Subpixel Image Quality Assessment Syncretizing Local Subpixel and Global Pixel Features

Jin Zeng, *Student Member, IEEE,* Lu Fang, *Member, IEEE,* Jiahao Pang, *Student Member, IEEE,* Houqiang Li, *Senior Member, IEEE,* Feng Wu, *Fellow, IEEE*

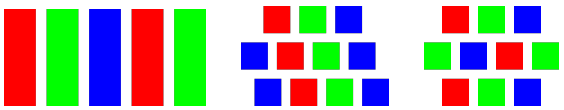
***Abstract*—Subpixel rendering technology increases the apparent resolution of the LCD/OLED screen by exploiting its physical properties that a pixel is composed of RGB individually addressable subpixels. Due to its intrinsic intercoordination between apparent luminance resolution and color fringing artifact, a common way of subpixel image assessment is subjective evaluation. In this paper, we propose a unified subpixel image quality assessment metric called SubPixel image Assessment (SPA) which syncretizes local subpixel and global pixel features. Specifically, comprehensive subjective studies are conducted to acquire data of user preferences. Accordingly, a collection of low-level features are designed under extensive perceptual validation, capturing subpixel features and pixel features which reflect local details and global distance with original image. With the features and their measurements as the basis, SPA is obtained which leads to a good representation of subpixel image characteristics. Experimental results justify the effectiveness and superiority of SPA. Besides, SPA is well adopted in different applications, including content adaptive sampling and metric-guided image compression.**

***Index Terms*—subpixel rendering, quality measurement, subpixel feature**

I. Introduction

A. Motivation

In patterned displays like color LCD/OLED, a pixel is composed of several color elements emitting the primary colors such as red, green and blue, which are called subpixels [1]. Fig. 1 shows various subpixel arragements on the displays where RGB stripe in Fig. 1(a) is the most typical one for computer monitors. By appropriately controlling subpixels, the apparent display resolution can be increased when rendering texts or images [2]. Various subpixel rendering algorithms for text or image display are springing up due to the market needs.

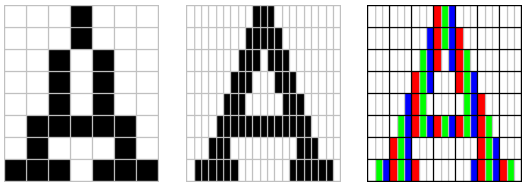


(a) (b) (c)

Fig. 1. Example of subpixel arrangements: (a) stripe; (b) diagonal; (c) triangular.

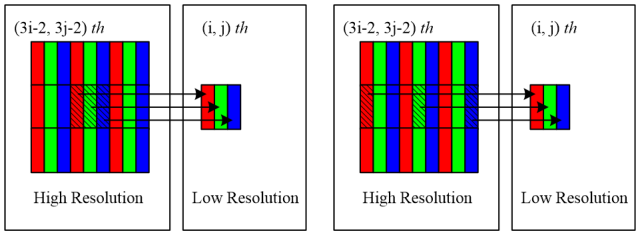
ClearType [3], a subpixel-based font display technology announced by Microsoft, is an example of how subpixel rendering works [1] for improving the readability of small text on LCD displays. In Fig. 2(a), when rendering the capital letter 'A' at pixel level, the edge appears blocky though it is obvious that a smooth edge should be presented. However when rendering at subpixel level, the edge appears much smoother and the shape is better preserved as can be seen in Fig. 2(b). Fig. 2(c) shows how subpixel rendering makes a difference. It borrows subpixels from adjacent pixels and increases the apparent horizontal resolution by three times.

Jin Zeng, Lu Fang and Jiahao Pang are with the Hong Kong University of Science and Technology (jzengab@connect.ust.hk; eefang@ust.hk; jpang@connect.ust.hk). Lu Fang, Houqiang Li and Feng Wu are with University of Science and Technology of China.



(a) (b) (c)

Fig. 2. (a) Pixel-based rendering with jaggy edge; (b) subpixel rendering with smooth edge; (c) subpixel structure.



(a) (b)

Fig. 3. (a) Direct Pixel-based Downsampling (DPD) pattern; (b) Direct Subpixel-based Downsampling (DSD) pattern.

Meanwhile, for image rendering, subpixel rendering provides users with images of higher sharpness and contrast. A simple example of how subpixel rendering works is shown in Fig. 3, where the original high resolution image is downsampled three times to fit the screen size, implying that one pixel is generated out of every 3×3 block. Direct Pixel-based Downsampling (DPD) [4] copies R, G, B values from ONE pixel in the 3×3 block, while Direct Subpixel-based Downsampling (DSD) [4] gets R, G, B values from THREE pixels with horizontal shift. Intuitively, more information is preserved in DSD image, while annoying color fringing artifacts may occur around sharp edges, since subpixels can not be treated as simply luminary element whose color can be ignored.

Therefore, the difficulty for subpixel-based image rendering is to balance luminance sharpness and color distortion [2], [5], and existing methods propose various anti-aliasing algorithms to address this problem. For example, Direct Subpixel-

(a) (b) (c)

Fig. 4. Magnified results of (a) PDAF with blurry edge; (b) DSDFA with sharp but color fringing artifact corrupted edge; (c) SDLCAF with color-free edge but less sharp than DSDFA.

based Downsampling with Anti-aliasing Filter (DSDFA) [4] proposes filter design based on frequency domain analysis, leading to sharp results though with small amount of color distortion. Fig. 4(b) shows the DSDFA result which is much sharper than the blurry result of pixel-based method, Pixel-based Downsampling with Anti-aliasing Filter (PDAF) [4] in Fig. 4(a), but the color fringing artifact along the grass is noticeable. Subpixel-based Downsampling with Luminance-Chrominance Anti-aliasing Filter (SDLCAF) [6] proposes different filters on Y, U, V channels to suppress color error and preserve high sharpness as depicted in Fig. 4(c) where color error is reduced but sharpness is unavoidably affected.

In light of this, it is hard to justify which method stands out of the others objectively. In other words, comprehensive assessment is necessary to measure the overall quality of subpixel-based rendered image (hereinafter called subpixel image). However, the work on the evaluation of subpixel image is far from enough due to several issues:

- Majority of existing evaluations are subjective-based [2], due to lack of ground truth subpixel images.
- Existing objective measurements merely account for limited features [7], which are far from comprehensive enough to cover all the characteristics of subpixel images.
- Given these one-sided assessments, none of unified metric exists, since the importance for each of them is not fully examined and justified.

Consequently, regarding the missing of systematic objective measurement for subpixel images, we design a mechanism for evaluating subpixel image rendering algorithms based on subjective study and metric estimation procedure. Specifically, our contributions are as follows:

- Unambiguous user preferences are obtained after extensive user study, guided by which, the visual quality of a subpixel image is decomposed into fundamental local subpixel features and global pixel features, each with a properly designed measurement.
- A unified metric called *SubPixel image Assessment* (SPA) is estimated by integrating individual features, which has been demonstrated to cover the attributes of subpixel images, and is comprehensive and consistent with human visual preference.
- SPA is effectively implemented in applications including content adaptive subpixel-based downsampling and subpixel image compression.

B. Related Works

1) Subpixel-Based Font Rendering: Subpixel rendering techniques arose from the problem of monochromatic font rendering on LCD/OLED displays, and various advanced font rendering techniques have been proposed. The Microsoft ClearType [3] uses the subpixel elements to triple horizontal spatial resolution and improve the readability of small text on LCD displays, though without further processing ClearType font induces color artifacts. Therefore, Platt *et al.* [8] adopted the S-CIELAB spatial filters [9] for designing spatio-chromatic filters to balance luminance resolution and color error.

Different from ClearType which is characterized by tight grid-fitting, Apple Quartz 2D [10] puts more emphasis on maintaining the shape of the typeface [11]. In addition, FreeType [12] is an open source software development library for font rendering addressing subpixel accuracy and is used by Android, iOS, Linux, *etc*. Besides, Anti-grain geometry [13] is another open source software library for 2D rendering, which provides text rendering service characterizing horizontal RGB subpixel anti-aliasing and vertical hinting.

2) Subpixel-Based Image Rendering: The simplest method is Direct Subpixel-based Downsampling (DSD) [4], directly copying R, G, B components from the original image without pre-filtering as shown in Fig. 3(b). Diagonal Direct Subpixel-based Downsampling (DDSD) [4] is similar though the downsampling pattern is in diagonal direction. Thus these two methods suffer from color fringing artifacts. For color error reduction, Kim *et al.* proposed in [14] to apply one directional anti-aliasing filter on the original image before downsampling, which degrades the sharpness though color error is reduced. Engelhardt *et al.* derived Multi-Sampled Anti-Aliasing filters (MSAA) [15] based on contrast sensitivity functions in [8]. In [7] and [4], Fang *et al.* designed anti-aliasing filters based on Minimizing Mean-Squared-Error (MMSE) and Frequency-domain Analysis (FA) respectively, leading to the methods called DSDMMSE/DDSDMMSE [7] and DSDFA/DDSDFA [4], characterizing high sharpness. Subpixel-based Downsampling with Luminance-Chrominance Anti-aliasing Filter (SDLCAF) [6] derived different filters for Y, U, V channels in order to suppress color error while maintaining sharpness.

