PSIM: Perceptual Similarity Index Measure

Md Eimran Hossain Eimon
Florida Atlantic University
meimon2021@fau.edu

Hari Kalva
Florida Atlantic University
hkalva@fau.edu

Abstract

*Human perception integrates information across multiple spatial scales, rapidly detecting coarse, appearance-level distortions while relying on high-acuity mechanisms to identify near-threshold deviations. Existing full-reference image quality assessment (FR-IQA) models often fail to capture this balance: some emphasize global structure at the expense of fine detail, while others excel at local fidelity but overlook global perceptual changes. To address this gap, we introduce **P**erceptual **S**imilarity **I**ndex **M**easure (PSIM), a perceptually motivated FR-IQA model that utilizes an advanced Multi-level Wasserstein Distortion (MWD) model. MWD evaluates image distortions across a hierarchy of spatial supports, enabling the model to capture both Most Apparent Distortions (MAD), which are large-scale and visually dominant changes that attract immediate attention, and Least Apparent Distortions (LAD), which are subtle, fine-grained deviations that become noticeable only under closer inspection. By computing Wasserstein distortions over progressively larger receptive fields, PSIM provides a unified representation of both coarse and fine perceptual regimes. Comprehensive evaluations across multiple public IQA datasets demonstrate that PSIM outperforms most existing state-of-the-art metrics while requiring significantly fewer FLOPs. Moreover, it generalizes strongly in cross-dataset evaluations and exhibits robustness to small spatial shifts. Further analysis demonstrates that PSIM achieves competitive performance in assessing perceptual color differences.*

1. Motivation

Human judgments of image quality arise from the hierarchical and multi-scale organization of the human visual system (HVS). The structure of the HVS is shaped by evolutionary pressures rather than deliberate design. The nonuniform distribution of rods and cones across the retina reflects adaptation to survival-critical demands. Peripheral rods support rapid detection of global motion, contrast changes, and approaching threats, while densely packed foveal cones

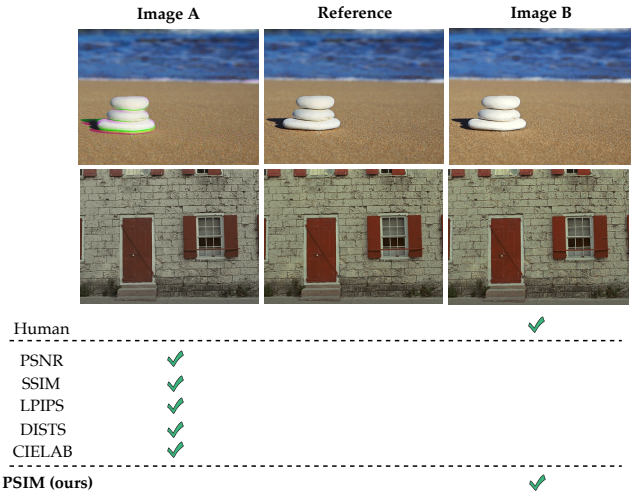

Figure 1. *Which image is closer to the reference in terms of perceived quality and color appearance?* **Top row:** Image A is distorted by chromatic aberration, whereas Image B is distorted by color quantization. **Bottom row:** Image A is distorted by color saturation change, while Image B is degraded by multiplicative Gaussian noise. Images are best viewed on a display with a luminance range of 5–300 cd/m^2 and a γ exponent of 2.4.

enable fine-scale analysis required for tool manipulation, reading, and discerning subtle textures. This biological architecture produces complementary perceptual sensitivities across spatial scales. Large distortions that alter global appearance, such as strong blur, broad color shifts, or prominent structural changes, tend to be detected early because they strongly activate coarse spatial-frequency channels and saliency mechanisms. Prior work in saliency modeling shows that global contrast, color, and orientation differences elicit fast pre-attentive responses [22, 23]. Studies on spatial-frequency processing demonstrate that low-frequency information is processed earlier and has a dominant role in initial perception [21]. Research on scene gist further confirms that global structure and appearance are extracted within a very short time window [8, 13, 37]. These findings support the interpretation of such distortions as Most Apparent Distortions (MAD), which influence the initial impression of quality before fine details are observed.

As distortion magnitude decreases and approaches the near-threshold regime, perception relies more heavily on foveal, high-acuity channels that specialize in resolving fine spatial detail. Subtle texture inconsistencies, faint ringing artifacts, and small structural deviations become noticeable only with closer inspection, corresponding to Least Apparent Distortions (LAD). Psychophysical studies of contrast and color sensitivity support this distinction between coarse early responses and fine later detection [36, 51].

Importantly, perception does not proceed in a rigid sequence from peripheral analysis to foveal scrutiny. Instead, the HVS performs parallel, coarse-to-fine processing across spatial-frequency channels with different receptive-field sizes and contrast sensitivities. Phenomena such as spatial-frequency sensitivity, near-threshold detection behavior, and the varying salience of distortions across scales [21, 38] support the view that image-quality judgments reflect an integration of signals across different pathways.

Despite this, many full-reference IQA metrics treat distortions uniformly and do not distinguish between coarse global appearance changes and fine local deviations. Metrics focused on global structure may underweight subtle distortions, while local fidelity measures may miss large-scale changes. Color-difference models often analyze chroma independently, even though perceptual appearance depends on the joint encoding of luminance, chroma, and structure.

Figure 1 highlights the dynamic relationship between LAD and MAD in shaping perceived image quality. In the top example, the chromatic aberration in the left image is immediately salient in peripheral vision, making the right image the clear perceptual match. In the bottom example, the saturation shift in the left image is obvious, but the multiplicative Gaussian noise in the right image remains effectively invisible without foveal scrutiny. Traditional metrics such as PSNR and SSIM [18, 56], as well as learned deep-feature metrics such as LPIPS [61] and DISTS [6], do not explicitly account for these distinct perceptual regimes and therefore struggle when global and local cues conflict. We also tested CIELAB [34], a perceptually uniform color space widely used for color-difference assessment. Despite its perceptual uniformity, it too fails to match human preferences in these challenging examples.

