





# Multiscale Sliced Wasserstein Distances as Perceptual Color Difference Measures

Jiaqi He<sup>1,2</sup>, Zihua Wang<sup>3</sup>, Leon Wang<sup>4</sup>, Tsein-I Liu<sup>4</sup>, Yuming Fang<sup>5</sup>,  
Qilin Sun<sup>6\*</sup>, and Kede Ma<sup>1,2</sup>

<sup>1</sup> Department of Computer Science, City University of Hong Kong

<sup>2</sup> Shenzhen Research Institute, City University of Hong Kong

<sup>3</sup> Department of Engineering, Shenzhen MSU-BIT University

<sup>4</sup> Guangdong OPPO Mobile Telecommunications Corp., Ltd.

<sup>5</sup> School of Information Management, Jiangxi University of Finance and Economics

<sup>6</sup> School of Data Science, The Chinese University of Hong Kong (Shenzhen)

jqhe00@mail.ustc.edu.cn   zihua.wang@my.cityu.edu.hk

{leon.wang, simon}@oppo.com   fa0001ng@e.ntu.edu.sg

sunqilin@cuhk.edu.cn   kede.ma@cityu.edu.hk

**Abstract.** Contemporary color difference (CD) measures for photographic images typically operate by comparing *co-located* pixels, patches in a “perceptually uniform” color space, or features in a learned latent space. Consequently, these measures inadequately capture the human color perception of misaligned image pairs, which are prevalent in digital photography (*e.g.*, the same scene captured by different smartphones). In this paper, we describe a perceptual CD measure based on the multiscale sliced Wasserstein distance, which facilitates efficient comparisons between *non-local* patches of similar color and structure. This aligns with the modern understanding of color perception, where color and structure are inextricably interdependent as a unitary process of perceptual organization. Meanwhile, our method is easy to implement and training-free. Experimental results indicate that our CD measure performs favorably in assessing CDs in photographic images, and consistently surpasses competing models in the presence of image misalignment. Additionally, we empirically verify that our measure functions as a metric in the mathematical sense, and show its promise as a loss function for image and video color transfer tasks. The code is available at <https://github.com/real-hjq/MS-SWD>.

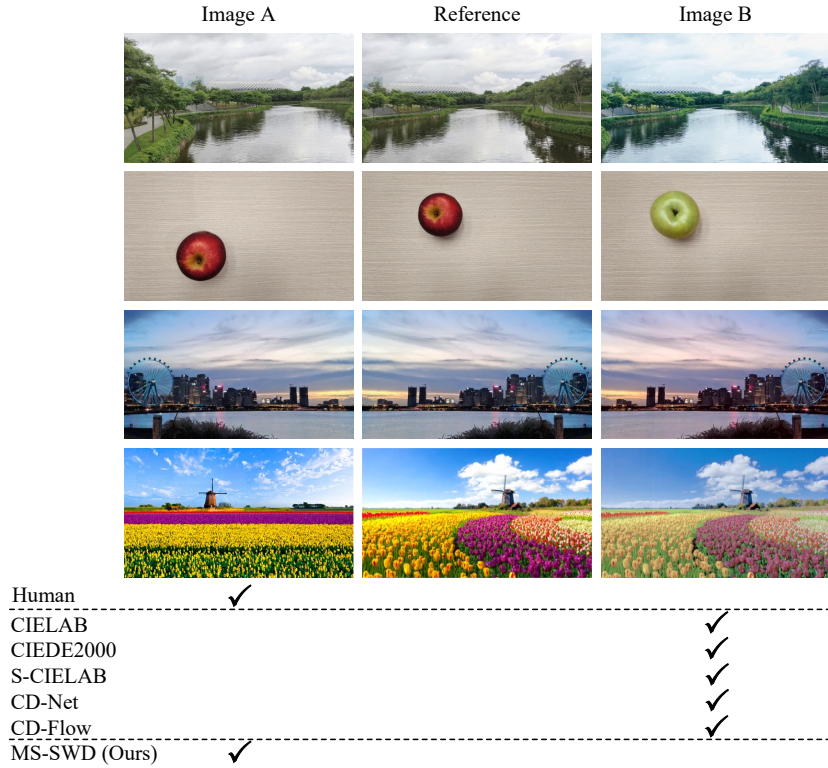
**Keywords:** Color difference assessment · Sliced Wasserstein distance · Multiscale analysis

## 1 Introduction

Measuring perceptual color differences (CDs) in photographic images is a prerequisite in many image processing and computer vision tasks [43]. The predominant and scientifically well-founded theme is the pursuit of a perceptually

---

\* Corresponding author.



**Fig. 1:** Which image is closer to the reference in terms of color appearance? Contemporary CD measures that seek *co-located* comparisons often fail to explain human judgments. The proposed MS-SWD measure based on the multiscale sliced Wasserstein distance aligns with human color perception in these four challenging cases of image misalignment: global motion due to camera movement (first row), local motion due to object displacement (second row), horizontal flipping (third row), and similar natural scenes from different viewpoints (last row).

uniform color space. Within such a space, numerical distances of two color points correspond directly to perceptual differences, regardless of their positions within the color spectrum. CIELAB and CIELUV, introduced by the Commission Internationale de l’Éclairage (CIE) in 1976, represent two of the pioneering perceptually uniform color spaces [31]. CD metrics (*e.g.*, CIELAB  $\Delta E_{ab}^*$ ) derived from these color spaces have been rapidly adopted in various industrial sectors. However, subsequent analysis revealed that these color spaces are insufficient for accurately quantifying small to medium CDs [18]. In response, more sophisticated metrics (*e.g.*, CIEDE2000 [29]) were introduced to address various aspects of perceptual non-uniformity, whose rectified parameters were determined by fitting chromaticity discrimination (*i.e.*, MacAdam) ellipses [29] obtained from subjective experiments.

Traditional CD metrics have demonstrated efficacy in predicting perceived differences between uniformly colored patches [33]. A straightforward adaptation for assessing photographic images of natural scenes involves averaging the CDs between *co-located* pixels [9]. However, this naïve extension shows a marginal correlation with human color perception, particularly when various sources of image misalignment are present (see Fig. 1).

Over the past decades, an extensive body of psychophysical and perceptual studies [2, 6, 26, 45] has provided a more compelling understanding of color perception: *color, structure, and motion are inextricably interdependent as a unitary process of perceptual organization* [22, 42]. Drawing inspiration from these scientific insights, researchers have started to incorporate spatial modeling as a crucial component of CD measures [9, 21, 34, 51–53, 57]. For instance, Zhang and Wandell [57] described a spatial extension of CIELAB  $\Delta E_{ab}^*$  by applying lowpass filtering in an opponent color space as a preprocessing step. Wang *et al.* [52] adopted a deep learning approach, training a lightweight neural network for “color space transform”, followed by a learned Mahalanobis metric for distance calculation. Again, these models are designed to compare *co-located* patches or features, making them susceptible to image misalignment (see Fig. 1).