3) Objective Measure for Subpixel Image Evaluation: Existing objective measures for subpixel image evaluation account for different features. For example, Luminance Sharpness Measure (LSM) [7] calculates the luminance high frequency energy of the subpixel image to reflect sharpness. Color Distortion Measure (CDM) [4] computes Peak Signal-to-Noise Ratio (PSNR) between original image and subpixel image in color channels to quantify color distortion. Luminance Aliasing Measure (LAM) [4] computes the discontinuity energy to measure aliasing artifact. These measures are well adopted in subpixel rendering algorithm evaluation [6], [16], [17] but only account for single feature respectively, failing to predict the overall image quality.

Moreover, since subpixel image is an approximation of original image, existing image quality evaluations that take the original image as a reference are applicable as well. The PSNR metric [18] measures squared error, though it may not match human visual perception. On the other hand, perception-based Structural Similarity Index (SSIM) [19] considers image

degradation as perceived variation in structural information. Another metric is S-CIELAB [9], which is a spatial extension of CIELAB [20] with an addition of a color separation and a spatial filter procedure. S-CIELAB takes into account the viewing conditions and specific features of the human visual system, and is well applied in filter design for subpixel-based text rendering [8], [21]. However, these metrics require the original image as a full reference, and do not consider subpixel-features, *e.g.* sharpness, contrast.

Different from existing evaluation methods, the proposed SPA is a unified objective measure which integrates comprehensive subpixel-image features whose weights are adjusted based on the user preference. In addition, SPA is of low computational complexity and well adopted in various applications.

The rest of the paper is organized as follows: in Section II, a crowdsourcing user study is conducted, incorporating nine subpixel-based and two pixel-based rendering algorithms, to find the importance of the principal features. Section III discusses the measurements designed to assess subpixel-image features. In Section IV, SPA is obtained as a mapping from the feature vector to image quality score that well matches user rating. The performance evaluation for each feature and SPA is illustrated as well. Applications of SPA are demonstrated in Section V. Lastly Section VI draws a conclusion of the work.

II. Subjective Study

Given that the subpixel image quality depends on multiple features, the feature weights need to be adjusted based on user preference when integrated into SPA so as to achieve a perceptually consistent metric. Therefore, we conduct a subjective study in order to: first, confirm the argument made in previous works that the subpixel image quality depends on sharpness, color artifacts, *etc.*; moreover, collect user preference data to tune the importance of each feature in the metric estimation stage. In the following context, the details of subjective study will be elaborated. First, the database setup is briefly explained, followed by the experimental design for obtaining pairwise comparison data. Next, unreliable users are detected and screened out based on subject-variability analysis. Lastly, the global score is inferred from the paired comparison data which shows consistency with previous works and serves as valid guideline for feature design.

A. Database Setup

The database consists of images with various situations, including indoor scenes, street views, buildings, human faces, animals, stationary objects, and some specially designed test images, *etc*. A total of 40 representative images are selected, and each one is processed with 11 image rendering algorithms, including 9 representative subpixel-based rendering algorithms, which are DSD [4], DDSD [4], DSDFA [4], DDSDFA [4], DSDMMSE [7], DDSDMMSE [7], SDLCAF [6], Kim [14], MSAA [15], and 2 pixel-based rendering algorithms, which are DPD [4] and PDAF [4].

To avoid resizing testing images on users' own screens, which may introduce extra distortion, we resize the images in advance to fit the screen resolution. According to the statistics of the browser screen resolution [22], 78% of online users are using screens with resolutions no smaller than 1366×768 while 1366×768 contributes 31%. Hence, the resulting images are set to be 800-pixel width, which is the normal size for images viewed on a 1366×768 browser[1]. *In sum, the database consists of 40 image groups, each containing 11 versions of the same original image.*

B. Pairwise Comparison Data Collection with Crowdsourcing

Given the database, there are various ways to rate the images. Considering the fact that giving absolute scores may lead to inconsistency among users, and giving an order of 11 items is of great difficulty, we employ *pairwise comparison* to ease the evaluation [23]. In addition, the pairwise comparison data is obtained with *crowdsourcing* [23], [24].

Specifically, we construct a website for image evaluation[2]. A pair of images are displayed at native resolution, and the user is asked to click on the image to toggle between the two images then decide the preferred one or choose the no-preference option if no difference is noticed. Each user is required to finish at least 5 sessions, and each session has 20 pairwise images that are randomly assigned. Therefore, only a fraction of the $\binom{11}{2} = 55$ pairs in the same image group is exposed to one user.

Each pair is compared by as least 20 users and only once will each pair be viewed by the same user. To ensure the user data quality, one *testing image pair* is inserted in every session, which is a comparison between a heavy-noise corrupted image and the result of state-of-art DDSDFA [4]. We reject the results from users who prefer noisy image to DDSDFA image for more than twice. At the end of data collection stage, we obtain data of 2,305 sessions from 292 users, among which, 7.8% users are rejected according to our rejecting criteria.

C. User Reliability Assessment with Within/Between-Subject Inconsistency

Crowdsourcing takes advantage of the Internet crowd so that researchers can conduct experiments with a more diverse set of users at a lower economic cost than traditional QoE (Quality of Experience) experiments under laboratory conditions. However, the disadvantage is that the users carry out experiments without supervision, and they may give inaccurate/dishonest results, thus many studies stress the necessity of filtering out unreliable users to ensure data quality [23]–[25].

Besides the rejecting mechanism in Section II-B, we also conduct user screening based on *subject-variability analysis*. In our experiment, each user views with only a small fraction of the image pairs, and for two different users, the number of pairs viewed by both is usually very small, so it is infeasible to directly compute subject variability based on preference judgements that different subjects made on the same set of

[1]Most of the aforementioned subpixel-based or pixel-based algorithms allow users to change resulting image size, and for those who don't, we resize the original image first using bicubic.

[2]The link is *http://imageevaluation.sinaapp.com/index.php*. Since the subjective study is over, the website is modified to show sample images via the link *http://imageevaluation.sinaapp.com/index.php?c=main&a=sample*.

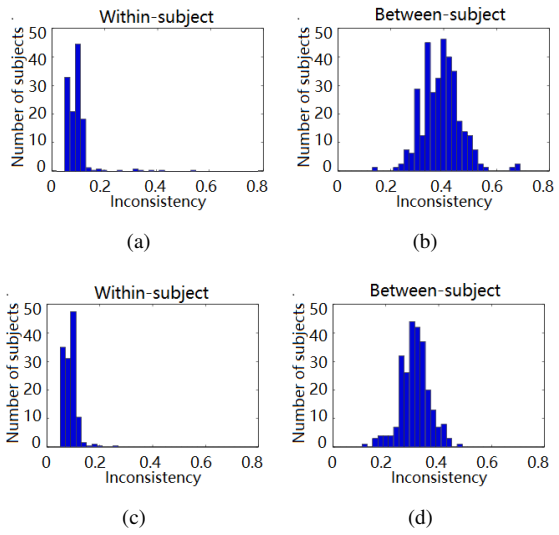


Fig. 5. Histogram of subject inconsistency: (a) within-subject before screening; (b) between-subject before screening; (c) within-subject after screening; (d) between-subject after screening.

images. Here we adopt the user reliability assessment method from [24] which fits our experimental setting.

After receiving paired comparison data, the within-subject and between-subject inconsistency are derived with HodgeRank [26], which respectively account for the inconsistency within the user's own comparison data, and the inconsistency between the user's data and the global average rating [24].

The histograms for within/between-subject inconsistency are shown in Fig. 5(a) and Fig. 5(b). As can be seen, some subjects have extraordinarily large inconsistency, *e.g.* with within-subject inconsistency larger than 0.3, or between-subject inconsistency larger than 0.6. A closer inspection at these subjects' records shows careless or random decisions. Therefore we screen out the subjects with top 5% within or between subject inconsistency, and recompute the inconsistency with "clean" data which gives the new inconsistency histograms in Fig. 5(c) and Fig. 5(d) where the outliers are greatly reduced, validating the effectiveness of user screening.

D. Global Score Estimation

Given the filtered paired comparison data, we need to rescale the local data to a global ranking score. To be specific, for each original image, $\{I_x\}_{x=1}^K$ are the set of results generated with K different image rendering algorithms where $K = 11$ in our case. For each image pair, the pairwise comparion data is represented by $(x, y, w_{xy}, t_{xy}, w_{yx})$ where $x, y \in \{1, \ldots, K\}$ are the image indices, w_{xy} is the number of users who prefer x to y, t_{xy} is the number of ties between x and y. To infer the global scores $\Lambda = \{\lambda_x\}_{x=1}^K$, Generalized Bradley-Terry Model with ties [27] is adopted, and the probability that one prefers x over y, and that one regards the two indifferent are

$$P(x > y) = \frac{\lambda_x}{\lambda_x + \theta\lambda_y}, \tag{1}$$

$$P(x = y) = \frac{(\theta^2 - 1)\lambda_x\lambda_y}{(\lambda_x + \theta\lambda_y)(\lambda_y + \theta\lambda_x)}, \tag{2}$$

where $\theta > 1$ is the threshold parameter. The notation "$x > y$" represents that x beats y, and "$x = y$" means x ties y. Then the likelihood of (λ_x, λ_y, θ) is,

$$\begin{aligned} L(\lambda_x, \lambda_y, \theta) &= P(w_{xy}, t_{xy}, w_{yx} | \lambda_x, \lambda_y, \theta) \\ &= \binom{w_{xy} + w_{yx} + t_{xy}}{w_{xy}} P(x > y)^{w_{xy}} \cdot \\ &\binom{w_{yx} + t_{xy}}{t_{xy}} P(x = y)^{t_{xy}} (1 - P(x > y) - P(x = y))^{w_{yx}}. \end{aligned} \tag{3}$$

Then, the likelihood of (Λ, θ) is

$$L(\Lambda, \theta) = \prod_{x=1}^{K} \prod_{y>x}^{K} P(w_{xy}, t_{xy}, w_{yx} | \lambda_x, \lambda_y, \theta). \tag{4}$$

By maximizing the likelihood with Expectation-Maximization (EM) algorithm [27], the global score Λ is obtained.