These observations motivate a representation that explicitly models perceptual distortions across multiple spatial scales. To address this need, we propose the Multi-level Wasserstein Distortion (MWD) framework. MWD measures discrepancies over progressively larger spatial supports, mirroring the hierarchical organization of the HVS and capturing distortions associated with both MAD and LAD. This multi-level formulation enables PSIM to account for global appearance, fine detail, and perceptual color differences within a unified and biologically inspired structure.

Our contributions are summarized as follows:

- We introduce a bio-inspired Multi-level Wasserstein Distortion (MWD) framework, based on optimal transport theory, to measure distortions across multiple spatial scales.
- We provide a formal characterization of Most Apparent Distortions (MAD) and Least Apparent Distortions (LAD), modeling how coarse appearance cues and fine near-threshold deviations jointly shape image quality judgments.
- Our proposed model demonstrates strong alignment with human perceptual judgments while requiring significantly reduced computational cost.
- We benchmark our proposed model across multiple public datasets and show that it outperforms most existing methods, generalizes effectively across datasets, remains robust to small spatial shifts, and provides competitive assessments of perceptual color differences.

2. Related Work

For decades, attempts to replicate early stages of human visual processing have shaped IQA research, and closer approximations of these mechanisms have consistently led to improved performance. FSIM [59], grounded in psychophysical evidence [9, 35], and CW-SSIM [45], which relies on complex wavelet representations that capture cortical response properties supported by neurological experiments [5], both illustrate how biologically inspired models can approximate aspects of human perception. Similarly, we draw inspiration from psychophysical findings showing that the HVS relies on multiple strategies when evaluating image quality. The Most Apparent Distortion (MAD) model [30] was the first to formalize this behavior by classifying distortions into apparent and near-invisible categories using a multiscale pyramid and Log-Gabor analysis. While this multi-strategy perspective motivates our approach, MAD depends on a handcrafted distortion model. In contrast, our proposed PSIM adopts a more advanced distortion formulation that jointly captures foveal and peripheral distortions while leveraging the "unreasonable effectiveness" of deep feature representations [61].

The idea of comparing images through Wasserstein distances—or proxies computed between feature-space distributions—is not new [17, 20, 39, 44, 48, 49]. Recent work has applied Wasserstein distance across a wide range of vision tasks, including image retrieval, texture synthesis, generative modeling, and style transfer, largely due to its ability to capture structural and statistical differences more faithfully than Euclidean or KL-based measures [50]. Prior work also shows that Wasserstein distortion offers a unified framework that smoothly balances pixel-level fidelity and perceptual realism [1, 42]. A recent IQA model [31] employs one-dimensional Wasserstein distances between deep

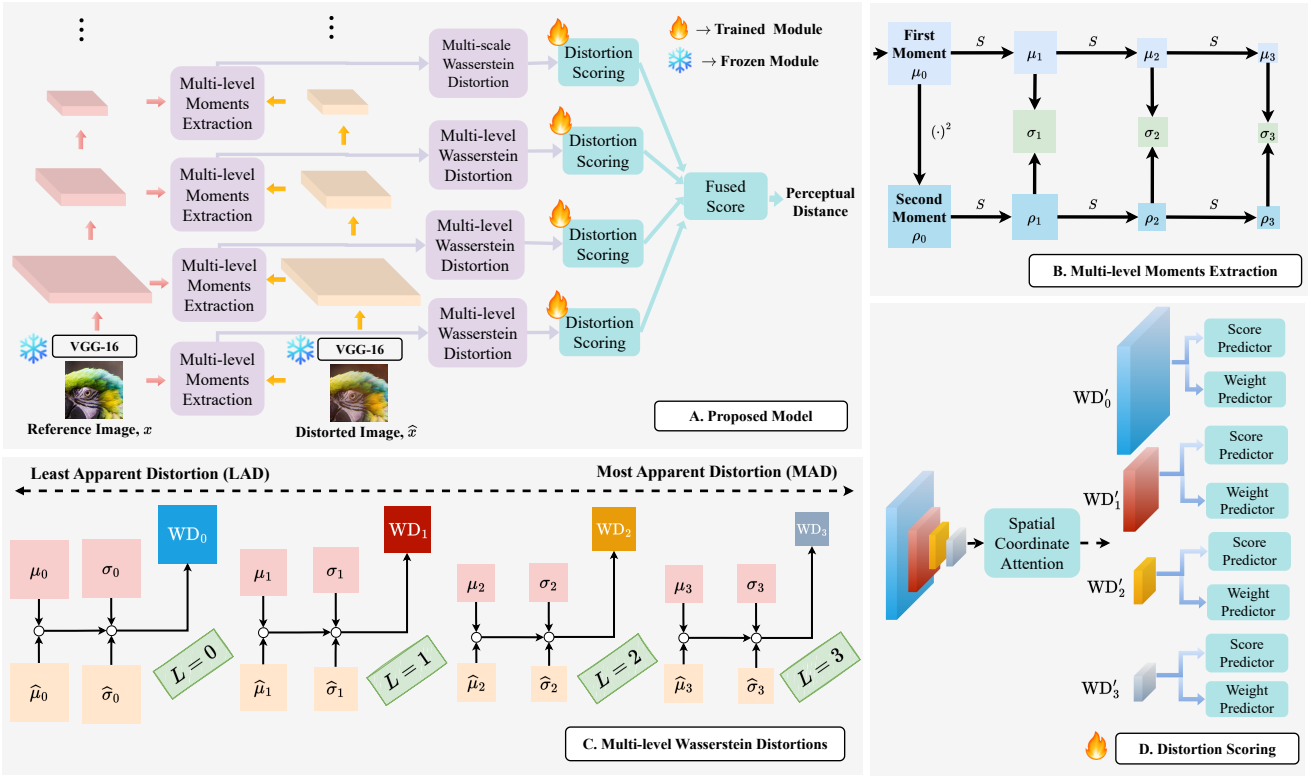

Figure 2. Overview of the proposed Perceptual Similarity Index Measure (PSIM). **A.** Proposed Model **B.** Multi-level Moments Extraction, **C.** Multi-level Wasserstein Distortions that capture *Least Apparent Distortion (LAD)* to *Most Apparent Distortion (MAD)*, and **D.** Distortion Scoring. VGG-16 [47] is used as the deep feature extractor, and only the *Distortion Scoring* module is trained.

feature distributions. However, such one-dimensional comparisons fail to capture spatially structured distortions, since perceptually distinct artifacts can produce similar marginal histograms despite having substantially different spatial organization and visual impact.