In this paper, we introduce a perceptual CD measure that facilitates efficient comparisons between *non-local* patches of similar color appearance and structural information. Our measure is primarily inspired by the seminal work of Eneka and Weiss [17], who generated natural images by direct patch distribution matching. In a similar spirit, we compute the perceptual CD between two photographic images as the statistical distance of their patch distributions across multiple scales. To compare two images, we start by building two Gaussian pyramids in a perceptually more uniform CLELAB color space. Next, we opt for the sliced Wasserstein distance (SWD) [38] to calculate the CD between the images at each scale. Finally, we average these CD values across all scales to obtain the overall CD estimate. The resulting measure, the multiscale SWD (MS-SWD), is conceptually simple and respects the modern view that color and structure interact inextricably in visual cortical processing. Meanwhile, MS-SWD is easy to implement and training-free.

We validate the proposed MS-SWD on the large-scale SPCD dataset [52]. Remarkably, even without training, MS-SWD excels in evaluating CDs in photographic images, especially when there is large image misalignment. Additionally, we empirically show that MS-SWD behaves as a metric in the mathematical sense, and serves as a valid loss function for perceptual optimization in image and video color transfer tasks.

## 2 Related Work

In this section, we present an overview of two areas of research closely related to our work: CD measures and patch matching methods in computer vision.

## 2.1 CD Measures

The development of CD measures has a rich history. In 1976, CIE recommended the CIELAB color space [39], in which the Euclidean distance,  $\Delta E_{ab}^*$ , has been widely accepted as the de facto CD metric. Shortly after its introduction, researchers realized that CIELAB is not perfectly perceptually uniform. This led to the proposal of more sophisticated metrics such as CMC (1:c) [12], CIE94 [32], and CIEDE2000 [29]. These metrics generally assume a standard viewing environment, *e.g.*, using the standard illuminant D65 and the 2° standard observer, with reference to a white background, and do not explicitly account for varying viewing conditions and ambient environments. To address this, metrics like CIECAM02 [30] and CIECAM16 [27] were developed to predict changes in color appearance under varying viewing conditions. CIELAB-based methods are best suited for matching uniformly colored patches.

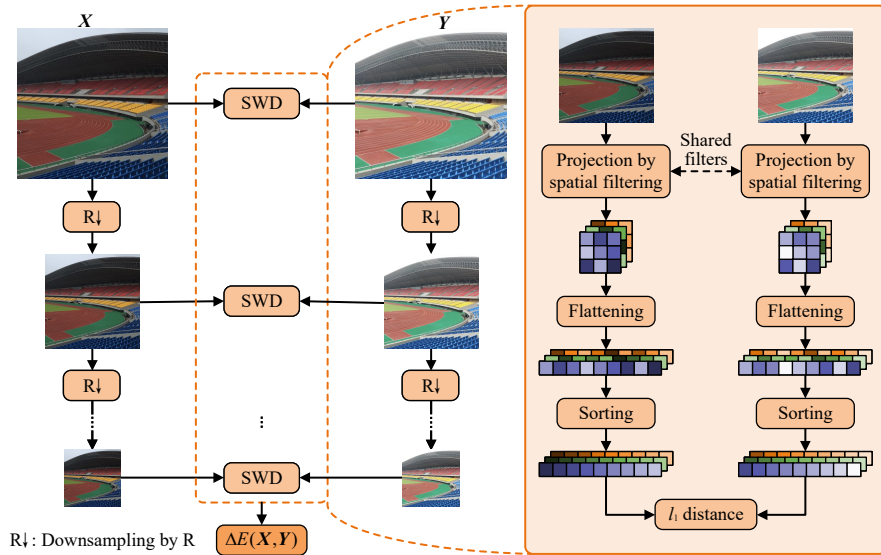
When assessing CDs in photographic images of natural scenes, humans tend to compare similar regions, co-located or not, in terms of color appearance and structural information within a broader spatial context [22, 42]. Zhang and Wandell [57] made one of the first attempts to extend CIELAB to S-CIELAB by incorporating spatial lowpass filtering as front-end preprocessing. Choudhury *et al.* [11] designed preprocessing filters based on the contrast sensitivity functions (CSFs) of the human eye. Hong *et al.* [21] computed the weighted sum of pixel-wise CDs, prioritizing spatially homogeneous regions that cover large areas or have large predicted CDs. Similarly, Ortiz-Jaramillo *et al.* [34] weighted patch-wise CDs using an image segmentation map computed from local binary patterns. The effectiveness of these spatially extended CD measures has been demonstrated only on small-scale private datasets with a few hand-picked images. Close to ours, Lee *et al.* [25] enabled *non-local* CD assessment by histogram intersection<sup>1</sup>, which, however, completely throws away spatial information that is crucial for human color perception. Wang *et al.* [52] demonstrated on the large-scale SPCD dataset that these simple spatial extensions may not yield noticeable performance improvements. As a result, they took a deep learning approach, and trained CD-Net [52] and CD-Flow [9] directly on SPCD. Inspired by [17], we tackle CD assessment of photographic images through multiscale patch distribution matching. MS-SWD enables efficient *non-local* patch comparisons without using any specialized training.

## 2.2 Patch Matching Methods

Patch matching is a fundamental technique in computer vision with diverse applications such as image denoising, image stitching, texture synthesis, image and video completion, 3D reconstruction, and object recognition. In patch matching, the search for patch nearest neighbors is often computationally intensive due to the need to explicitly establish bidirectional mappings [5, 24, 47]. In recent years,

<sup>1</sup> Histogram intersection measures the similarity between two normalized histograms by summing the minimum values of corresponding bins.





**Fig. 2:** System diagram of the proposed MS-SWD for perceptual CD assessment.

generative adversarial networks (GANs) and their derivatives have largely overtaken traditional patch matching methods. SinGAN [41] and InGAN [46] are representative examples that indirectly match patch distributions of two images by training patch-based discriminators. To seek a direct (patch) distribution matching without involving time-consuming training, SWD has been explored in various image generation tasks, in the raw pixel domain [13, 17, 23], wavelet domain [38], and VGG feature domain [40]. Our MS-SWD measure draws significant inspiration from [17] but for a different purpose (*i.e.*, CD assessment) with a different motivation (*i.e.*, non-local patch comparison).

### 3 MS-SWDs as Perceptual CD Measures

In this section, we first introduce the necessary preliminaries - SWD, and then present in detail our MS-SWD measure for perceptual CD assessment. Fig. 2 shows the system diagram of MS-SWD.