E. Discussion on Subjective Scores

By examining the average global score for each method over all 40 images, as depicted in Fig. 6, we arrive at the following observations:

- High sharpness is accompanied with color distortion in subpixel images. For example, the top three methods (DSDMMSE, DDSDFA, DDSDMMSE) have high sharpness and contrast but also small amount of color distortion as mentioned in [4], [7], but the subjective scores are still high, indicating that color fidelity can be sacrificed to some extent for luminance sharpness. This is consistent with the human visual system (HVS) perception, *i.e.*, HVS is less sensitive to high frequency chrominance error than to high frequency luminance error [28].
- If the image suffers from severe color distortions, even if sharpness and contrast are high, the subjective quality will still be low. For example, DSD and DDSD are methods without anti-aliasing filtering, therefore they have high sharpness but suffer obvious color distortions [4], leading to low subjective scores.
- For the image with small color error but low sharpness, *e.g.* Kim and pixel-based method PDAF, the image quality is still not appealing.

Thus the above observations are consistent with the arguments in existing works that there is a trade-off between luminance sharpness/contrast and color fidelity, which further implies that one single feature is far from enough to evaluate the overall subpixel image quality. This motivates us to design a collection of low-level features to cover the pros and cons of subpixel images as discussed in Section III.

F. Effect of Image Content

To see if the image content affects the judgement, for each image, we examine the rank order of the 11 methods in terms of their subjective scores, and find that DSD and DDSD have larger variations: for most images, DSD and DDSD (hereinafter combined and written as DSD/DDSD) have much lower scores than DSDMMSE and DDSDMMSE (hereinafter combined and written as DSD-/DDSDMMSE),

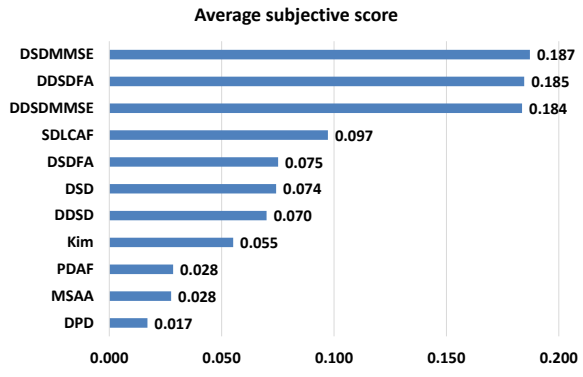


Fig. 6. Average subjective scores for different methods.

but for some, DSD/DDSD exhibit competitive results as DSD-/DDSDMMSE.

Thus, we divide the 40 original images in the database into three categories (Category-I contains 31 images, Category-II contains 8 images, and Category-III contains 1 image whose rank ordering is distinct from all the others), and the score curve for each image is plotted out in the corresponding subfigure in Fig. 7, where the horizontal and vertical axis are the method index and the subjective score respectively, and each curve corresponds to one image. Specifically, to categorize the images, we perform the following procedures: first, for each original image, we have a 11-dimensional score vector according to the subjective scores of the 11 methods; then for every two original images, we calculate the correlation coefficient [29] between their score vectors; next, we construct a full graph with each original image as the vertex, and correlation coefficient as the edge weight, and then apply spectral clustering algorithm in [30] to obtain the three categories.

From Fig. 7, we can see that Category-I and Category-II share similar score pattern for most methods, but the major difference is that in Category-II, method 1 (DDSD) and method 5 (DSD) achieve competitive scores as method 3 (DDSDMMSE) and method 7 (DSDMMSE) respectively, while in Category-I, DDSD and DSD are of relatively low scores. Category-III can be treated as an outlier because the results of method 8 (Kim), 9 (PDAF), and 10 (SDLCAF) are with highest scores. The Category-III image is shown in Fig. 8(d) which contains many regular textures like scales, parallel curves, *etc.*, so methods characterizing high sharpness exhibit obvious color error, while methods characterizing color error suppression (Kim, PDAF, SDLCAF) have higher scores, which is contrary to most images. Examining the image contents in Category-I and Category-II, we further find that:

- In Category-I, the difference between DSD/DDSD and DSD-/DDSDMMSE appears more obvious than in Category-II. Since DSD-/DDSDMMSE is the pre-filtered version of DSD/DDSD, suppressing the color error and enhancing the contrast, the image quality is usually enhanced thus scores of DSD-/DDSDMMSE images are higher than DSD/DDSD as shown in Fig. 7(a).
- However in Category-II the difference between DSD/DDSD and DSD-/DDSDMMSE is less obvious because Category-II typically exhibits: 1) fine and irregular texture, such as leaves, grass, field of flowers; 2) flat region, such as sky, wall; 3) night view. The corresponding representative images are shown in Fig. 8. With these contents, the filtering will have less obvious effect: 1) irregular texture already has high contrast, so the enhancement is not obvious; the color artifact could be large but does not appear apparent due to the complicated texture; 2) in case of flat region, subpixel rendering does not make a difference; 3) night view is too dark to exhibit the difference induced by filtering. Therefore, the scores of DSD/DDSD and DSD-/DDSDMMSE are similar due to the effect of these contents. Note that the images with irregular texture are the majority in Category-II while those with flat region and night view are not.

In sum, image content can affect the subjective score, and this also motivates us to include more detailed features to capture such effect so that SPA is capable of evaluating subpixel images of various contents.

III. Feature Design

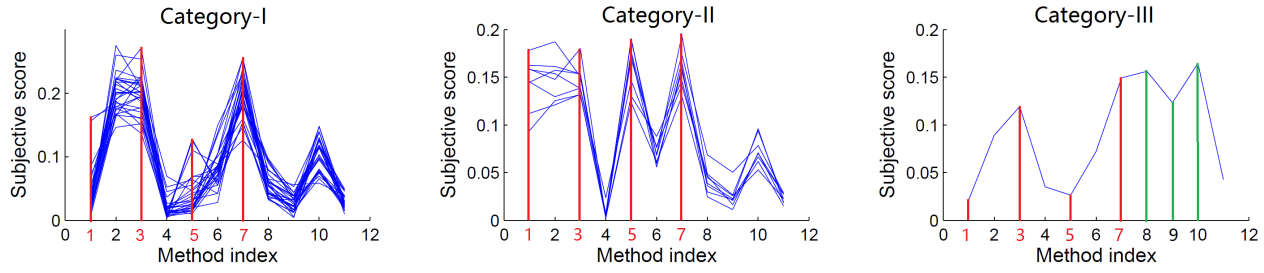
In this section, we design multiple low-level features for comprehensive description of subpixel images, which are quantified by the proposed measurements. The features are classified into two categories: *subpixel features* which capture the local details of subpixel images, referring to luminance sharpness and contrast, as well as the artifacts of color fringing, staircase, aliasing, and noise; *pixel features* that describe the global similarity between the original image and the resulting image generated by the rendering algorithm. The definitions of notations to be used in the following context are listed in Table I.

TABLE I
Definitions of Notations.

Notation	Definition
I_0	Original image of size $M \times N$
I_x	Subpixel image of size $m \times n$
(i, j)	Pixel index
b_l	The lth block of I_x when I_x is divided into L non-overlapping blocks, $l \in \{1, \ldots, L\}$
$u \times u$	Block size
Φ	Channel index for YCbCr space, $\Phi \in \{\mathrm{Y}, \mathrm{Cb}, \mathrm{Cr}\}$
Ω	Channel index for RGB space, $\Omega \in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}$
I_x^Φ, I_x^Ω	Φ or Ω channel of subpixel image I_x
b_l^Φ, b_l^Ω	Φ or Ω channel of image block b_l

A. Local Subpixel Features

Local subpixel features describe the attributes of the subpixel images, including the merits, *i.e.* high sharpness and contrast, and the artifacts and distortions as well, *i.e.* color



(a) (b) (c)

Fig. 7. Subjective score curves for images in: (a) Category-I; (b) Category-II; (c) Category-III. The method indices are: 1-DDSD, 2-DDSDFA, 3-DDSDMMSE, 4-DPD, 5-DSD, 6-DSDFA, 7-DSDMMSE, 8-Kim, 9-PDAF, 10-SDLCAF, 11-MSAA.

(a) (b) (c) (d)

Fig. 8. Representative images in Category-II with contents: (a) grass and trees with fine and irregular textures; (b) sky with flat region; (c) night view. Image in Category-III: (d) chart.

fringing, staircase, aliasing, and noise. The details and measurements for these features are elaborated as follows.