Moreover, inspired by the fovea–periphery organization of the HVS, many compression frameworks adopt an economical strategy that allocates more bits to high-fidelity reconstruction near predicted fixation regions, while encoding peripheral content more coarsely [52–54]. In some cases, peripheral regions could even be replaced at the decoder with perceptually similar synthesized textures to preserve subjective quality at low bitrates, emphasizing perceptual realism over exact pixel fidelity. This trade-off between realism and fidelity aligns with the properties of Wasserstein distance, which can capture both structural consistency and perceptual plausibility within a unified measure. In addition, sliced Wasserstein distance [43], which projects high-dimensional data onto multiple 1D slices, has shown strong promise in perceptual color-difference [15], suggesting that Wasserstein distance can effectively capture color and appearance variations that traditional metrics often overlook. Collectively, these findings inspired us to adopt the Wasserstein distortion for perceptual modeling.

3. Proposed Method

We present the **Perceptual Similarity Index Measure (PSIM)**, as shown in Figure 2, a full-reference image quality assessment model that quantifies perceptual distance between a reference image x and its distorted version $\hat{x}$. PSIM mimics the well-established property of the human visual system (HVS): receptive fields in the ventral stream grow with eccentricity [10].

PSIM begins by extracting a hierarchy of deep feature representations. Formally, for an image pair $(x, \hat{x})$, where x denotes the reference image and $\hat{x}$ its distorted version, a deep encoder $\Phi(\cdot)$ extracts multi-level feature activations:

$$\mathcal{F}(x) = \{\, f_i = \Phi_i(x) \,\}_{i=0}^{M} \tag{1}$$

$$\mathcal{F}(\hat{x}) = \{\, \hat{f}_i = \Phi_i(\hat{x}) \,\}_{i=0}^{M} \tag{2}$$

where Φ_i denotes the i-th stage of the encoder $\Phi(\cdot)$. Each feature pair $(f_i, \hat{f}_i)$ is then passed to the *Multi-level Moments Extraction* stage to compute local statistical summaries across scales. For brevity, we omit the index when unambiguous and write $f_i, \hat{f}_i$ simply as $f, \hat{f}$.

