p_A > p_B) \\ & + 0.5\, \mathbb{1}(\hat{y}_A = \hat{y}_B) \end{aligned} \tag{19}$$

## 8. Visualization

To illustrate how PSIM organizes distortions across perceptual scales, we conduct a visual analysis using the TID2013 dataset. The model is trained on KADID-10k and then applied to all distorted versions of each reference image. For a given reference, we compute the Multi-level Wasserstein Distortion scores for all associated distortions and sort them according to their $(L2{+}L3)$ values, which correspond to coarser perceptual scales with larger receptive fields. Distorted images with high $(L2{+}L3)$ scores typically contain degradations that "pop out" immediately—artifacts that are noticeable even without deliberate fixation. These examples represent the Most Apparent Distortion (MAD) cases.

Conversely, we also rank images according to their $(L0{+}L1)$ scores and select those with the lowest values. These distorted images often appear almost indistinguishable from the reference at first glance; their deviations become noticeable only after deliberate foveal inspection. Such examples constitute the Least Apparent Distortion (LAD) cases, capturing distortions that require focused attention to perceive.

Figure 5 presents representative examples from this procedure. Our model reliably identifies distorted images with large, salient degradations as MAD, while subtle, fine-grained alterations are categorized as LAD. This illustrates how our proposed model leverages multi-scale information to judge image quality.

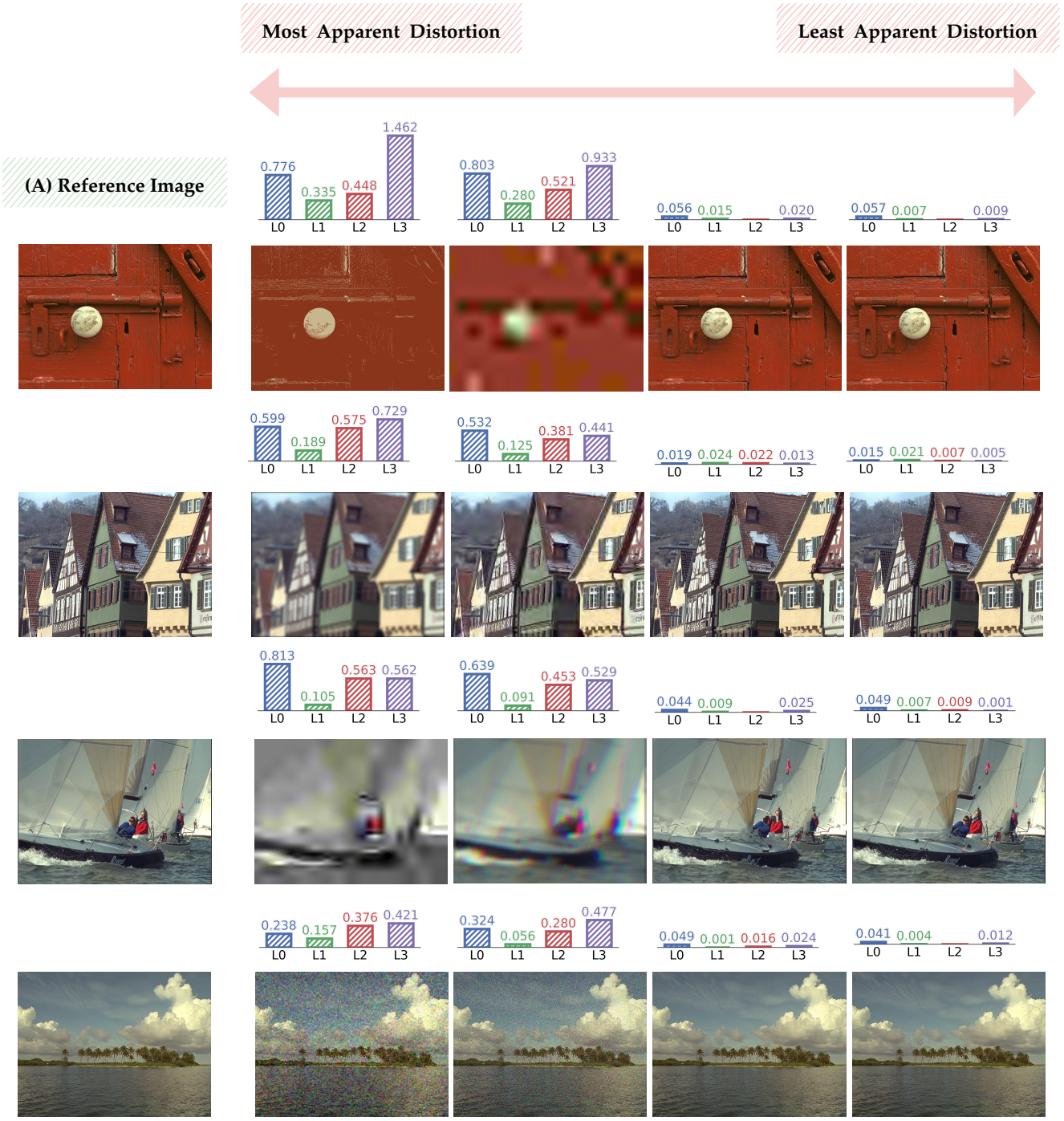


Figure 5. Multi-level Wasserstein Distortion (MWD) breakdown for an original image and its distorted versions from the TID2013 dataset. Bars L0–L3 represent distortion measured at progressively larger perceptual scales: L0–L1 capture fine, near-threshold (LAD) deviations, while L2–L3 capture coarse, most-apparent (MAD) distortions. Bar height indicates each level's contribution to the overall PSIM perceptual distance. For each row, the first column shows the reference image. The next two columns contain the distorted images with the highest MAD score (largest $L2{+}L3$). The last two columns contain the distorted images with the lowest LAD score (smallest $L0{+}L1$). Images are best viewed on a display with a luminance range of 5–300 cd/m$^2$ and a $\gamma$ exponent of 2.4.