1) Sharpness: Subpixel images are visually appealing in terms of higher apparent luminance resolution, appearing much sharper than pixel-based rendered images. It implies that sharpness is one of the crucial features in describing subpixel details. Since subpixel images are distinguished from pixel-based images mainly in high frequency details, the luminance sharpness measure (LSM) [7] computes the average of the directional high-frequency energy in Y channel,

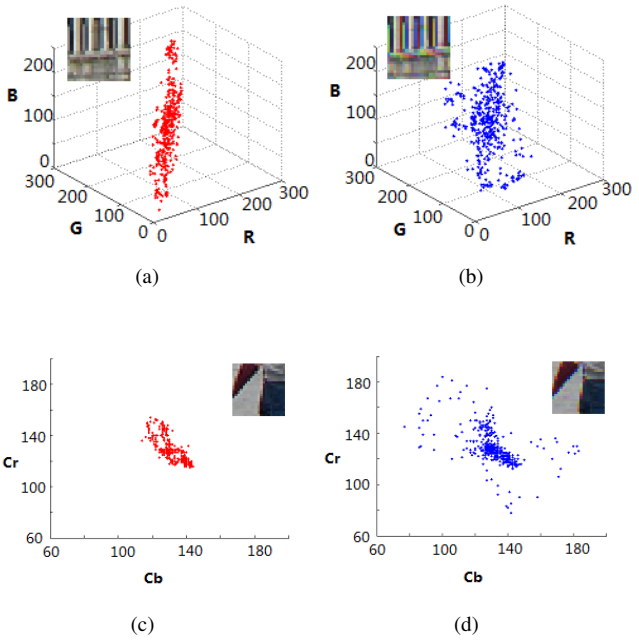
$$\mathrm{LSM}_a(I_x) = \frac{1}{mn}\sum_{d=1}^{4} \parallel H_a^d * I_x^{\mathrm{Y}} \parallel_2, \tag{5}$$

where H_a^d is the first order gradient filter with kernel a = [-1 1] in the dth direction, *i.e.*, horizontal ($d = 1$), vertical ($d = 2$), diagonal ($d = 3$), and anti-diagonal ($d = 4$). Similarly, the second order gradient filter b = [-1 2 1] is implemented to define LSM_b. From (5), we can see a larger value of LSM indicates more high-frequency energy, leading to higher apparent luminance sharpness.

2) Contrast: As noted in [4], for the case where two images have similar LSM values, the one with higher contrast provides better visual quality. The luminance contrast measure (LCM) is then characterized by the local variance of the image luminance, which is computed in the Y channel of each image block b_l,

$$\mathrm{LCM}(b_l) = \sqrt{\frac{1}{u^2}\sum_{i,j=1}^{u}(b_l^{\mathrm{Y}}(i,j)-\mu)^2}, \quad \mu = \frac{1}{u^2}\sum_{i,j=1}^{u} b_l^{\mathrm{Y}}(i,j), \tag{6}$$

where the block size u is chosen to be 30. $\mathrm{LCM}(I_x)$ is given by the average of $\mathrm{LCM}(b_l)$ over all blocks. Therefore, a higher LCM value suggests higher luminance contrast.



(a) (b) (c) (d)

Fig. 9. Image patch (magnified) and its distribution in RGB space: (a) PDAF result with no color error; (b) DSD result with abrupt color. Another image patch (magnified) and its distribution in CbCr space: (c) PDAF result with no color error; (d) DSD result with abrupt color.

3) Abrupt Color: High sharpness and contrast come along with the annoying color fringing artifacts in subpixel images. In particular, the artifacts usually happen around sharp edges, generating annoying colors. Such color distortion is the dominant color artifact in subpixel image, which is named as *abrupt color* here. Since accurate estimation of abrupt color remains unsolved, we design multiple measures that may work under different situations so as to better assess abrupt color. The measures are described below.

- *Abrupt color measure with color line model* ($\mathrm{ACM}_{\mathrm{CL}}$)

  While the color line assumption for natural color images claims that colors of a local region typically form a line in color space [31], the existence of abrupt color expands the color distribution, though the main orientation still preserves. As shown in Fig. 9(a), the result of PDAF algorithm contains no color noise, while the result of DSD algorithm in Fig. 9(b) exhibits severe color error. As further depicted by the color space distribution, the

PDAF patch presents stronger linearity, while the DSD one is expanded in distribution. Thus, the abrupt color can be measured by examining to which extent the color line is expanded. Specifically, for the local $u \times u$ image block b_l to be measured, let V_l be the vectorized version of b_l. Note that V_l is a $u^2 \times 3$ matrix, then singular value decomposition (SVD) is applied to V_l, *i.e.*,

$$V_l = U_l \Sigma_l W_l^{\mathrm{T}}, \tag{7}$$

where the singular values of V_l presented in decreasing order, $[\sigma_1, \sigma_2, \sigma_3]$, reveals the color distribution. For instance, if the colors are linearly distributed, σ_1 tends to be much larger than σ_2 and σ_3. However if the line is expanded, σ_2 will be relatively large. Thus, the average of the ratios of σ_2 to σ_1 over all the image blocks can be calculated to define $\mathrm{ACM_{CL}}$,

$$\mathrm{ACM_{CL}}(I_x) = \frac{1}{L}\sum_{l=1}^{L}\frac{\sigma_2(l)}{\sigma_1(l)}. \tag{8}$$

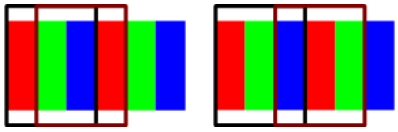
Here the block size u is set to be 8. Based on (8), $\mathrm{ACM_{CL}} \in [0,\ 1]$, and a higher value indicates more severe color distortion.

- *Abrupt color measure with k-medoids clustering* ($\mathrm{ACM_K}$)
For natural images, the number of dominant colors within a local region is usually no more than two [31]. The occurrence of abrupt color leads to more color clusters. For instance, Fig. 9(c) shows that the result of PDAF with no color error has relatively compact color distribution in (Cb, Cr) space, while the result of DSD corrupted by color noise has loose distribution as shown in Fig. 9(d). Such observation implies that if exotic dominant color—the dominant color induced by color fringing artifact and not intrinsic in the original image—is detected in a local region, or the color distribution distracts severely from the kernel, the probability of abrupt color increases.
Inspired by the observation, we first transform the image from RGB space to YCbCr space, and divide it into 8×8 blocks. Then for each block b_l, k-medoids clustering [32] is applied with Cb and Cr components as the two-dimensional input data. The cluster number is chosen to be three, since the number of dominant colors within a local region is usually no more than three even when abrupt color exists.
Let $(c_k^{\mathrm{Cb}},\ c_k^{\mathrm{Cr}})$ be the (Cb, Cr) value of the kth medoid where $k \in \{1, 2, 3\}$, and v_k be the number of pixels within the kth cluster, then $\mathrm{ACM_K}(b_l)$ is given by the variance of the medoids,

$$\mathrm{ACM_K}(b_l) = \frac{1}{2u^2}\sum_{\Phi\in\{\mathrm{Cb,Cr}\}}\sum_{k=1}^{3} v_k (c_k^{\Phi} - \mu^{\Phi})^2, \tag{9}$$

$$\mu^{\Phi} = \frac{1}{u^2}\sum_{k=1}^{3} v(k) c_k^{\Phi}. \tag{10}$$

Then, $\mathrm{ACM_K}(I_x)$ is calculated by averaging $\mathrm{ACM_K}(b_l)$ over all the blocks. According to (9) and (10), when abrupt color occurs, the variance of medoids increases, which leads to a larger value of $\mathrm{ACM_K}$.



(a) (b)

Fig. 10. Illustration of subpixel recombination: (a) SR_1, (b) SR_2, where black rectangle boxes out the current pixel, and the brown rectangle boxes out the recombined pixel.

- *Abrupt color measure with subpixel recombination* ($\mathrm{ACM_{SR}}$)
As shown in Fig. 10(a), the recombination of G, B subpixels in current pixel and R subpixel in the next pixel on the right will generate a new pixel. If varying far from the new pixel, the current pixel is highly likely to exhibit abrupt color since pixels with abrupt color are quite different from neighboring pixels. In this way, the abrupt color can be measured by calculating the amount of pixels prone to exhibit abrupt color, via generating the new pixels with *subpixel recombination*.
Apart from the recombination in Fig. 10(a), there can be various recombination patterns which are adjusted according to the subpixel layouts of displays. Without loss of generality, we implement two different patterns for RGB stripe display, with Fig. 10(a) denoted as SR_1, and Fig. 10(b) by recombining B subpixel of the current pixel with the R, G subpixels of the neighboring right pixel, denoted as SR_2. The recombination operation generates two new images, denoted as I_{SR_1} and I_{SR_2}, then the absolute difference between input image I_x and subpixel recombination images computed in (Cb, Cr) channels is denoted as E_{SR_1} and E_{SR_2} respectively,

$$E^{\Phi}_{\mathrm{SR_k}}(i,j) = |I_x^{\Phi}(i,j) - I^{\Phi}_{\mathrm{SR_k}}(i,j)|, \tag{11}$$

where $k = 1, 2$ is the index of the recombination pattern, and $\Phi \in \{\mathrm{Cb}, \mathrm{Cr}\}$. If the difference is larger than the threshold in either channel, the corresponding pixel is marked as an abrupt color corrupted pixel,