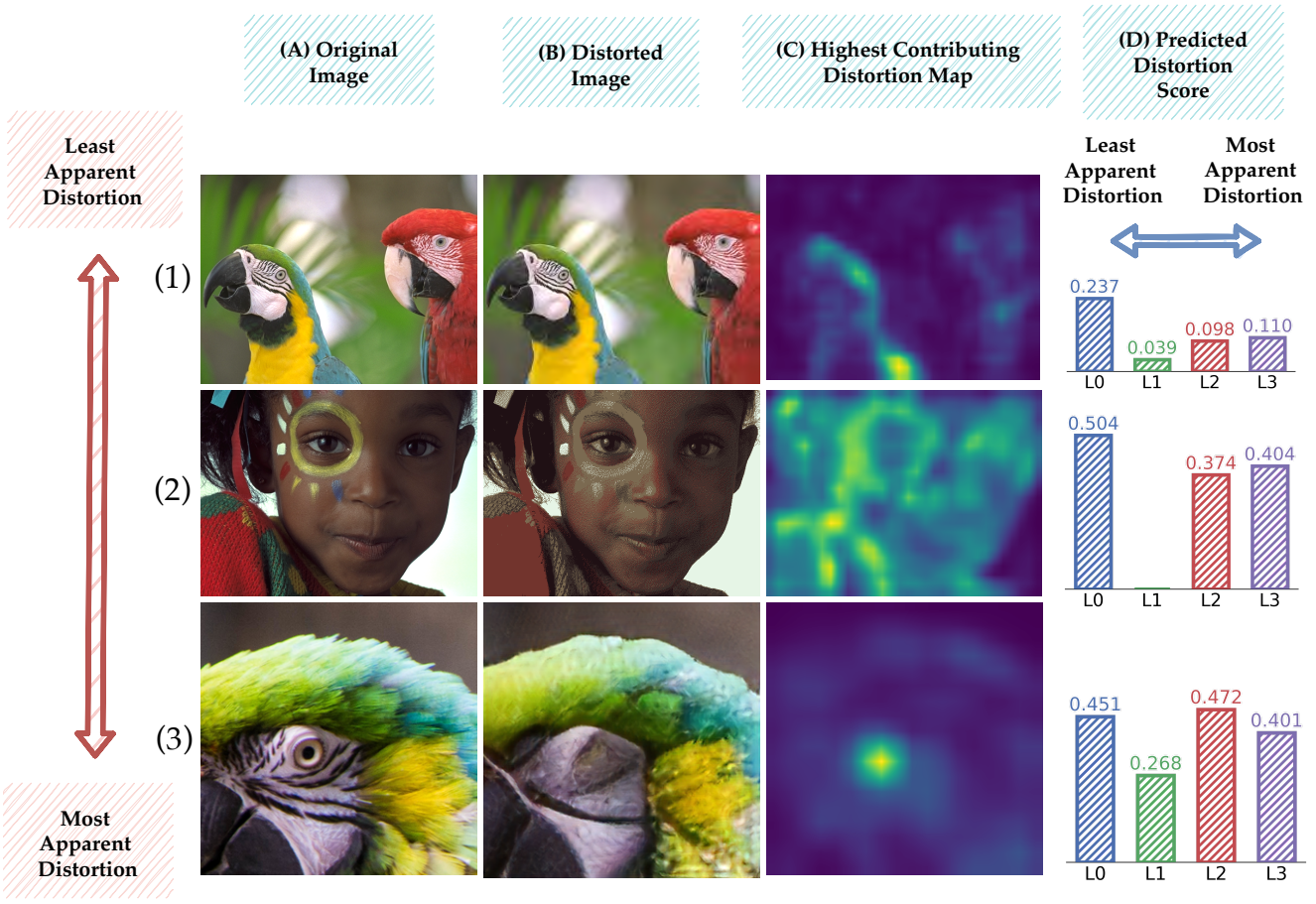

Figure 3. Visualization of the distortion maps. Rows (1)–(3) illustrate a "perceptual progression" from *Least Apparent Distortion* (LAD) to *Most Apparent Distortion* (MAD). Columns: (A) original image, (B) distorted image, (C) highest-contributing distortion map, and (D) predicted distortion scores across Multi-level Wasserstein Distortion (MWD) levels. The bar plot in (D) illustrates how the Human Visual System (HVS) may integrate different cues with varying receptive fields when judging image quality—where L0 captures foveal, fine-grained distortions requiring fixation (a more effortful process), and L3 reflects coarse, globally salient degradations (things that simply "pop out").

3.1. Multi-Level Moments Extraction

To emulate how receptive fields in the ventral visual stream expand with eccentricity, we construct a hierarchy of locally aggregated feature statistics whose spatial support grows with level L. We have used four levels of spatial support in our proposed model. For each feature map f (and similarly $\hat{f}$), we apply a low-pass and subsampling operator $\mathcal{S}_L$ to estimate local first and second raw moments:

$$\mu_L = \mathcal{S}_L(f), \quad \rho_L = \mathcal{S}_L(f^2) \tag{3}$$

from which the local variance field follows as:

$$\sigma_L^2 = \rho_L - \mu_L^2 \tag{4}$$

The initialization $\mu_0 = f$ and $\sigma_0^2 = 0$ anchors the finest, foveal level, while deeper levels ($L{>}0$) progressively capture coarser, peripheral statistics. The same procedure applied to $\hat{f}$ yields $(\hat{\mu}_L, \hat{\sigma}_L^2)$.

3.2. Multi-Level Wasserstein Distortion

At each level L, we measure the perceptual discrepancy between corresponding local statistics of the reference and distorted features using the 2-Wasserstein distance. Assuming local Gaussianity, the squared 2-Wasserstein distance between two multivariate Gaussians $\mathcal{N}(\mu, C)$ and $\mathcal{N}(\hat{\mu}, \hat{C})$ is:

$$\|\mu - \hat{\mu}\|_2^2 + \mathrm{Tr}\big(C + \hat{C} - 2(C^{1/2}\hat{C}C^{1/2})^{1/2}\big) \tag{5}$$

Assuming channel-wise independence, the covariances C and $\hat{C}$ are diagonal, and (5) simplifies to:

$$WD_L = (\mu_L - \hat{\mu}_L)^2 + (\sigma_L - \hat{\sigma}_L)^2 \tag{6}$$

Each WD_L map thus measures the perceptual distance between features at level L.

3.3. Distortion Scoring

The distortion maps WD_L are subsequently modulated by learned Spatial Coordinate Attention (SCA) [19] module $\mathcal{A}_L(\cdot)$ that emphasizes perceptually salient distortion regions:

$$WD'_L = \mathcal{A}_L(WD_L) \tag{7}$$

Each attended map WD'_L is globally pooled to a descriptor $\mathbf{z}_L$, summarizing the level's overall distortion. Two parallel MLPs then predict the score q_L and weight w_L:

$$q_L = \psi_q(\mathbf{z}_L) \tag{8}$$

$$w_L = \mathrm{ReLU}\big(\psi_w(\mathbf{z}_L)\big) \tag{9}$$

3.4. Fused Perceptual Distance

Finally, the overall perceptual distance is computed as a normalized, weighted aggregation of the level-wise distortion scores:

$$\mathrm{PSIM}(x, \hat{x}) = \frac{\sum_L w_L\, q_L}{\sum_L w_L} \tag{10}$$

Equation (10) formalizes perceptual cue integration: foveal levels (L=0) account for fine structural fidelity, whereas peripheral levels (L=3) capture globally apparent distortions. Examples of distortion maps generated by our model are shown in Figure 3.

4. Experiments

4.1. Evaluated Datasets

We evaluate the proposed PSIM model on multiple public full-reference IQA benchmarks, summarized in Table 1. The general quality datasets include CSIQ [30], TID2013 [40], KADID-10k [32], BAPPS [61], PieAPP [41], and PIPAL [14, 24]. Since human perception remains largely invariant under small spatial shifts, we further evaluate PSIM on BAPPS-ST [12], a spatially shifted variant of the BAPPS dataset designed to test shift tolerance. Moreover, we argue that an effective IQA model should also capture perceptual color differences well, as the human visual system is highly sensitive to color consistency. To this end, we evaluate PSIM on two datasets: TID2013-Color, a subset of the original TID2013 traditionally used to test color difference generalization [4, 57], which includes color distortions such as quantization noise, color dithering, and chromatic aberration; and the Smartphone Photography Color Difference (SPCD) [57] dataset, a large-scale benchmark for smartphone color difference evaluation reflecting device-dependent color variations across vendors. Unless otherwise stated, PSIM is trained on the KADID-10k dataset.

4.2. Performance Evaluation Metrics

PLCC and SRCC. PLCC quantifies the linear correlation between predicted scores $\hat{y}$ and subjective mean opinion scores (MOS) y, while SRCC evaluates the rank correlation. Following standard practice [6], we apply a four-parameter logistic mapping before computing PLCC:

$$\hat{y}' = \frac{\beta_1 - \beta_2}{1 + \exp(-(\hat{y} - \beta_3)/|\beta_4|)} + \beta_2 \tag{11}$$

where $\{\beta_i \mid i = 1, 2, 3, 4\}$ are fitted with least-square losses between $\hat{y}'$ and ground-truth labels y, and are initialized with $\beta_1 = \max(y)$, $\beta_2 = \min(y)$, $\beta_3 = \mu(\hat{y})$, $\beta_4 = \sigma(\hat{y})/4$. Here, $\sigma(\cdot)$ is the standard deviation.