#### 3.1 SWD

Among various statistical distances between two probability distributions, the Wasserstein distance enjoys several advantages, including 1) intuitive interpretation (as the minimum “cost” of transforming one distribution into another), 2) sensitivity to distribution shape (by computing the actual geometric distances between points in the distributions), 3) robustness to support differences (even

when the supports of the two distributions do not overlap), and 4) smooth gradients for optimization [4]. The 1-Wasserstein distance (also known as the earth mover’s distance) between two probability distributions  $\mu$  and  $\nu$  is defined as

$$\text{WD}(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \gamma} \|\mathbf{x} - \mathbf{y}\|_1, \quad (1)$$

where  $\Gamma(\mu, \nu)$  denotes the set of all joint distributions (couplings)  $\gamma$  whose marginals are  $\mu$  and  $\nu$ . The Wasserstein distance is notoriously challenging to implement due to its high computational complexity, especially when working with empirical distributions represented by high-dimensional samples<sup>2</sup>. To reduce the computational complexity and improve the scalability and robustness to high dimensions, Rabin *et al.* [38] introduced SWD by projecting the high-dimensional data onto a lower-dimensional subspace and then calculating the Wasserstein distance therein. When the projected space is one-dimensional, SWD can be mathematically expressed as

$$\text{SWD}(\mathbf{U}, \mathbf{V}) = E_{\mathbf{w} \sim \mathcal{U}(\mathbb{S}^{N-1})} \text{WD}(\mathbf{U}\mathbf{w}, \mathbf{V}\mathbf{w}), \quad (2)$$

where  $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{M \times N}$ ,  $M$  is the number of samples to represent the empirical distributions, and  $N$  is the sample dimension.  $\mathbb{S}^{N-1} := \{\mathbf{w} \in \mathbb{R}^{N \times 1} \mid \|\mathbf{w}\|_2 = 1\}$  for any  $N \geq 2$  is the unit hyper-sphere,  $\mathcal{U}(\mathbb{S}^{N-1})$  is the uniform distribution defined over  $\mathbb{S}^{N-1}$ , and  $\mathbb{E}_{\mathbf{w}}$  is the expectation over the random unit vector  $\mathbf{w}$ . In Eq. (2), the one-dimensional Wasserstein distance can be efficiently calculated by *sorting* the projected samples and computing the  $\ell_1$ -distance between the sorted samples [44]. SWD typically reduces the computational complexity from  $\mathcal{O}(M^{2.5})$  [36] to  $\mathcal{O}(M \log M)$ .

### 3.2 MS-SWD for CD Assessment

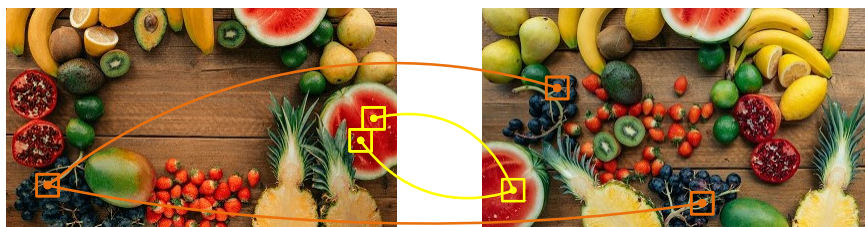
Let  $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$  and  $\mathbf{Y} \in \mathbb{R}^{H \times W \times 3}$  be two photographic images that are possibly misaligned, where  $H$  and  $W$  are the image height and width, respectively. We first construct two Gaussian pyramids,  $\{\mathbf{X}^{(i)}\}_{i=1}^K$  and  $\{\mathbf{Y}^{(i)}\}_{i=1}^K$  by iteratively applying a Gaussian filter and downsampling the filtered image by a factor of  $R$ , where  $\mathbf{X}^{(i)}, \mathbf{Y}^{(i)} \in \mathbb{R}^{\lfloor H/2^{i-1} \rfloor \times \lfloor W/2^{i-1} \rfloor \times 3}$  and  $K$  denotes the number of scales. We then represent  $\mathbf{X}^{(i)}$  and  $\mathbf{Y}^{(i)}$ , for  $1 \leq i \leq K$ , in the CIELAB color space, where we observe significant performance gains over the sRGB color space. Although spatial pre-filtering of  $\mathbf{X}^{(i)}$  and  $\mathbf{Y}^{(i)}$  based on CSFs [11, 57] can also be applied, it does not yield noticeable improvements and is therefore excluded from our current implementation.

For ease of mathematical description, we also use an alternative notation for  $\mathbf{X}^{(i)}$ , denoted as  $\mathbf{X}_{\text{col}}^{(i)} \in \mathbb{R}^{M \times (N \times 3)}$ , in which we rearrange the  $M$  overlapping image patches of size  $\sqrt{N} \times \sqrt{N} \times 3$  into columns. The transformation from  $\mathbf{X}^{(i)}$  to  $\mathbf{X}_{\text{col}}^{(i)}$  can be efficiently achieved using the `img2col()` operator, a common tool in image processing for implementing convolutions.

<sup>2</sup> This corresponds to solving a large-scale linear programming problem, which is painfully slow.

**Algorithm 1** MS-SWD for Perceptual CD Assessment

- 
- 1: **Input:** A pair of photographic images that are possibly misaligned,  $(\mathbf{X}, \mathbf{Y})$ , the number of scales,  $K$ , and the number of random projections,  $P$
  - 2: **Output:** Predicted CD,  $\Delta E(\mathbf{X}, \mathbf{Y})$
  - 3: Build Gaussian pyramids  $\{\mathbf{X}^{(i)}\}_{i=1}^K$  and  $\{\mathbf{Y}^{(i)}\}_{i=1}^K$ , where  $\mathbf{X}^{(1)} = \mathbf{X}$  and  $\mathbf{Y}^{(1)} = \mathbf{Y}$
  - 4: Convert  $\{\mathbf{X}^{(i)}\}_{i=1}^K$  and  $\{\mathbf{Y}^{(i)}\}_{i=1}^K$  from the sRGB to CIELAB color space
  - 5:  $\Delta E \leftarrow 0$
  - 6: **for**  $i \leftarrow 1$  **to**  $K$  **do**
  - 7:     **for**  $j \leftarrow 1$  **to**  $P$  **do**
  - 8:          $\mathbf{w} \sim \mathcal{U}(\mathbb{S}^{N \times 3 - 1})$
  - 9:          $\mathbf{w} \leftarrow \text{unflat}(\mathbf{w})$                      ▷ “unflat()” converts a vector into a tensor
  - 10:          $\mathbf{x} \leftarrow \text{flat}(\text{Conv2d}(\mathbf{X}^{(i)}, \mathbf{w}, \text{'reflect'}))$      ▷ “flat()” is the inverse of “unflat()”
  - 11:          $\mathbf{y} \leftarrow \text{flat}(\text{Conv2d}(\mathbf{Y}^{(i)}, \mathbf{w}, \text{'reflect'}))$
  - 12:          $\Delta E \leftarrow \Delta E + \frac{1}{M} \|\text{sort}(\mathbf{x}) - \text{sort}(\mathbf{y})\|_1$
  - 13:     **end for**
  - 14: **end for**
  - 15:  $\Delta E(\mathbf{X}, \mathbf{Y}) \leftarrow \frac{1}{KP} \Delta E$
- 