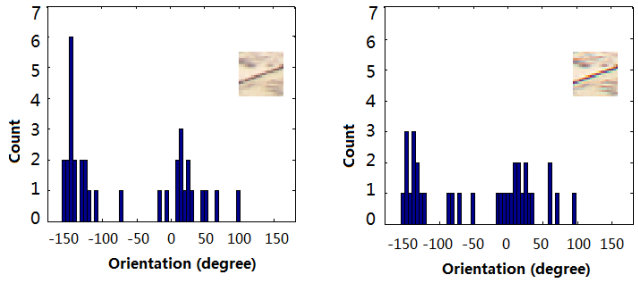
$$M^{\Phi}_{\mathrm{SR_k}}(i,j) = \begin{cases} 1, & \text{if } E^{\Phi}_{\mathrm{SR_k}}(i,j) \geq T^{\Phi}, \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

$$M_{\mathrm{SR_k}}(i,j) = M^{\mathrm{Cb}}_{\mathrm{SR_k}}(i,j) \vee M^{\mathrm{Cr}}_{\mathrm{SR_k}}(i,j), \tag{13}$$

where the threshold $(T^{\mathrm{Cb}},\ T^{\mathrm{Cr}})$ is set to be (20, 20). Then, the average of the absolute difference for the masked pixels leads to the value of $\mathrm{ACM_{SR}}$, and a higher value of $\mathrm{ACM_{SR}}$ indicates more severe color artifact,

$$\mathrm{ACM_{SR}}(I_x) = \frac{1}{2mn}\sum_{\Phi}\sum_{k=1}^{2}\sum_{i,j=1}^{m,n} E^{\Phi}_{\mathrm{SR_k}}(i,j) \times M_{\mathrm{SR_k}}(i,j). \tag{14}$$

4) Staircase Artifact: In addition to abrupt color artifact, staircase artifact is another visual distortion that undermines the virtue of subpixel images. The artifact usually happens at sloping edges, where the edge appears jagged thus inconsistent with the original structure. For example in Fig. 11(b), the result



(a) (b)

Fig. 11. Image patch (magnified) and the histogram of orientation for edge pixels: (a) Kim's result with negligible staircase artifact; (b) DDSDFA result degraded by staircase artifact.

of DDSDFA has a sharper edge but also more obvious staircase artifact at the sloping line, while in Fig. 11(a), the result of Kim is less sharp but the edge appears smooth.

To quantify the staircase artifact, we adopt subpixel recombination in Y channel to detect staircase artifact corrupted pixels, which gives *staircase artifact measure with subpixel recombination* ($\mathrm{SAM_{SR}}$). Similar to $\mathrm{ACM_{SR}}$, if the difference $E_{\mathrm{SR_k}}$ in Y channel is larger than the threshold, the corresponding pixel is marked out, and the average of the absolute difference for these marked pixels gives the value of $\mathrm{SAM_{SR}}$,

$$M^{\mathrm{Y}}_{\mathrm{SR_k}}(i,j) = \begin{cases} 1, & \text{if } E^{\mathrm{Y}}_{\mathrm{SR_k}}(i,j) \geq T^{\mathrm{Y}}, \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

$$\mathrm{SAM_{SR}}(I_x) = \frac{1}{mn} \sum_{k=1}^{2} \sum_{i,j=1}^{m,n} E^{\mathrm{Y}}_{\mathrm{SR_k}}(i,j) \times M^{\mathrm{Y}}_{\mathrm{SR_k}}(i,j), \tag{16}$$

where the threshold T^{Y} is set to be 20. Therefore a larger value of $\mathrm{SAM_{SR}}$ is associated with more staircase artifact. Nevertheless, $\mathrm{SAM_{SR}}$ captures not only the staircase, but also more general cases like noise, edges, *etc*. Thus for more accurate assessment, we introduce another measurement that more specifically describes the staircase structure.

When staircase artifact is presented, the gradient orientation of pixels along the edge will exhibit larger variance. For example, in Fig. 11 we show the Y channel orientation histograms of edge pixels in two image blocks: one from Kim with negligible staircase artifact, and the other one from DDSDFA with noticeable staircase artifacts. It is obvious that Fig. 11(a) has a distribution with smaller variance, while the other one in Fig. 11(b) owns much larger variance. Hence the variance is the clue that tells whether staircase artifact occurs or not, and we propose *staircase measure with gradient orientation variance* ($\mathrm{SAM_{GO}}$) which is calculated in local blocks,

$$\mathrm{SAM_{GO}}(b_l) = \frac{1}{|M_e|} \sum_{i \in M_e} (\theta(i) - \mu)^2, \quad \mu = \frac{1}{|M_e|} \sum_{i \in M_e} \theta(i), \tag{17}$$

where M_e is the set of edge pixels marked out via canny edge detector [33], and $\theta(i)$ is the gradient orientation at ith pixel in M_e. The block size u is chosen to be 8, and the staircase measure for the whole image is given by the average of $\mathrm{SAM_{GO}}(b_l)$ over all blocks.

5) Aliasing: To measure aliasing artifacts, we adopt Luminance Aliasing Measure (LAM) proposed in [4]. Aliasing usually appears as broken lines and sawtooth along edges, so LAM measures the *discontinuity energy*. Specifically, PDAF result is used as the reference because lines and edges in PDAF are continuous (not broken) [34]. First, LAM identifies edge pixels with PDAF result, and then in both the PDAF image and the input image I_x computes the square difference between each edge pixel and its neighbouring edge pixels, the global sum of which gives D_{PDAF} and D_x. The difference, $D_x - D_{\mathrm{PDAF}}$, gives the discontinuity energy because when broken line occurs, D_x will be larger than D_{PDAF} and the difference will increase. Therefore, larger LAM value indicates larger aliasing artifact.

6) Noise: As noted in [35], large derivatives are concentrated at a small number of pixels for natural images, leaving the majority of image pixels constant. In other words, gradient magnitudes are sparse if noise does not present. We thus denote the mean norm of gradient magnitudes, which we call $\mathrm{NM_{Sp}}$ to quantify the noise level,

$$\mathrm{NM_{Sp}}(I_x) = \sum_{\Omega \in \{R,G,B\}} \left(\frac{1}{mn} \sum_{i,j=1}^{m,n} |\nabla I^{\Omega}_x(i,j)|^p \right)^{1/p}, \tag{18}$$

where $\Omega = \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}$ represents the RGB channel, and $|\nabla I^{\Omega}_x(i,j)|$ is the gradient magnitude of Ω channel.

B. Global Pixel Features

Pixel features measure the global similarity between the original image and the subpixel image. Even though the original image is not a proper ground truth as it does not exhibit subpixel characteristics, it is still a valid reference for the subpixel image. However, the size of original image and subpixel images can be different due to rescaling, so many existing metrics, *e.g.* PSNR [18], SSIM [19], and S-CIELAB [9], fail to measure the similarity. In the following context, several *scale-invariant* global measurements are proposed to capture the pixel features.

1) YCbCr Histogram Distance: YCbCr histogram distance measures the Euclidean distance between the YCbCr component histograms of the original image and the subpixel image, given by

$$\mathrm{D^{\Phi}_{HIS}} = \sqrt{\sum_{b=1}^{B} (h^{\Phi}_0(b) - h^{\Phi}_x(b))^2}, \tag{19}$$

where $\Phi \in \{\mathrm{Y}, \mathrm{Cb}, \mathrm{Cr}\}$, so the measure contains three dimensions for Y, Cb, Cr perspectively. h_0 and h_x are the normalized histograms for original image and the subpixel image respectively, and $b = 1, \cdots, B$ is the index of bins. Accordingly, the measure reveals the similarity between the two images in YCbCr space distribution.

2) HOG Distance: Histogram of Orientated Gradients (HOG) [36] is locally normalized histogram of gradient orientation, which shows extreme high efficiency in object detection. We adopt it as a luminance feature and compute

D_{HOG}, which is the Euclidean distance between HOG of the original image and that of the subpixel image.

3) Frequency Domain Distribution Distance: Frequency domain distribution distance (D^{Φ}_{Freq}, $\Phi \in \{Y, Cb, Cr\}$) accounts for the distribution similarity in frequency domain. Specifically, the image is transformed into YCbCr color space, followed with Discrete-Time Fourier Transform (DTFT) and the computation of the histogram distribution for each channel. Similar to D_{HIS}, D_{Freq} computes the Euclidean distance between the histograms of original image and subpixel image for Y, Cb and Cr respectively.

C. Discussion on Feature Score

The abbreviations of the feature measures are summarized in Table II.

TABLE II
ABBREVIATIONS AND FULL NAMES OF FEATURE MEASURES.

Abbreviation	Full name
LSM	Luminance sharpness measure
LCM	Luminance contrast measure
ACM_{CL}	Abrupt color measure with color line model
ACM_{K}	Abrupt color measure with k-medoids clustering
ACM_{SR}	Abrupt color measure with subpixel recombination
SAM_{SR}	Staircase artifact measure with subpixel recombination
SAM_{GO}	Staircase artifact measure with gradient orientation
LAM	Luminance aliasing measure
NM_{Sp}	Noise measure with sparsity
D_{HIS}	YCbCr histogram distance
D_{HOG}	HOG distance
D_{Freq}	Frequency domain distribution distance

Equipped with the proposed features, we examine how these features evaluate different image rendering methods. Specifically for each feature, we compute the metric scores and find the rankings for different methods (listed in Table III), and arrive at the following observations:

- Sharpness (LSM) and contrast (LCM) measures are similar, where subpixel-based methods with luminance enhancement (*e.g.* DSD-/DDSDMMSE) have high scores, and pixel-based method without anti-aliasing filtering (*e.g.* DPD) also has high score; methods with blurry results (*e.g.* Kim, MSAA and PDAF) have low scores.
- Abrupt color measures (ACM_{CL}, ACM_{K}, ACM_{SR}) give good scores to pixel-based methods (DPD and PDAF), but poor scores for DDSD which is without anti-aliasing filtering.
- Staircase (SAM_{SR} and SAM_{GO}), aliasing (LAM) and noise measure (NM_{Sp}) give good scores to blurry results (Kim, MSAA and PDAF), and poor scores to DPD and DSD-/DDSDMMSE.
- For D_{HIS}, Y channel distance gives good score to DPD, but relatively poor scores to all the others. HOG distance (D_{HOG}) gives good result to DPD and DSD-/DDSDMMSE, but poor score to blurry results (Kim, MSAA, and PDAF), which is opposite to results of D_{Freq} in Y channel. In addition, results of D_{HIS} and D_{Freq} in Cb and Cr channels are similar to those of abrupt color measures.