STRESS. The standardized residual sum of squares (STRESS) [11, 25] measures the agreement between predicted and ground-truth perceptual differences, and is defined as:

$$\mathrm{STRESS} = 100\sqrt{\frac{\sum_{i=1}^{I}(\Delta E_i - F\,\Delta V_i)^2}{F^2\sum_{i=1}^{I}(\Delta V_i)^2}} \tag{12}$$

where I is the number of test pairs, and F is the scale correction factor between the predicted color differences ΔE and the ground-truth color differences ΔV:

$$F = \frac{\sum_{i=1}^{I}(\Delta E_i)^2}{\sum_{i=1}^{I}\Delta E_i \Delta V_i} \tag{13}$$

Lower STRESS indicates a tighter fit between predicted and subjective judgments, typically used in evaluation of color metrics.

Rank-Flip Rate (r_{rf}). To quantify spatial robustness, we compute the rank-flip rate, following [12], which counts how often the relative ranking between two distorted images changes after a small spatial shift:

$$r_{\mathrm{rf}} = \frac{1}{N}\sum_{l=1}^{N}\left(s_1^l < s_2^l\right) \neq \left(\hat{s}_1^l < \hat{s}_2^l\right) \tag{14}$$

where N is the number of samples, s_1^l and s_2^l denote the original scores for the two distorted images in the l-th sample, and $\hat{s}_1^l$ and $\hat{s}_2^l$ denote the corresponding scores after a small shift. A lower r_{rf} indicates higher robustness to spatial misalignment.

Table 1. Datasets used in this study.

Test	Dataset	#Ref	#Dist	#Ratings
General	CSIQ	30	866	5k
	TID2013	25	3,000	524k
	KADID-10k	81	10.1k	304k
	BAPPS	–	187.7k	484k
	PieAPP	200	20k	2.3M
	PIPAL	250	29k	1.13M
Shift-Tolerance	BAPPS-ST	–	187.7k × 3	484k × 3
Color-Difference	TID2013-Color	25	375	87.5k
	SPCD	701	10k	200k

Table 2. Performance comparison on the BAPPS validation set in terms of 2AFC accuracy (range [0, 1]), where higher values indicate stronger perceptual alignment with human judgments. The best and second-best results are highlighted in orange and blue, respectively. All results are obtained using the authors' official implementations. PSIM achieves the highest overall alignment, capturing both synthetic and real-world distortions.

Method	Synthetic distortions			Distortions by real-world algorithms					All ↑
	Traditional	CNN-based	All	Super resolution	Video deblurring	Colorization	Frame interpolation	All	
Human	0.808	0.844	0.826	0.734	0.671	0.688	0.686	0.695	0.739
PSNR	0.573	0.801	0.687	0.642	0.590	0.624	0.543	0.600	0.629
SSIM [56]	0.605	0.806	0.705	0.647	0.589	0.624	0.573	0.608	0.641
MS-SSIM [55]	0.585	0.768	0.676	0.638	0.589	0.524	0.572	0.581	0.613
VSI [60]	0.630	0.818	0.724	0.668	0.592	0.597	0.568	0.606	0.646
MAD [30]	0.598	0.770	0.684	0.655	0.593	0.490	0.581	0.580	0.615
VIF [46]	0.556	0.744	0.650	0.651	0.594	0.515	0.597	0.589	0.610
FSIMc [59]	0.627	0.794	0.710	0.660	0.590	0.573	0.581	0.601	0.638
NLPD [29]	0.550	0.764	0.657	0.655	0.584	0.528	0.552	0.580	0.606
GMSD [58]	0.609	0.772	0.690	0.677	0.594	0.517	0.575	0.591	0.624
DeepWSD [31]	0.594	0.788	0.691	0.630	0.588	0.569	0.547	0.584	0.619
DSD [27]	0.580	0.698	0.639	0.654	0.560	0.515	0.595	0.581	0.600
A-FINE [3]	0.706	0.804	0.755	0.684	0.580	0.561	0.620	0.611	0.659
DeepIQA [26]	0.703	0.794	0.748	0.660	0.582	0.585	0.598	0.606	0.654
PieAPP [41]	0.725	0.769	0.747	0.685	0.582	0.594	0.598	0.615	0.659
AHIQ [28]	0.605	0.763	0.684	0.653	0.583	0.531	0.604	0.593	0.623
WaDIQaM-FR [2]	0.637	0.795	0.716	0.657	0.584	0.581	0.578	0.600	0.639
ST-LPIPS [12]	0.719	0.812	0.766	0.696	**0.609**	**0.629**	**0.631**	**0.641**	0.682
LPIPS [61]	0.760	**0.828**	0.794	0.705	0.605	0.625	0.630	**0.641**	0.692
DISTS [6]	**0.772**	0.822	**0.797**	**0.710**	0.600	0.627	0.625	**0.641**	**0.693**
PSIM (Ours)	**0.799**	**0.839**	**0.819**	**0.712**	**0.606**	**0.642**	**0.626**	**0.647**	**0.704**

Table 3. Cross- and intra-dataset results of PSIM and representative IQA models, evaluated in terms of PLCC and SRCC (higher is better). Cross-dataset experiments train models on KADID-10k or PIPAL and evaluate them on unseen datasets (CSIQ, TID2013), assessing their ability to generalize beyond the training data. Intra-dataset evaluations follow the official splits of PieAPP and PIPAL. "–" denotes unavailable or inapplicable results.