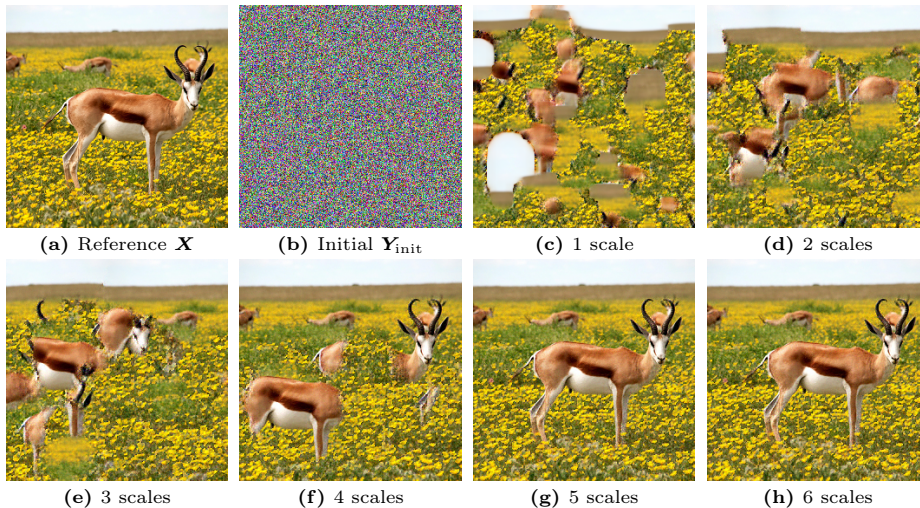


**Fig. 3:** The  $\text{sort}()$  operator in MS-SWD enables efficient comparisons of non-local patches with similar color appearance and structural information. Each curve represents a different random projection; the patches at the two ends of the curve share the same rank (*i.e.*, correspondence) after sorting, thus subject to CD calculation.

After generating the Gaussian pyramids  $\{\mathbf{X}_{\text{col}}^{(i)}\}_{i=1}^K$  and  $\{\mathbf{Y}_{\text{col}}^{(i)}\}_{i=1}^K$  in CIELAB, we calculate the predicted CD,  $\Delta E(\mathbf{X}, \mathbf{Y})$ , between the two images  $\mathbf{X}$  and  $\mathbf{Y}$  by averaging SWD from Eq. (2) across all scales:

$$\Delta E(\mathbf{X}, \mathbf{Y}) = \frac{1}{K} \sum_{i=1}^K \text{SWD}(\mathbf{X}_{\text{col}}^{(i)}, \mathbf{Y}_{\text{col}}^{(i)}) = \frac{1}{KP} \sum_{i=1}^K \sum_{j=1}^P \text{WD}(\mathbf{X}_{\text{col}}^{(i)} \mathbf{w}^{(j)}, \mathbf{Y}_{\text{col}}^{(i)} \mathbf{w}^{(j)}), \quad (3)$$

where the expectation in Eq. (2) is approximated by the average over a set of  $P$  random unit projections  $\{\mathbf{w}^{(j)}\}_{j=1}^P$ . It is important to note that the matrix multiplication,  $\mathbf{X}_{\text{col}}^{(i)} \mathbf{w}^{(j)}$  (and  $\mathbf{Y}_{\text{col}}^{(i)} \mathbf{w}^{(j)}$ ) can be implemented by a single convolution. The computational procedure of MS-SWD for perceptual CD assessment is given in Algorithm 1, in which we particularly emphasize the importance of the  $\text{sort}()$  operator in Step 12. Provided that the two images differ primarily in color appearance, patches of similar color and structure, whether co-located or



**Fig. 4:** Illustration of multiscale analysis in ensuring pixel-level image fidelity. Images (c)-(h) are generated by minimizing  $\Delta E(\mathbf{X}, \mathbf{Y})$  with respect to  $\mathbf{Y}$  to match (a) the reference image  $\mathbf{X}$ , starting from (b) the initial Gaussian noise image  $\mathbf{Y}_{\text{init}}$  and for different values of  $K$ .

not, are likely to have similar ranks (*i.e.*, correspondences) in the projected space after sorting, thus subject to CD calculation. This 1) facilitates efficient non-local patch comparisons without the need to compute patch nearest neighbors [17], and meanwhile 2) respects the modern view of human color perception [2] that color and structure are inextricably interdependent as a unitary process of perceptual organization [22, 42].

Multiscale analysis is another crucial aspect of our approach, although, through our internal subjective testing, it seems that human color perception of photographic images remains fairly stable under varying viewing conditions related to image scale (*e.g.*, display resolution and viewing distance). This is because matching the single-scale patch distribution is unlikely to guarantee image fidelity at the pixel level. As illustrated in Fig. 4, we begin with a Gaussian noise image  $\mathbf{Y}_{\text{init}}$  of the same size as the reference image  $\mathbf{X}$ , and iteratively refine  $\mathbf{Y}_{\text{init}}$  by minimizing Eq. (3) of varying  $K$  using gradient-based optimization. With a limited number of scales, the optimized image fails to recover the structural details of the reference, and exhibits perceptually annoying distortions (*e.g.*, object discontinuity), despite the MS-SWD value being close to zero. Our empirical observations indicate that for a  $256 \times 256$  image, using five scales suffices to recover the reference image within the human perceptual threshold.

## 4 Experiments

In this section, we first compare the proposed MS-SWD with existing CD measures on SPCD [52, 55]. We then perform a series of ablation studies to validate

**Table 1:** Performance evaluation of CD measures on the SPCD dataset. The top section lists standard CD formulae derived from uniformly colored patches. The second section contains CD measures adapted for photographic images. The third section includes general-purpose image quality models. The fourth section consists of just-noticeable difference measures. The top two methods are highlighted in boldface.

Method	Perfectly aligned pairs			Non-perfectly aligned pairs			All		
	STRESS↓	PLCC↑	SRCC↑	STRESS↓	PLCC↑	SRCC↑	STRESS↓	PLCC↑	SRCC↑
CIELAB [39]	31.280	0.790	0.774	30.009	0.683	0.577	31.952	0.714	0.665
CIE94 [32]	34.643	0.786	0.772	30.147	0.692	0.572	34.305	0.709	0.654
CIEDE2000 [29]	29.862	<b>0.827</b>	<b>0.821</b>	30.650	0.653	0.561	31.431	0.725	0.685
CIECAM02 [30]	<b>24.779</b>	0.823	0.820	29.339	0.679	0.612	<b>27.151</b>	<b>0.748</b>	0.725
CIECAM16 [27]	<b>23.901</b>	0.818	0.820	29.934	0.661	0.600	<b>26.817</b>	0.743	<b>0.726</b>
S-CIELAB [57]	29.977	0.824	0.819	32.057	0.627	0.522	32.760	0.699	0.657
Lee05 [25]	58.652	0.728	0.735	56.515	0.636	0.637	58.031	0.697	0.710
Hong06 [21]	60.361	0.732	0.811	57.466	0.538	0.462	61.242	0.609	0.634
Ouni08 [35]	29.864	<b>0.826</b>	<b>0.821</b>	30.657	0.653	0.561	31.435	0.722	0.685
PieAPP [37]	41.550	0.502	0.511	39.619	0.483	0.410	41.896	0.467	0.451
LPIPS [56]	40.972	0.767	0.766	46.402	0.272	0.237	64.407	0.448	0.396
4LIP [3]	29.368	0.743	0.714	<b>27.559</b>	<b>0.730</b>	<b>0.638</b>	29.197	0.715	0.663
DISTS [16]	33.417	0.725	0.722	33.244	0.571	0.495	37.236	0.582	0.549
A-DISTS [14]	38.190	0.661	0.663	42.488	0.387	0.365	51.360	0.424	0.384
ST-LPIPS [20]	37.234	0.810	0.813	43.912	0.399	0.362	50.579	0.535	0.512
DeepWSD [28]	31.760	0.539	0.540	43.342	0.055	0.015	49.705	0.136	0.180
Chou07 [10]	52.463	0.780	0.793	37.704	0.645	0.518	49.581	0.667	0.615
Butteraugli [1]	42.691	0.615	0.589	48.764	0.205	0.193	54.801	0.372	0.354
PIM-5 [7]	58.737	0.685	0.695	48.454	0.556	0.482	60.346	0.455	0.480
MS-SWD (Ours)	34.040	0.778	0.755	<b>28.363</b>	<b>0.841</b>	<b>0.805</b>	32.781	<b>0.794</b>	<b>0.772</b>

the key design choices of MS-SWD. Finally, we explore the use of MS-SWD in guiding image and video color transfer.