In summary, the feature scores match our expectations. Further evaluation for each feature is conducted in Section IV-B.

D. Effect of Image Content

To see if image content affects the feature score, we follow similar procedures as in Section II-F, *i.e.*, for each feature, we compute the correlation coefficient between the scores of every two images, and the average values over all $\binom{40}{2}$ image pairs are shown in Table IV:

- For most features, different images exhibit similar score tendency leading to high correlations, except for the aliasing measure (LAM), HOG distance (D_{HOG}), and frequency domain distance in Y channel (D^{Y}_{Freq}).
- For LAM, large variation happens at DSD and DDSD, which are subpixel methods without color correction or contrast enhancement filters. A closer examination on images causing large variation shows that these images exhibit contents like fine and irregular texture, flat region, or night view, similar to those images in Category-II as mentioned in Section II-F. Therefore, the reason why these images have distinguished behavior is very likely to be that the filtering is not making an obvious difference between DSD/DDSD and the methods with filtering, so they tend to share similar LAM.
- However for D_{HOG} and D^{Y}_{Freq}, not only DSD/DDSD, but also PDAF and Kim whose results are the smoothest among all methods, have large variation, leading to a small correlation coefficient. In this case, it is less clear how image content affects the metric because the variation is too large and it is hard to classify the images. However, leaving out the images causing variation, we still observe a score tendency that is consistent with the average score rank shown in Table III.

Furthermore, to verify if the features capture the effects of image contents on the subjective judgements as discussed in Section II-F, we compare the feature scores for images in Category-I and II. In Table V and Table VI, selected feature scores for DSD/DDSD, DSD-/DDSDMMSE, and also their differences are listed for Category I and II. We can see that the luminance difference, *e.g.* contrast (LCM), staircase (SAM_{GO}), aliasing (LAM), for Category-II is mostly smaller than that for Category-I, though color difference (ACM_{CL}) is larger, which is consistent with the effect of irregular texture. Since night view and flat region images are not the majority type of images in Category-II, the features mainly account for characteristics of fine texture images.

In sum, the effect of image content on the rank order for different features as well as the subjective judgements, is consistent with our expectation and well captured by the features. Therefore, by integrating the features, SPA is capable of evaluating images with different contents.

IV. METRIC ESTIMATION

Given the user preference data and the feature collection, SPA is achieved as a mapping function from the feature

TABLE III
AVERAGE FEATURE SCORES AND THE ACCORDING RANKS FOR DIFFERENT METHODS.

Methods	LSM		LCM		ACM_{CL}		ACM_{K}		ACM_{SR}		SAM_{SR}		SAM_{GO}		LAM	
	score	rank	score	rank	score	rank	score	rank	score	rank	score	rank	score	rank	score	rank
DDSD	0.057	5	0.125	5	0.249	**11**	0.022	**10**	0.00764	**9**	0.014	8	0.147	7	0.010	7
DDSDFA	0.048	7	0.122	8	0.225	**9**	0.022	**9**	0.00690	8	0.012	4	0.145	5	0.006	4
DDSDMMSE	0.068	**3**	0.130	**2**	0.228	**10**	0.023	**11**	0.00892	**11**	0.016	**10**	0.152	**9**	0.013	**9**
DPD	0.068	**2**	0.128	**3**	0.089	**1**	0.011	**2**	0.00607	6	0.014	**9**	0.161	**11**	0.016	**11**
DSD	0.054	6	0.124	6	0.189	8	0.018	7	0.00662	7	0.012	6	0.152	8	0.009	6
DSDFA	0.044	8	0.123	7	0.161	5	0.016	6	0.00605	5	0.012	3	0.142	4	0.007	5
DSDMMSE	0.069	**1**	0.132	**1**	0.169	6	0.019	8	0.00865	**10**	0.016	**11**	0.154	**10**	0.016	**10**
Kim	0.042	**9**	0.121	**9**	0.157	4	0.017	5	0.00601	4	0.012	5	0.135	**2**	0.003	**3**
PDAF	0.034	**11**	0.118	**11**	0.010	**1**	0.003	**1**	0.00321	**1**	0.008	**1**	0.132	**1**	0.000	**1**
SDLCAF	0.057	4	0.126	4	0.095	**3**	0.012	**3**	0.00539	**3**	0.013	7	0.146	6	0.010	8
MSAA	0.037	**10**	0.119	**10**	0.06	4	0.005	**2**	0.00503	**2**	0.010	**2**	0.137	**3**	0.001	**2**

Methods	NM_{Sp}		D^{Y}_{HIS}		D^{Cb}_{HIS}		D^{Cr}_{HIS}		D_{HOG}		D^{Y}_{Freq}		D^{Cb}_{Freq}		D^{Cr}_{Freq}	
	score	rank	score	rank	score	rank	score	rank	score	rank	score	rank	score	rank	score	rank
DDSD	0.087	9	0.010	4	0.085	**10**	0.099	**10**	0.071	5	0.146	8	0.389	**11**	0.404	**11**
DDSDFA	0.081	7	0.009	**3**	0.080	**9**	0.094	**9**	0.071	6	0.135	5	0.375	**9**	0.380	**9**
DDSDMMSE	0.096	**10**	0.012	8	0.091	**11**	0.107	**11**	0.070	**3**	0.158	**9**	0.386	**10**	0.390	**10**
DPD	0.084	8	0.001	**1**	0.001	**1**	0.001	**1**	0.070	**1**	0.176	**11**	0.090	**1**	0.089	**1**
DSD	0.081	6	0.007	**2**	0.061	7	0.068	7	0.071	4	0.143	6	0.349	7	0.350	8
DSDFA	0.075	4	0.011	6	0.057	6	0.063	6	0.073	8	0.106	4	0.245	6	0.235	5
DSDMMSE	0.096	**11**	0.013	**10**	0.070	8	0.078	8	0.070	**2**	0.172	**10**	0.357	8	0.346	7
Kim	0.071	**3**	0.012	7	0.057	5	0.062	5	0.076	**11**	0.101	**1**	0.205	4	0.193	4
PDAF	0.062	**1**	0.014	**11**	0.018	**2**	0.012	**2**	0.075	**10**	0.102	**2**	0.170	**3**	0.181	**3**
SDLCAF	0.080	5	0.011	5	0.048	**3**	0.047	**3**	0.073	7	0.144	7	0.140	**2**	0.156	**2**
MSAA	0.068	**2**	0.013	**9**	0.053	4	0.058	4	0.073	**9**	0.106	**3**	0.244	5	0.243	6

TABLE IV
AVERAGE CORRELATION COEFFICIENT FOR DIFFERENT FEATURES.

Feature	LSM	LCM	ACM_{CL}	ACM_{K}	ACM_{SR}	SAM_{SR}	SAM_{GO}	LAM
Coefficient	0.931	0.916	0.944	0.962	0.921	0.921	0.839	**0.776**
Feature	NM_{Sp}	D^{Y}_{HIS}	D^{Cb}_{HIS}	D^{Cr}_{HIS}	D_{HOG}	D^{Y}_{Freq}	D^{Cb}_{Freq}	D^{Cr}_{Freq}
Coefficient	0.949	0.809	0.924	0.928	**0.608**	**0.402**	0.792	0.816

TABLE V
AVERAGE FEATURE SCORES OF DDSD, DDSDMMSE, AND THEIR DIFFERENCES OVER IMAGES IN CATEGORY-I (C1) AND CATEGORY-II (C2).

Feature	DDSD		DDSDMMSE		Difference	
	C1	C2	C1	C2	C1	C2
LCM	0.126	0.113	0.132	0.118	0.006	0.005
ACM_{CL}	0.233	0.285	0.219	0.249	-0.014	-0.035
SAM_{GO}	0.146	0.150	0.152	0.154	0.006	0.004
LAM($\times 10^{-3}$)	9.35	9.84	12.39	11.69	3.04	1.85

TABLE VI
AVERAGE FEATURE SCORES OF DSD, DSDMMSE, AND THEIR DIFFERENCES OVER IMAGES IN CATEGORY-I (C1) AND CATEGORY-II (C2).

Feature	DSD		DSDMMSE		Difference	
	C1	C2	C1	C2	C1	C2
LCM	0.126	0.113	0.135	0.121	0.009	0.008
ACM_{CL}	0.172	0.228	0.159	0.191	-0.013	-0.038
SAM_{GO}	0.150	0.155	0.154	0.156	0.004	0.001
LAM($\times 10^{-3}$)	8.73	8.33	15.38	13.75	6.65	5.42

space to a scalar score that well matches the user rating. In this section, we first evaluate the performance of each proposed feature measure in capturing its according feature. Also we show that single feature cannot predict the subpixel image quality comprehensively, which motivates us to pool the features into one unified metric with weight coefficients optimized with pairwise comparison data. Furthermore, the performance of the proposed SPA is evaluated with an independent testing dataset. For performance analysis on the feature measures and SPA, we begin with the discussion of the proposed evaluation method.

A. Position-Aware Rank Distance Metric

An intuitive way to evaluate the feature measures and SPA is adopting the idea of rank distance measurement to compare our score results with the ground truth score.