	Cross-Dataset Experiments								**Intra-Dataset Experiments**			
Train Dataset	**KADID-10k**				**PIPAL**				**PieAPP**		**PIPAL**	
Test Dataset	CSIQ		TID2013		CSIQ		TID2013		PieAPP		PIPAL	
Method	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑
PieAPP [41]	0.877	0.892	0.859	0.876	–	–	–	–	**0.842**	**0.831**	0.597	0.607
WaDIQaM-FR [2]	0.901	0.909	0.834	0.831	0.834	0.822	0.786	0.739	0.439	0.352	0.548	0.553
LPIPS [61]	0.896	0.876	0.749	0.670	0.857	0.858	0.790	0.760	0.654	0.641	0.633	0.595
DISTS [6]	0.928	0.929	0.855	0.830	**0.862**	0.859	0.803	**0.765**	0.725	0.693	0.687	0.655
AHIQ [28]	**0.955**	**0.951**	**0.899**	**0.901**	0.861	**0.865**	**0.804**	0.763	**0.840**	**0.838**	**0.823**	**0.813**
PSIM (Ours)	**0.949**	**0.950**	**0.909**	**0.897**	**0.935**	**0.925**	**0.861**	**0.838**	0.835	0.816	**0.730**	**0.701**

4.3. Evaluations

How well does PSIM perform as a general image quality model? PSIM achieves strong perceptual alignment with human judgments across diverse distortions and datasets. On the BAPPS dataset—where human pairwise preference serves as the ground truth—traditional measures such as PSNR and SSIM plateau below 0.65 in 2AFC accuracy, while PSIM attains 0.70, as shown in Table 2. Beyond the BAPPS dataset, PSIM generalizes effectively to large-scale subjective datasets, PieAPP and PIPAL, as shown in Table 3 (Intra-Dataset Experiments). The latter remains particularly challenging due to the prevalence of GAN-based distortions. Despite relying on a lightweight VGG-16 backbone, PSIM ranks second overall, which demonstrates the efficacy of our proposed model.

How well does PSIM generalize across datasets? PSIM demonstrates strong cross-dataset generalization. As shown in Table 3 (Cross-Dataset Experiments), PSIM consistently outperforms widely adopted learned perceptual metrics, e.g., LPIPS and DISTS in cross-dataset evaluations. Remarkably, it even surpasses AHIQ, which employs a substantially heavier ResNet50 [16] and ViT [7] backbone. It should be noted in Table 3, results that AHIQ achieves the highest scores in intra-dataset evaluations, but its performance drops noticeably under cross-dataset experiments, indicating a tendency to overfit to the training distribution. In contrast, PSIM shows superior generalization performance across datasets.

How efficient is PSIM in terms of computation? Despite its high accuracy, PSIM remains computationally lightweight, as shown in Figure 4. Compared to large hybrid backbones such as AHIQ (ResNet50 and ViT), PSIM achieves comparable performance while requiring 55.3% fewer FLOPs and 84.9% fewer parameters.

How robust is PSIM under small spatial shifts? Shift robustness is evaluated using the rank-flip rate (r_{rf}), which measures how often image-pair rankings change after small pixel shifts. On BAPPS-ST, PSIM achieves r_{rf} of 2.28, 1.83, and 3.02 for 1–3 px shifts, outperforming traditional metrics and performing comparably to the shift-invariant ST-LPIPS (0.57–1.50), as shown in Table 4.

How well does PSIM perceive color differences? Color perception is evaluated on the SPCD and TID2013-Color datasets. As shown in Table 5 and Table 6, PSIM consistently achieves higher correlations and lower STRESS than existing IQA models. On SPCD (Table 5), PSIM's performance is lower than CIELAB but outperforms all other IQA metrics. On TID2013-Color (Table 6), it surpasses not only CIELAB but also recent learned color-difference models such as CD-Net [57], CD-Flow [4], and MS-SWD [15].

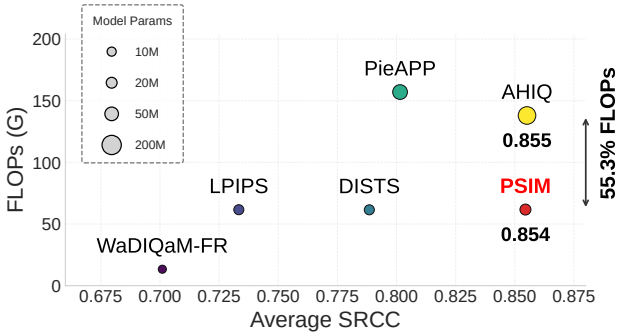

Figure 4. Efficiency benchmark. Average SRCC, computed by averaging SRCC scores across all datasets (from Table 3), is plotted against FLOPs (in G) using a $3 \times 224 \times 224$ input. Our proposed model, PSIM, on average achieves performance comparable to the heavy hybrid backbone–based IQA model AHIQ while requiring 55.3% fewer FLOPs and 84.9% fewer parameters.

Table 4. Evaluation of shift-tolerance (r_{rf}). Smaller values indicate stronger robustness to spatial misalignment on BAPPS-ST. Following [12], BAPPS-ST is created by horizontally shifting each distorted image in BAPPS [61].

Method	$r_{\mathrm{rf}} \downarrow$		
	1-pixel	2-pixel	3-pixel
PSNR	3.09	6.39	9.56
SSIM [56]	3.16	7.20	13.73
MS-SSIM [55]	**2.22**	5.83	10.66
VSI [60]	6.28	17.34	15.16
VIF [46]	4.77	7.97	10.73
FSIMc [59]	5.76	18.37	13.72
NLPD [29]	5.48	10.13	16.45
GMSD [58]	7.04	9.79	14.22
CW-SSIM [45]	3.91	6.88	9.47
GTI-CNN [33] §	3.95	4.91	7.88
PieAPP [41]	2.83	3.19	3.81
WaDIQaM-FR [2]	3.22	4.64	6.14
AHIQ [28]	4.05	5.23	7.21
LPIPS [61]	2.75	3.27	3.66
DISTS [6]	2.85	2.89	4.03
ST-LPIPS [12] §	**0.57**	**1.06**	**1.50**
PSIM (Ours)	2.28	**1.83**	**3.02**

(§) Explicitly designed to be shift-invariant.

4.4. Ablation Study

We evaluate the contribution of each level in the Multi-level Wasserstein Distortion (MWD). As shown in Table 7, using only the base level (L0) captures fine, foveal distortions but lacks broader contextual awareness, and consequently shows comparatively lower performance. Introducing higher levels (L1–L3) progressively improves perfor-

Table 5. Performance evaluation of perceptual color difference on the Smartphone Photography Color Difference (SPCD) dataset. Lower STRESS and higher PLCC/SRCC indicate better performance.