#### 4.1 Main Experiments

**SPCD Dataset.** We conduct the main experiments on SPCD [52, 55], which is the largest image dataset currently available for CD assessment. SPCD comprises 30,000 photographic image pairs that span diverse real-world picture-taking scenarios, featuring great variations in foreground elements, background complexity, lighting and weather conditions, and camera modes. Of these 30,000 pairs, 10,005 are non-perfectly aligned, captured by six flagship smartphones, while the remaining pairs are perfectly aligned with CDs induced through simulated color alterations.

**Implementation Details.** MS-SWD does not contain any trainable parameters; all its hyper-parameters are inherited directly from previous studies. These include the number of scales  $K = 5$ , the downsampling factor  $R = 2$ , and the filter size  $\sqrt{N} = 11$  with a stride of 1 from the MS-SSIM paper [54], and the number of random unit projections  $P = 128$  from the GPDM paper [17]. Throughout all experiments, we resize the images to  $256 \times 256$  for testing.

**Evaluation Criteria.** We employ three evaluation criteria: standardized residual sum of squares (STRESS) [19], Pearson linear correlation coefficient (PLCC), and Spearman’s rank correlation coefficient (SRCC). STRESS assesses the pre-

diction accuracy and statistical significance, and is defined as

$$\text{STRESS} = 100 \sqrt{\frac{\sum_{i=1}^I (\Delta E_i - F \Delta V_i)^2}{F^2 \sum_{i=1}^I (\Delta V_i)^2}}, \quad (4)$$

where  $I$  is the number of test pairs, and  $F$  is the scale correction factor between predicted CDs,  $\Delta E$  and ground-truth CDs,  $\Delta V$ :

$$F = \frac{\sum_{i=1}^I (\Delta E_i)^2}{\sum_{i=1}^I \Delta E_i \Delta V_i}. \quad (5)$$

A smaller value of STRESS indicates a tighter fit. PLCC and SRCC measure the prediction linearity and monotonicity, respectively, with a larger value indicating better correlation. Before calculating PLCC, we linearize model predictions by fitting a four-parameter logistic function, as suggested in [50].

**SPCD Results.** We compare MS-SWD against 19 state-of-the-art methods, categorized as follows: 1) CD metrics derived from uniformly colored patches, including CIELAB [39], CIE94 [32], CIEDE2000 [29], CIECAM02 [30], and CIECAM16 [27]; 2) CD measures designed for photographic images, including S-CIELAB [57], Lee05 [25], Hong06 [21], and Ouni08 [35]; 3) general-purpose image quality models, including PieAPP [37], LPIPS [56], FLIP [3], DISTS [16], A-DISTS [14], ST-LPIPS [20], and DeepWSD [28]; 4) just-noticeable difference methods, including Chou07 [10], Butteraugli [1], and PIM-5 [7]. We use the official implementations provided by the original authors for CIECAM02, CIECAM16, Butteraugli, PIM-5, and the seven general-purpose image quality models. For the remaining methods, we use the implementations provided by Jaramillo *et al.* [34].

From the results in Table 1, we have several key observations. First, the majority of CD methods exhibit diminished performance for non-perfectly aligned pairs due to co-located comparisons, even when the misalignment is imperceptible to the human eye. Second, CD formulae recommended by CIE, along with their spatial extensions S-CIELAB and Ouni08, deliver outstanding correlation with human color perception, especially on perfectly aligned pairs. This provides a strong indication of the practical applicability of the CIELAB color space. Third, general-purpose image quality models and just-noticeable difference measures fail to accurately predict CDs in photographic images. Finally, the proposed MS-SWD significantly surpasses all competing methods on the non-perfectly aligned pairs, and achieves the overall best performance in terms of PLCC and SRCC without training on perceptual CD data. This highlights the importance of non-local patch comparisons in CD assessment.

**Robustness Results to Geometric Transformations.** To further verify the robustness of MS-SWD to geometric transformations, we follow the experimental procedure in [9], and augment SPCD by 1) randomly shifting one image relative to the other by up to 5% pixels in both axes, 2) enlarging one image by a factor of 1.1, and 3) horizontally flipping one image (see the third row of Fig. 1). These transformations are applied to the non-perfectly aligned pairs in

**Table 2:** Performance evaluation of CD measures on the augmented SPCD dataset by geometric transformations. “Translation” involves randomly shifting one image relative to the other by up to 5% of pixels in both axes. “Dilation” refers to enlarging one image by a factor of 1.1. “Flipping” means horizontally flipping one image.

Method	Translation			Dilation			Flipping		
	STRESS↓	PLCC↑	SRCC↑	STRESS↓	PLCC↑	SRCC↑	STRESS↓	PLCC↑	SRCC↑
CIELAB [39]	38.271	0.386	0.304	38.667	0.361	0.260	42.956	0.168	0.094
CIE94 [32]	37.271	0.435	0.318	37.539	0.419	0.274	41.715	0.210	0.113
CIEDE2000 [29]	38.365	0.377	0.284	38.619	0.362	0.240	42.770	0.170	0.079
S-CIELAB [57]	39.048	0.349	0.253	39.262	0.332	0.216	42.960	0.151	0.065
Lee05 [25]	56.466	0.632	<b>0.633</b>	56.529	0.636	<b>0.637</b>	56.515	<b>0.636</b>	<b>0.637</b>
Hong06 [21]	57.521	0.297	0.206	55.609	0.284	0.177	56.718	0.154	0.098
LPIPS [56]	45.853	0.048	0.018	43.882	0.083	0.109	43.545	0.074	0.104
DISTS [16]	37.303	0.362	0.289	37.519	0.317	0.252	37.287	0.316	0.233
CD-Net [52]	29.737	0.659	0.567	29.848	0.656	0.542	39.325	0.295	0.221
CD-Flow [9]	<b>29.188</b>	<b>0.719</b>	0.569	<b>29.065</b>	<b>0.705</b>	0.584	<b>36.546</b>	0.393	0.263
MS-SWD (Ours)	<b>28.353</b>	<b>0.836</b>	<b>0.798</b>	<b>28.144</b>	<b>0.833</b>	<b>0.793</b>	<b>26.132</b>	<b>0.836</b>	<b>0.788</b>