The Spearman's rank correlation [37] and Kendall τ distance [38] are among the most widely used ranking comparison methods. However the problem is, they do not account for the *position weight* and the *displacement*. Giving a correct order of those relatively good results is more meaningful than ranking those bad results correctly, so higher rank results should be assigned a higher position weight. Moreover, the displacement should be taken into account since larger punishment should be given when the metric mistakes in ranking the two images with bigger distance than with smaller distance.

In light of this, we propose *position-aware Kendall τ distance*, so that the rank comparison serves for our performance evaluation more accurately. Specifically, let s be the scores given by features or SPA, and σ be the ground truth. The

position weight is given by,

$$w(x) = \sigma(x) - \sigma_{min}, \tag{20}$$

where x is the image index, and σ_{min} is the minimum value in σ. Similar to the Kendall τ distance, we find the pairwise disagreements between two rankings, *i.e.*, $(\sigma(x) - \sigma(y))(s(x) - s(y)) < 0$, which gives the set of pairs $D_{(\sigma,s)}$. The proposed position-aware Kendall τ distance is defined as

$$K(\sigma, s) = \sum_{(x,y)\in D_{(\sigma,s)}} w(x)w(y)|\sigma(x) - \sigma(y)|, \tag{21}$$

where $w(x)w(y)$ accounts for the *position weight*, and $|\sigma(x) - \sigma(y)|$ takes in the *displacement punishment*. The result is normalized by $K(\sigma, -\sigma)$, which is the worst case by reversing the true score to achieve the maximum mismatch,

$$\overline{K(\sigma, s)} = \frac{K(\sigma, s)}{K(\sigma, -\sigma)}. \tag{22}$$

B. Single Feature Evaluation

As stated in Section III, each feature is dedicated to represent one characteristic of the subpixel image. To evaluate the performance of each feature using proposed position-aware rank distance metric, we specifically synthesize a few datasets, where each of them contains images with one particular artifact or attribute.

- *Blurring*: artificial motion blur kernels (45-degree motion in counter-clockwise direction) with different sizes are performed on original sharp images, leading to a series of blurry images.
- *Contrast adjustment*: color intensities of original images are adjusted by saturating certain percentage of the values at low and high intensities, so as to increase the contrast. As we change the percentage of values to be saturated, a set of images with different contrast are generated.
- *Color distortion*: subpixel recombination image is generated, containing color errors that are typical in subpixel images. The original image and recombination image are fused with different weighting factors, leading to different levels of color distortions.
- *Aliasing*: DPD images are generated to produce aliasing artifacts, which are further resized to original size using bicubic interpolation. Similar to color distortion group, the linear combination of original image and resized DPD image, produces a series of images with different level of aliasing artifacts.
- *Noise corruption*: Gaussian noise with different levels is artificially added to the original image. Specifically, we change the noise variance to create images with different noise level.

Sample images in the five synthetic datasets with certain attributes are shown in Fig. 12. For each dataset, there are 10 images with different levels of the corresponding attribute. In Fig. 12, we show a pair of images: one exhibits the lowest level of the attribute, while the other exhibits the highest level of the attribute. Since we artificially create these extra attributes, we give *true scores* to the synthesized images according to the level of attribute added. Then these images are processed by four different methods, DPD, PDAF, DDSD, and DDSDMMSE, which are the four typical methods of pixel-based and subpixel-based rendering. The resulting images share the same scores as the original synthesized images, which are compared to the feature given scores for feature evaluation.

(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Fig. 12. Testing images (patches) in synthetic datasets with attribute of: (a)(f) blurring; (b)(g) contrast adjustment; (c)(h) color distortion; (d)(i) aliasing; (e)(j) noise corruption. The first-row images exhibit the lowest level of the attributes, and the second-row images exhibit the highest level of the attributes.

Each feature is then evaluated on the corresponding dataset. Specifically, blurring dataset tests LSM; contrast adjustment dataset tests LCM; color distortion dataset tests abrupt color measures ($\mathrm{ACM_{CL}}$, $\mathrm{ACM_{K}}$, $\mathrm{ACM_{SR}}$); aliasing dataset tests $\mathrm{SAM_{SR}}$, $\mathrm{SAM_{GO}}$, LAM; noise corruption dataset tests $\mathrm{NM_{Sp}}$, and pixel features ($\mathrm{D_{HIS}}$, $\mathrm{D_{HOG}}$, $\mathrm{D_{Freq}}$).

The rank distance measured by the proposed metric in Section IV-A is shown in the second column of Table VII. It can be seen that the rank distance for all features are small, indicating the features well capture the attributes which they are designed for. In addition, we calculate the Pearson correlation coefficient [29] between the feature score and the subjective score, and the results are listed in the last column of Table VII. The correlations are very low, indicating that a feature alone cannot predict the image quality, motivating our approach of combining all features into a unified metric.

C. Metric Estimation with Logistic Regression

To derive SPA, we first define it as a mapping function from a feature space $\mathbf{f}$ to a scalar in $\mathbb{R}^+$. Feature space $\mathbf{f}$ is defined as a concatenation of the features defined earlier. We intend to fit the mapping function $s(\mathbf{f})$ using Logistic Regression [39] based on the user rating. Suppose we have a pair of images, I_x and I_y where $x, y \in \{1, \ldots, K\}$ are the image indices, with feature vectors $\mathbf{f}_x$ and $\mathbf{f}_y$, then according to [39], the probability that a user chooses x over y is,

$$P(x > y) = \frac{1}{1 + \exp(-\mathbf{c}^{\mathrm{T}}(\mathbf{f}_x - \mathbf{f}_y))}, \tag{23}$$

where $\mathbf{c}$ is the linear combination coefficient vector to be estimated with logistic regression.

As denoted in Section II-D, w_{xy} is the number of users preferring image x to y, w_{yx} is the one for preferring image y

TABLE VII
SINGLE FEATURE PERFORMANCE: 2ND COLUMN, RANK DISTANCE WITH SCORES OF SYNTHETIC IMAGES; 3RD COLUMN, CORRELATION COEFFICIENT WITH SUBJECTIVE SCORES.

Feature	Rank distance with synthetic score	Correlation coefficient with subjective score
LSM	0	0.567
LCM	0	0.614
$\mathrm{ACM_{CL}}$	3.99×10^{-5}	0.492
$\mathrm{ACM_{K}}$	0.118	0.640
$\mathrm{ACM_{SR}}$	0.020	0.731
$\mathrm{SAM_{SR}}$	0	0.665
$\mathrm{SAM_{GO}}$	0.051	0.361
LAM	0.014	0.421
$\mathrm{NM_{Sp}}$	0	0.718
$\mathrm{D^{Y}_{HIS}}$	0.056	0.248
$\mathrm{D^{Cb}_{HIS}}$	0.087	0.632
$\mathrm{D^{Cr}_{HIS}}$	0.028	0.655
$\mathrm{D_{HOG}}$	0.084	-0.329
$\mathrm{D^{Y}_{Freq}}$	3.59×10^{-4}	0.236
$\mathrm{D^{Cb}_{Freq}}$	0.007	0.547
$\mathrm{D^{Cr}_{Freq}}$	0.013	0.553

to x, and the tie number is $t_{xy} = t_{yx}$. To derive the likelihood of $\mathbf{c}$ using logistic model, the outcome is treated as a binomial distribution, and the tie number is equally divided into w_{xy} and w_{yx}, *i.e.*,

$$w'_{xy} = w_{xy} + t_{xy}/2, \quad w'_{yx} = w_{yx} + t_{xy}/2, \tag{24}$$

$$L(\mathbf{c}) = \prod_{(x,y)} \binom{w'_{xy} + w'_{yx}}{w'_{xy}} P(x > y)^{w'_{xy}} (1 - P(x > y))^{w'_{yx}}. \tag{25}$$

By maximizing the above likelihood with Iteratively Reweighted Least Squares (IRLS) algorithm [40], we obtain the parameter value $\mathbf{c}$ and the SPA score is given by,

$$s(\mathbf{f}) = \mathbf{c}^{\mathrm{T}}\mathbf{f}. \tag{26}$$

To avoid overfitting and redundancy, the reduction of feature dimension is conducted to find the suitable feature collection, under the guidance of k-fold cross validation [41]. Specifically, we randomly divide 40 image groups into 5 clusters, each containing 8 groups, where cluster 1-4 serve as training data, while cluster 5 works as testing data. By rotating the testing set, we repeat the ranking distance measurement for 5 times.

TABLE VIII
SCALED WEIGHT AND COMPUTATIONAL COMPLEXITY FOR SELECTED FEATURES.

Feature	Weight	Complexity
Abrupt color measure with subpixel recombination ($\mathrm{ACM_{SR}}$)	-5.097	$O(mn)$
Luminance contrast measure (LCM)	5.095	$O(mn)$
Staircase artifact measure with subpixel recombination ($\mathrm{SAM_{SR}}$)	4.714	$O(mn)$
Luminance aliasing measure (LAM)	-4.237	$O(\max\{mn, MN\})$
YCbCr histogram distance in Cb channel ($\mathrm{D^{Cb}_{HIS}}$)	2.327	$O(\max\{mn, MN\})$
Frequency domain distribution distance in Cb channel ($\mathrm{D^{Cb}_{Freq}}$)	0.976	$O(\max\{mn\log(mn), MN\log(MN)\})$

Table VIII depicts the optimal combination of the features. Each feature weight is rescaled with the standard deviation of the corresponding feature to reveal its importance, and listed in the decreasing order. As we can see, luminance contrast is a key factor to provide appealing visual quality, while abrupt color is the key artifact affecting the overall image quality. In addition, staircase artifact is also in the supporting role due to its high correlation with luminance contrast. Also, aliasing artifact is another artifact that reduce the image quality. Besides, two measurements for chrominance distance with original image are in the supporting roles though with small weights, indicating that color distortion will accompany with luminance contrast, and color fidelity can be moderately sacrificed for higher contrast. Therefore the results are consistent with our expectations. The linear combination of the selected features with their corresponding weights gives the proposed subpixel image quality assessment metric, SPA.