Method	STRESS ↓	PLCC ↑	SRCC ↑
Color Difference Metrics			
CIELAB [34]	**31.952**	**0.714**	**0.665**
General-purpose IQA Models			
WaDIQaM-FR [2]	57.276	0.464	0.449
PieAPP [41]	41.896	0.467	0.451
AHIQ [28]	53.329	0.308	0.209
LPIPS [61]	64.407	0.448	0.396
DISTS [6]	**37.236**	0.582	0.549
PSIM (Ours)	37.892	**0.657**	**0.636**

Table 6. Performance evaluation of perceptual color difference on the TID2013-Color dataset.

Method	STRESS ↓	PLCC ↑	SRCC ↑
Color Difference Metrics			
CIELAB [34]	18.950	0.739	0.749
Learned Color Difference Metrics			
CD-Net [57]	15.962	0.801	0.826
CD-Flow [4]	**14.110**	**0.837**	0.832
MS-SWD [15]	18.579	0.828	**0.839**
General-purpose IQA Models			
PieAPP [41]	20.918	0.620	0.653
LPIPS [61]	15.420	0.816	0.804
DISTS [6]	15.235	0.821	0.805
WaDIQaM-FR [2]	15.882	0.806	0.774
AHIQ [28]	16.230	0.802	0.784
PSIM (Ours)	**11.801**	**0.897**	**0.886**

mance by expanding the effective receptive field and capturing peripheral distortions. Adding the Spatial Coordinate Attention (SCA) module yields a further gain by adaptively emphasizing distortion regions according to their spatial relevance. Furthermore, the results show that multi-scale aggregation is essential for robust perceptual quality assessment. While lower levels focus on localized distortions, higher levels provide complementary contextual information that helps disambiguate visually similar artifacts across different regions. This interplay between local precision and global context enables more stable and consistent predictions across diverse distortion types.

Table 7. Ablation study of Multi-level Wasserstein Distortion (MWD) and Spatial Coordinate Attention (SCA) module. All models are trained on KADID-10k and evaluated on CSIQ.

Model	MWD Levels				SCA	PLCC ↑ / SRCC ↑
	L0	L1	L2	L3		
(1)	✓					0.918 / 0.929
(2)	✓	✓				0.926 / 0.932
(3)	✓	✓	✓			0.935 / 0.939
(4)	✓	✓	✓	✓		0.941 / 0.942
(5)	✓	✓	✓	✓	✓	**0.949 / 0.950**

5. Conclusion

PSIM provides strong perceptual alignment with human judgments by integrating both large-scale appearance cues and fine-grained structural deviations within a unified multi-level distortion formulation. By comparing feature statistics over progressively larger spatial supports, PSIM naturally captures distortions that are immediately noticeable in the periphery as well as subtle changes that require focused foveal inspection. Owing to this advanced distortion modeling, PSIM consistently outperforms most IQA metrics while requiring significantly lower computation. Moreover, PSIM demonstrates remarkably strong performance in two areas that are typically challenging for general-purpose IQA models. First, the model exhibits reliable sensitivity to perceptual color differences. The hierarchical moment comparisons inherently couple color and form in a manner that closely reflects human visual perception, enabling PSIM to perform well on both traditional and device-dependent color-difference benchmarks. Second, PSIM demonstrates strong spatial robustness. The multi-level Wasserstein formulation stabilizes perceptual responses to small spatial shifts, resulting in consistently low rank-flip rates.

6. Limitations

PSIM demonstrates competitive performance across a wide range of datasets and distortion types, but several limitations remain. First, PSIM is a full-reference metric and therefore requires access to an undistorted reference image; extending the multi-level Wasserstein distortion framework to no-reference or reduced-reference settings would broaden its applicability and make it suitable for a wider range of real-world scenarios. Second, while the multi-level design is inspired by the hierarchical organization of human perception, PSIM does not explicitly model task-dependent attention, fixation patterns, or individual observer variability, all of which can influence perceived quality under more complex viewing conditions.

References

[1] Jona Ballé, Luca Versari, Emilien Dupont, Hyunjik Kim, and Matthias Bauer. Good, cheap, and fast: Overfitted image compression with wasserstein distortion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23259–23268, 2025. 2

[2] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1): 206–219, 2017. 6, 7, 8

[3] Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12742–12752, 2025. 6

[4] Haoyu Chen, Zhihua Wang, Yang Yang, Qilin Sun, and Kede Ma. Learning a deep color difference metric for photographic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22242–22251, 2023. 5, 7, 8

[5] M Concetta Morrone and David Charles Burr. Feature detection in human vision: A phase-dependent energy model. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 235(1280):221–245, 1988. 2

[6] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 2, 5, 6, 7, 8

[7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[8] Li Fei-Fei, Aniruddha Iyer, Christof Koch, and Pietro Perona. What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1):1–29, 2007. 1

[9] Sylvain Fischer, Filip Šroubek, Laurent Perrinet, Rafael Redondo, and Gabriel Cristóbal. Self-invertible 2d log-gabor wavelets. *International Journal of Computer Vision*, 75(2): 231–246, 2007. 2

[10] Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201, 2011. 3

[11] Pedro A Garcia, Rafael Huertas, Manuel Melgosa, and Guihua Cui. Measurement of the relationship between perceived and computed color differences. *Journal of the Optical Society of America A*, 24(7):1823–1829, 2007. 5

[12] Abhijay Ghildyal and Feng Liu. Shift-tolerant perceptual similarity metric. In *European Conference on Computer Vision*, pages 91–107. Springer, 2022. 5, 6, 7

[13] Michelle R. Greene and Aude Oliva. Recognition of natural scenes from global properties. *Journal of Vision*, 9(1):1–12, 2009. 1

[14] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S Ren, Radu Timofte, Yuan Gong, Shanshan Lao, Shuwei Shi, Jiahao Wang, Sidi Yang, et al. Ntire 2022 challenge on perceptual image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 951–967, 2022. 5