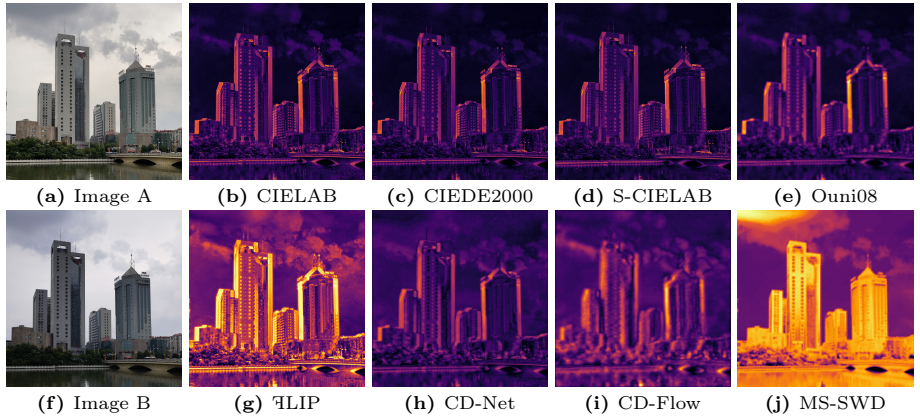
the SPCD dataset. From the results in Table 2, we find that all competing methods experience a significant performance drop, except for Lee05 which benefits from non-local CD assessment by histogram intersection. Although designed to be aware of geometric transformations, DISTS, CD-Net, and CD-Flow can only handle mild transformations. In stark contrast, the proposed MS-SWD is exceptionally robust to geometric transformations, even in the challenging case of horizontal flipping.

**Visualization of CD Maps.** We compare the CD maps generated by the proposed MS-SWD with seven other representative CD measures. Fig. 5 shows the visualization results for a non-perfectly aligned pair. It is evident that all competing methods are sensitive to image misalignment, leading to falsely large CDs along object boundaries. On the contrary, the proposed MS-SWD generates a more accurate CD map, correcting identifying areas of large CDs (*e.g.*, the clouds, buildings, and trees).

## 4.2 Ablation Studies

**Verification as an Empirical Metric.** We design computational experiments to verify that the proposed MS-SWD behaves empirically as a metric, which holds potential in perceptual optimization of color image processing algorithms. Non-negativity and symmetry are immediately apparent from Eq. (3). For the identity of indiscernibles (*i.e.*,  $\Delta E(\mathbf{X}, \mathbf{Y}) = 0 \iff \mathbf{X} = \mathbf{Y}$ ), we resort to the reference image recovery task [15] as a way of examining pixel-level image fidelity (see Fig. 4), where we find MS-SWD successfully recovers the reference image from all structured and non-structured initializations. For the triangle inequality (*i.e.*,  $\Delta E(\mathbf{X}, \mathbf{Y}) \leq \Delta E(\mathbf{X}, \mathbf{Z}) + \Delta E(\mathbf{Z}, \mathbf{Y})$ ), we test MS-SWD on 100,000 randomly selected image triplets of the same content from SPCD, and find no violations. In conclusion, we empirically establish that MS-SWD behaves as a metric in the mathematical sense.





**Fig. 5:** Comparison of CD Maps for a non-perfectly aligned pair, where a warmer color indicates a larger pixel-wise (or patch-wise) CD.

**Table 3:** Ablation analysis of the number of random linear unit projections in MS-SWD. The default setting is highlighted in boldface.

# of random projections	STRESS↓	PLCC↑	SRCC↑	Time (ms)
$P = 4$	31.849	0.804	0.779	3.7
$P = 16$	29.186	0.833	0.799	4.2
$P = 64$	28.425	0.841	0.805	6.2
$P = \mathbf{128}$	28.363	0.841	0.805	9.5
$P = 256$	28.318	0.842	0.806	15.3

**Number of Random Linear Unit Projections.** We investigate the effect of the number of random linear unit projections in MS-SWD, with  $P$  values selected from  $\{4, 16, 64, 128, 256\}$ . Table 3 shows the results on the non-perfectly aligned pairs from SPCD, where the average inference time is estimated using an NVIDIA A100 GPU. It is clear that the CD assessment performance of MS-SWD remains fairly stable when we decrease  $P$ , but excessively small  $P$  values will compromise the ability of MS-SWD to maintain pixel-level image fidelity. Therefore, we choose  $P = 128$  to balance prediction accuracy, metric property, and computational complexity.

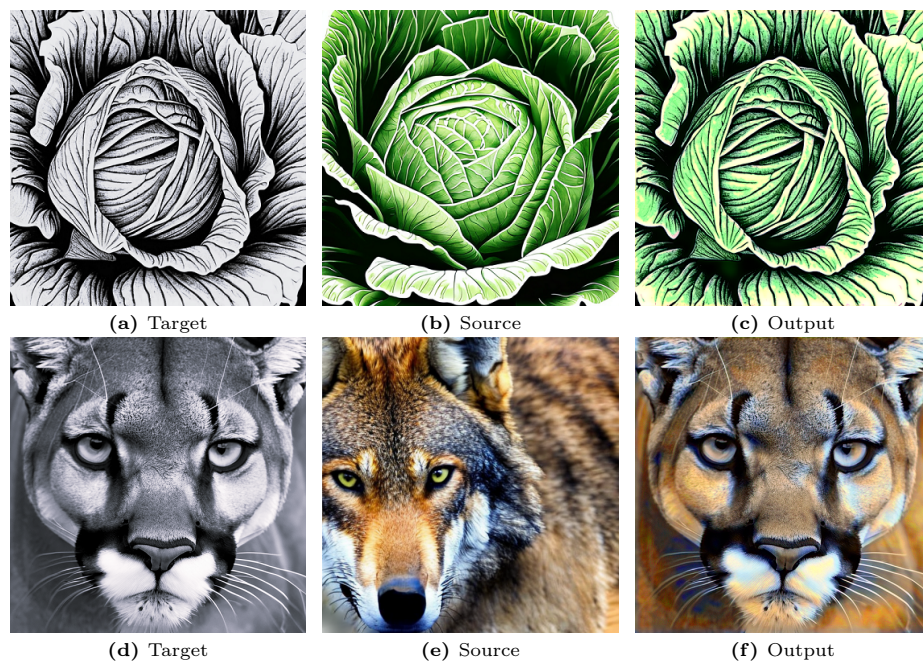
**Learnable Non-Linear Projections.** To further enhance MS-SWD, we explore replacing random linear unit projections with learnable non-linear projections. Inspired by CD-Net [52], we design a lightweight neural network for non-linear projection, including a front-end  $11 \times 11$  convolution layer and a back-end  $1 \times 1$  convolution layer with leaky ReLU in between. Training involves minimizing the PLCC loss using Adam optimizer, initialized with a learning rate of  $10^{-3}$  and decayed by a factor of 2 every 5 epochs. We train the network for 10 epochs using a mini-batch size of 30. We randomly partition SPCD into 70%, 10%, and 20% for training, validation, and testing, respectively, while ensuring content in-



**Table 4:** Ablation analysis of using learnable non-linear projections in place of random linear unit projections in MS-SWD.