[15] Jiaqi He, Zhihua Wang, Leon Wang, Tsein-I Liu, Yuming Fang, Qilin Sun, and Kede Ma. Multiscale sliced wasserstein distances as perceptual color difference measures. In *European Conference on Computer Vision*, pages 425–442. Springer, 2024. 3, 7, 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[17] Eric Heitz, Kenneth Vanhoey, Thomas Chambon, and Laurent Belcour. A sliced wasserstein loss for neural texture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9412–9420, 2021. 2

[18] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 2

[19] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021. 5

[20] Antoine Houdard, Arthur Leclaire, Nicolas Papadakis, and Julien Rabin. A generative model for texture synthesis based on optimal transport between feature distributions. *Journal of Mathematical Imaging and Vision*, 65(1):4–28, 2023. 2

[21] Howard C. Hughes, G. Nozawa, and Francis L. Kitterle. Global precedence, spatial frequency channels, and the processing of local–global stimuli. *Journal of Experimental Psychology: General*, 125(2):260–278, 1996. 1, 2

[22] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, 2000. 1

[23] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 1

[24] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European conference on computer vision*, pages 633–651. Springer, 2020. 5

[25] Brian W Keelan and Hitoshi Urabe. Iso 20462: a psychophysical image quality measurement standard. In *Image Quality and System Performance*, pages 181–189. SPIE, 2003. 5

[26] Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1676–1684, 2017. 6

[27] Idan Kligvasser, Tamar Shaham, Yuval Bahat, and Tomer Michaeli. Deep self-dissimilarities as powerful visual fingerprints. *Advances in Neural Information Processing Systems*, 34:3939–3951, 2021. 6

[28] Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang. Attentions

help cnns see better: Attention-based hybrid image quality assessment network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1140–1149, 2022. 6, 7, 8

[29] Valero Laparra, Johannes Ballé, Alexander Berardino, and Eero P Simoncelli. Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging*, 28: 1–6, 2016. 6, 7

[30] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006–011006, 2010. 2, 5, 6

[31] Xingran Liao, Baoliang Chen, Hanwei Zhu, Shiqi Wang, Mingliang Zhou, and Sam Kwong. Deepwsd: Projecting degradations in perceptual space to wasserstein distance in deep feature space. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 970–978, 2022. 2, 6

[32] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 5

[33] Kede Ma, Zhengfang Duanmu, and Zhou Wang. Geometric transformation invariant image quality assessment using convolutional neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6732–6736. IEEE, 2018. 7

[34] Marc Mahy, Luc Van Eycken, and André Oosterlinck. Evaluation of uniform color spaces developed after the adoption of cielab and cieluv. *Color Research & Application*, 19(2): 105–121, 1994. 2, 8

[35] Céline Mancas-Thillou and Bernard Gosselin. Character segmentation-by-recognition using log-gabor filters. In *18th International Conference on Pattern Recognition (ICPR'06)*, pages 901–904. IEEE, 2006. 2

[36] K. T. Mullen. The contrast sensitivity of human colour vision to red–green and blue–yellow chromatic gratings. *The Journal of Physiology*, 359:381–400, 1985. 2

[37] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 1

[38] Aude Oliva and Antonio Torralba. Building the gist of a scene. In *Progress in Brain Research*, pages 23–36. Elsevier, 2006. 2

[39] Francois Pitie, Anil C Kokaram, and Rozenn Dahyot. N-dimensional probability density function transfer and its application to color transfer. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 1434–1439. IEEE, 2005. 2

[40] Nikolay Ponomarenko, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Lina Jin, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Color image database tid2013: Peculiarities and preliminary results. In *European workshop on visual information processing (EUVIP)*, pages 106–111. IEEE, 2013. 5

[41] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018. 5, 6, 7, 8, 1

[42] Yang Qiu and Aaron B Wagner. Low-rate, low-distortion compression with wasserstein distortion. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 855–860. IEEE, 2024. 2

[43] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International conference on scale space and variational methods in computer vision*, pages 435–446. Springer, 2011. 3

[44] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000. 2

[45] Mehul P Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing*, 18(11):2385–2401, 2009. 2, 7

[46] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15 (2):430–444, 2006. 6, 7

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[48] Guillaume Tartavel, Gabriel Peyré, and Yann Gousseau. Wasserstein loss for image synthesis and restoration. *SIAM Journal on Imaging Sciences*, 9(4):1726–1755, 2016. 2

[49] Jonathan Vacher, Aida Davila, Adam Kohn, and Ruben Coen-Cagli. Texture interpolation for probing visual perception. *Advances in neural information processing systems*, 33: 22146–22157, 2020. 2

[50] Cédric Villani et al. *Optimal transport: old and new*. Springer, 2008. 2

[51] Brian A. Wandell. *Foundations of Vision*. Sinauer Associates, 1995. 2

[52] Zhou Wang and A.C. Bovik. Embedded foveation image coding. *IEEE Transactions on Image Processing*, 10(10): 1397–1410, 2001. 3

[53] Zhou Wang and Alan C Bovik. Foveated image and video coding. In *Digital Video Image Quality and Perceptual Coding*, pages 431–458. CRC Press, 2017.

[54] Zhou Wang, Ligang Lu, and A.C. Bovik. Foveation scalable video coding with automatic fixation selection. *IEEE Transactions on Image Processing*, 12(2):243–254, 2003. 3

[55] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pages 1398–1402 Vol.2, 2003. 6, 7

[56] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2, 6, 7

[57] Zhihua Wang, Keshuo Xu, Yang Yang, Jianlei Dong, Shuhang Gu, Lihao Xu, Yuming Fang, and Kede Ma.

Measuring perceptual color differences of smartphone photographs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10114–10128, 2023. 5, 7, 8

[58] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE transactions on image processing*, 23(2):684–695, 2013. 6, 7

[59] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011. 2, 6, 7

[60] Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing*, 23(10): 4270–4281, 2014. 6, 7

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2, 5, 6, 7, 8, 1