Method	Perfectly aligned pairs			Non-perfectly aligned pairs			All		
	STRESS↓	PLCC↑	SRCC↑	STRESS↓	PLCC↑	SRCC↑	STRESS↓	PLCC↑	SRCC↑
CD-Net [52]	20.891	0.867	0.870	22.543	0.818	0.776	21.431	0.846	0.842
CD-Flow [9]	<b>16.613</b>	<b>0.896</b>	<b>0.904</b>	<b>21.374</b>	0.856	0.794	<b>18.473</b>	0.871	0.865
MS-SWD (Learned)	21.870	0.894	0.896	22.359	<b>0.876</b>	<b>0.857</b>	22.364	<b>0.884</b>	<b>0.889</b>

Trainable parameters: CD-Net (0.01M), CD-Flow (60.49M), and MS-SWD (0.05M).

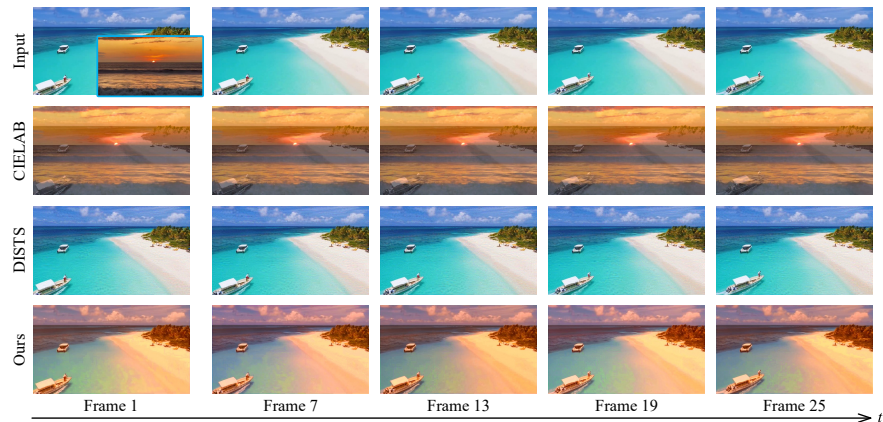
**Fig. 6:** Image color transfer results guided by MS-SWD.

dependence during dataset splitting. This procedure is repeated ten times, and the average results are reported. As shown in Table 4, our learned MS-SWD outperforms the most advanced CD-Flow with just 0.08% of its parameters.

### 4.3 Image and Video Color Transfer

In this subsection, we explore the application of the proposed MS-SWD in the image and video color transfer task. Our computational algorithm is straightforward: given a source color image (or video)  $\mathbf{X}$ , we aim to transfer its color appearance to the target grayscale image (or color video)  $\mathbf{Y}_{\text{init}}$  through the following optimization problem:

$$\mathbf{Y}^* = \arg \min_{\mathbf{Y}} \Delta E(\mathbf{X}, \mathbf{Y}), \quad (6)$$



**Fig. 7:** Comparison of video color transfer results. The first row displays five frames sampled from the target video, with the source image providing the desired color appearance shown in the bottom right corner of Frame 1.

starting from  $\mathbf{Y}_{\text{init}}$ . The results of image color transfer are shown in Fig. 6, demonstrating a successful color mapping from source to target, while preserving the underlying structure. The video color transfer outcomes are shown in Fig. 7. Using CIELAB, we transfer the color patterns as well as unwanted structure details to the target. DISTS fails in this task, often leaving the target largely unchanged. MS-SWD produces visually appealing video frames in terms of transferred color appearance, structure preservation, and temporal consistency.

## 5 Conclusion and Discussion

We have introduced MS-SWD, the multiscale sliced Wasserstein distance designed for measuring perceptual CDs in photographic images. Unlike traditional *co-located* comparisons prevalent in CD assessment, MS-SWD enables efficient comparisons between *non-local* patches of similar structure and color information, making it exceptionally robust to real-world image misalignment. MS-SWD is training-free using random linear unit projections, which can be replaced by learnable non-linear projections for improved performance.

We highlight the importance of multiscale analysis in MS-SWD for preserving pixel-level image fidelity, thereby demonstrating its empirical metric properties. An intriguing mathematical inquiry remains: whether MS-SWD is indeed a metric, given specific hyper-parameter configurations. Additionally, exploring alternative linear and non-linear image pyramids beyond the Gaussian pyramid, such as the (normalized) Laplacian pyramid [8], steerable pyramid [48], and VGG feature hierarchy [49], presents an interesting avenue. Last, there is potential to extend the non-local computation in MS-SWD (via the `sort()` operator) to measure other perceptual aspects of human vision (*e.g.*, image quality).

## Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (2023YFE0210700), the Hong Kong ITC Innovation and Technology Fund (9440390), the National Natural Science Foundation of China (62071407, 62375233, 62301323, 62441203, 62302423, and 62311530101), and the Shenzhen Natural Science Foundation (20231128191435002).

## References

1. Alakuijala, J., Obryk, R., Stoliarchuk, O., Szabadka, Z., Vandevenne, L., Wassenberg, J.: Guetzli: Perceptually guided JPEG encoder. arXiv preprint arXiv:1703.04421 (2017)
2. Albertazzi, L., Van Tonder, G.J., Vishwanath, D.: Perception Beyond Inference: The Information Content of Visual Processes. MIT Press (2011)
3. Andersson, P., Nilsson, J., Akenine-Möller, T., Oskarsson, M., Åström, K., Fairchild, M.D.: FLIP: A difference evaluator for alternating images. *ACM on Computer Graphics and Interactive Techniques* **3**(2), 1–23 (2020)
4. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning. pp. 214–223 (2017)
5. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics* **28**(3), 1–11 (2009)
6. Ben-Shahar, O., Zucker, S.W.: Hue geometry and horizontal connections. *Neural Networks* **17**(5-6), 753–771 (2004)
7. Bhardwaj, S., Fischer, I., Ballé, J., Chinen, T.: An unsupervised information-theoretic perceptual quality metric. In: International Conference on Neural Information Processing Systems. pp. 1–12 (2020)
8. Burt, P.J., Adelson, E.H.: The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications* **31**(4), 532–540 (1983)
9. Chen, H., Wang, Z., Yang, Y., Sun, Q., Ma, K.: Learning a deep color difference metric for photographic images. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 22242–22251 (2023)
10. Chou, C.H., Liu, K.C.: A fidelity metric for assessing visual quality of color images. In: International Conference on Computer Communications and Networks. pp. 1154–1159 (2007)
11. Choudhury, A., Wanat, R., Pytlarz, J., Daly, S.: Image quality evaluation for high dynamic range and wide color gamut applications using visual spatial processing of color differences. *Color Research & Application* **46**(1), 46–64 (2021)
12. Clarke, F.J.J., McDonald, R., Rigg, B.: Modification to the JPC79 colour-difference formula. *Journal of the Society of Dyers and Colourists* **100**(4), 128–132 (1984)
13. Deshpande, I., Zhang, Z., Schwing, A.: Generative modeling using the sliced Wasserstein distance. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3483–3491 (2018)
14. Ding, K., Liu, Y., Zou, X., Wang, S., Ma, K.: Locally adaptive structure and texture similarity for image quality assessment. In: ACM International Conference on Multimedia. pp. 2483–2491 (2021)

15. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision* **129**(4), 1258–1281 (2021)
16. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(5), 2567–2581 (2022)
17. Elnekave, A., Weiss, Y.: Generating natural images with direct patch distributions matching. In: *European Conference on Computer Vision*. pp. 544–560 (2022)
18. Fairchild, M.D.: *Color Appearance Models*. John Wiley & Sons, Ltd (2013)
19. Garcia, P.A., Huertas, R., Melgosa, M., Cui, G.: Measurement of the relationship between perceived and computed color differences. *Journal of the Optical Society of America A* **24**(7), 1823–1829 (2007)
20. Ghildyal, A., Liu, F.: Shift-tolerant perceptual similarity metric. In: *European Conference on Computer Vision*. pp. 91–107 (2022)
21. Hong, G., Luo, M.R.: New algorithm for calculating perceived colour difference of images. *The Imaging Science Journal* **54**(2), 86–91 (2006)
22. Kanizsa, G.: *Organization in Vision: Essays on Gestalt Perception*. Praeger Publishers (1979)
23. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: *International Conference on Learning Representations*. pp. 1–26 (2018)
24. Kolkin, N., Salavon, J., Shakhnarovich, G.: Style transfer by relaxed optimal transport and self-similarity. In: *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 10051–10060 (2019)
25. Lee, S.M., Xin, J.H., Westland, S.: Evaluation of image similarity by histogram intersection. *Color Research & Application* **30**(4), 265–274 (2005)
26. Lennie, P.: Color coding in the cortex. *Color Vision: From Genes to Perception* pp. 235–247 (1999)
27. Li, C., Li, Z., Wang, Z., Xu, Y., Luo, M.R., Cui, G., Melgosa, M., Pointer, M.: A revision of CIECAM02 and its CAT and UCS. In: *Color and Imaging Conference*. pp. 208–212 (2016)
28. Liao, X., Chen, B., Zhu, H., Wang, S., Zhou, M., Kwong, S.: DeepWSD: Projecting degradations in perceptual space to Wasserstein distance in deep feature space. In: *ACM International Conference on Multimedia*. pp. 970–978 (2022)
29. Luo, M.R., Cui, G., Rigg, B.: The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research & Application* **26**(5), 340–350 (2001)
30. Luo, M.R., Li, C.: CIECAM02 and its recent developments. *Advanced Color Image Processing and Analysis* pp. 19–58 (2013)
31. Mahy, M., Van Eycken, L., Oosterlinck, A.: Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV. *Color Research & Application* **19**(2), 105–121 (1994)
32. McDonald, R., Smith, K.J.: CIE94-A new colour-difference formula. *Journal of the Society of Dyers and Colourists* **111**(12), 376–379 (1995)
33. Moroney, N., Fairchild, M.D., Hunt, R.W., Li, C., Luo, M.R., Newman, T.: The CIECAM02 color appearance model. In: *Color and Imaging Conference*. pp. 23–27 (2002)
34. Ortiz-Jaramillo, B., Kumcu, A., Platasa, L., Philips, W.: Evaluation of color differences in natural scene color images. *Signal Processing: Image Communication* **71**, 128–137 (2019)

35. Ouni, S., Zagrouba, E., Chambah, M., Herbin, M.: A new spatial colour metric for perceptual comparison. In: International Conference on E-Systems Engineering, Communication and Information. pp. 413–428 (2008)
36. Pitie, F., Kokaram, A.C., Dahyot, R.: N-dimensional probability density function transfer and its application to color transfer. In: IEEE International Conference on Computer Vision. pp. 1434–1439 (2005)
37. Prashnani, E., Cai, H., Mostofi, Y., Sen, P.: PieAPP: Perceptual image-error assessment through pairwise preference. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1808–1817 (2018)
38. Rabin, J., Peyré, G., Delon, J., Bernot, M.: Wasserstein barycenter and its application to texture mixing. In: International Conference on Scale Space and Variational Methods in Computer Vision. pp. 435–446. (2012)
39. Robertson, A.R.: The CIE 1976 color-difference formulae. *Color Research & Application* **2**(1), 7–11 (1977)
40. Santos, C.N.d., Mroueh, Y., Padhi, I., Dognin, P.: Learning implicit generative models by matching perceptual features. In: IEEE International Conference on Computer Vision. pp. 4461–4470 (2019)
41. Shaham, T.R., Dekel, T., Michaeli, T.: SinGAN: Learning a generative model from a single natural image. In: IEEE International Conference on Computer Vision. pp. 4570–4580 (2019)
42. Shapley, R., Hawken, M.J.: Color in the cortex: Single-and double-opponent cells. *Vision Research* **51**(7), 701–717 (2011)
43. Sharma, G., Bala, R.: *Digital Color Imaging Handbook*. CRC Press (2017)
44. Shen, H.C., Wong, A.K.: Generalized texture representation and metric. *Computer Vision, Graphics, and Image Processing* **23**(2), 187–206 (1983)
45. Shevell, S.K., Kingdom, F.A.A.: Color in complex scenes. *Annual Review of Psychology* **59**, 143–166 (2008)
46. Shocher, A., Bagon, S., Isola, P., Irani, M.: InGAN: Capturing and retargeting the “DNA” of a natural image. In: IEEE International Conference on Computer Vision. pp. 4492–4501 (2019)
47. Simakov, D., Caspi, Y., Shechtman, E., Irani, M.: Summarizing visual data using bidirectional similarity. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2008)
48. Simoncelli, E.P., Freeman, W.T.: The steerable pyramid: A flexible architecture for multi-scale derivative computation. In: IEEE International Conference on Image Processing. pp. 444–447 (1995)
49. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations. pp. 1–14 (2015)
50. VQEG: Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment (2000), <http://www.vqeg.org>
51. Wang, Z., Xu, K., Ding, K., Jiang, Q., Zuo, Y., Ni, Z., Fang, Y.: CD-iNet: Deep invertible network for perceptual image color difference measurement. *International Journal of Computer Vision* (2024), to appear
52. Wang, Z., Xu, K., Yang, Y., Dong, J., Gu, S., Xu, L., Fang, Y., Ma, K.: Measuring perceptual color differences of smartphone photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(8), 10114–10128 (2023)
53. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004)

54. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Asilomar Conference on Signals, Systems & Computers. pp. 1398–1402 (2003)
55. Xu, K., Wang, Z., Yang, Y., Dong, J., Xu, L., Fang, Y., Ma, K.: A database of visual color differences of modern smartphone photography. In: IEEE International Conference on Image Processing. pp. 3758–3762 (2022)
56. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
57. Zhang, X., Wandell, B.A.: A spatial extension of CIELAB for digital color-image reproduction. *Journal of the Society for Information Display* **5**(1), 61–63 (1997)