In addition, the feature computational complexity is listed in Table VIII, where M and N are the height and width of the original image, m and n are the height and width of the subpixel image. Since subpixel image rendering is usually applied to image *down-sampling* [2], we assume $M \geq m$, $N \geq n$. Therefore, our proposed SPA metric is of complexity $O(MN\log(MN))$. As for the runtime, take the image in Fig. 8(a) for example, on a laptop computer with Intel Core i7 CPU and 8 GB RAM, the MATLAB implementation takes about 9.5 seconds to compute score of its DDSDMMSE result, where the original image and subpixel image are of size 3264×4912 and 544×818, respectively. The runtime can be further reduced with parallel computing.

To visually illustrate SPA score, Fig. 13 shows the results of DDSD, DDSDFA, DDSDMMSE, DPD, PDAF with the corresponding SPA scores in increasing order. PDAF has relatively low score due to the low luminance contrast, DPD suffers from low quality as well due to the severe aliasing artifact, while in DDSD color errors are disturbing. DDSDMMSE and DDSDFA have similar results, though DDSDMMSE outperforms DDSDFA occasionally due to the higher contrast and less obvious color error.

In addition, Table IX shows the rank distance between subjective score (given by the *training data*) and metric score with the evaluation method proposed in Section IV-A. The result is compared with the three existing methods, *i.e.* PSNR, SSIM, and S-CIELAB[3]. To compute S-CIELAB metric, we assume viewing distance of 20 inches, and display of 100 ppi for spatial calibration, using sRGB color space for color calibration. In addition, the S-CIELAB value is given by the average of error map, and reversed to compare with subjective score because for S-CIELAB small value indicates better result. Based on Table IX, we can see SPA is more consistent with user preference than existing methods. This is because PSNR and SSIM simply rate the images based on similarity with original image while ignoring visual quality, and pixel-based methods (PDAF and DPD) are given high scores. Even if S-CIELAB considers viewing conditions and the perception

[3]Since PSNR, SSIM and S-CIELAB require that the image to be measured owns the same size as the reference, the resulting image is resized to be original size with bicubic method, for comparing with the original image.

(a) (b) (c) (d) (e)

Fig. 13. Cropped patch: (a) DPD, score 23.08; (b) PDAF, score 23.96; (c) DDSD, score 27.10; (d) DDSDFA, score 28.30; (e) DDSDMMSE, score 28.98.

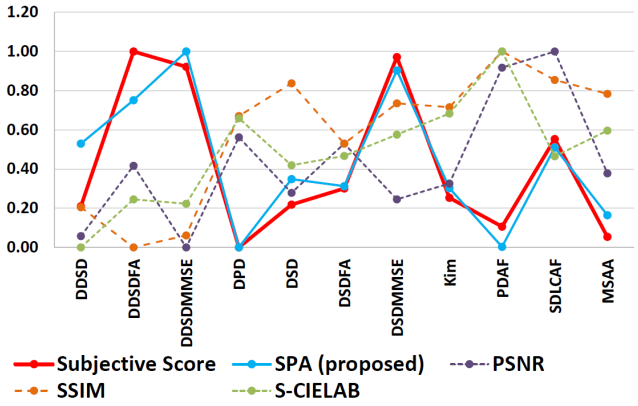


Fig. 14. Comparison of subjective user score and objective metric score.

of human visual system, it is still giving good results to pixel-based methods because it takes original image as the full reference. Most subpixel methods fail in the evaluation using PSNR, SSIM, S-CIELAB, which is opposite to the user preference.

TABLE IX
AVERAGE RANK DISTANCE BETWEEN SUBJECTIVE SCORE AND METRIC SCORE ON TRAINING DATASET AND TESTING DATASET.

Metric	PSNR	SSIM	S-CIELAB	SPA
Training data	0.706	0.626	0.681	0.069
Testing data	0.724	0.598	0.651	0.077

D. Metric Evaluation with Independent Testing Dataset

Apart from the evaluation on the training dataset, we include another group of 10 images for further validation of SPA. We obtain paired comparison data with the same approach as in Section II, and calculate global scores that represent users preference. To compare the objective metric score with the subjective user score, we rescale the scores to be within [0, 1] via $s' = (s - \min(s))/(\max(s) - \min(s))$. Note that S-CIELAB measures the error where a smaller value indicates better quality, we reverse and rescale its value via $s' = (\max(s) - s)/(\max(s) - \min(s))$.

We then plot subjective score and objective metric score in Fig. 14, and the rank distance between them is also listed in Table IX. Examining the plots in Fig. 14, we have following observations:

- The proposed SPA matches well with the subjective score and successfully distinguishes the methods with top scores, *e.g.* DDSDMMSE, DSDMMSE, DDSDFA.
- However, the other three objective metrics are inconsistent with user preference—they usually over-rate pixel-based methods (*e.g.* DPD and PDAF) with good scores, while under-rate subpixel-based methods (*e.g.* DDSDMMSE and DDSDFA) with relatively lower scores. This is mainly due to the fact that these three metrics do not take subpixel-features into consideration.

Also as shown in Table IX, SPA has a much lower rank distance than existing schemes. In light of this, the proposed SPA is superior over existing objective assessment methods in terms of better consistency with the user preference, due to the comprehensive measurements of both pixel and subpixel features.

V. APPLICATIONS OF PROPOSED METRIC

A. Content Adaptive Subpixel-Based Downsampling

When downsampling images with subpixel rendering algorithms, various sampling patterns are available that suit different image content, therefore in the following context, we demonstrate how the proposed SPA is applied in adaptively choosing sampling pattern to improve visual quality of the resulting image.

1) **Motivation:** Various subpixel-based sampling patterns have been proposed such as Direct Subpixel-based Downsampling (DSD), Diagonal Direct Subpixel-based Downsampling (DDSD), Anti-diagonal Direct Subpixel-based Downsampling (ADDSD) [7], and Subpixel-based Downsampling balancing Luminance and Chrominance (SDLC) [6]. Existing methods for subpixel-based image downsampling, *e.g.* DSD-/DDSDFA, DSD-/DDSDMMSE, adopt single sampling pattern. However as discussed in [5], single sampling pattern cannot handle all image content. For example, in regions with horizontal edges, DSD does not improve the apparent resolution because it merely samples in horizontal direction, which is parallel to the horizontal edges. Therefore, to achieve improved resolution in this case,

TABLE X
SUBJECTIVE SCORES OF CONTENT ADAPTIVE SUBPIXEL-BASED DOWNSAMPLING RESULTS GENERATED WITH DIFFERENT METRICS.

Image	PSNR	SSIM	S-CIELAB	SPA
Church	0.122	0.087	0.378	**0.413**
Museum	0.103	0.240	**0.344**	0.313
Fishmarket	0.275	0.109	0.307	**0.309**
Window	0.094	0.092	0.358	**0.456**
Island	0.111	0.024	**0.455**	0.410
Guell	0.121	0.088	**0.472**	0.319
Chart	0.339	0.114	0.207	**0.340**
Font	0.182	0.099	0.280	**0.439**
Average	0.168	0.107	0.350	**0.375**

DDSD is preferred by changing the sampling direction from horizontal to diagonal. To automatically select the appropriate sampling pattern, we adopt SPA for sampling pattern selection and this method is called *metric-guided Content Adaptive Subpixel-based Downsampling* (CASD).

2) **Procedure of the algorithm:** We first apply anti-aliasing filer derived in [7] and then downsample the image with four sampling patterns, resulting in four downsampled images, $\{I_x\}$ where $x \in \{\mathrm{DSD}, \mathrm{DDSD}, \mathrm{ADDSD}, \mathrm{SDLC}\}$ is the sampling pattern index, and the resulting images are divided into 8×8 blocks. Then for blocks at the same location, for example the l-th block in each resulting image, the scores for these blocks are calculated with SPA, and we find the one with highest score and choose the corresponding sampling pattern x_o as the optimal one for l-th block. The final result is achieved by fusing all the blocks generated via the locally selected sampling pattern.
3) **Evaluation:** To evaluate the performance of SPA, we also adopt other metrics for sampling pattern selection, *i.e.* PSNR, SSIM, and S-CIELAB, and conduct subjective study involving 20 participants. Each participant is asked to do pairwise comparison of the images generated with the four metrics. 8 testing images are included and the subjective scores calculated by Generalized Bradley-Terry Model with Ties [27] are listed in Table X. On average, SPA outperforms the other metrics. One subjective example is shown in Fig. 15. Results of PSNR and SSIM in Fig. 15(a)(b) are of relatively low sharpness, where the texture of the window in the red box appears less distinct. The S-CIELAB result in Fig. 15(c) is sharp but with noticeable color artifacts. On the contrary, SPA provides sharp result in Fig. 15(d) with less noticeable color error.

In addition, to see the advantage of CASD over existing methods with single sampling patterns, we add another subjective study involving 20 users. Each user is asked to compare the results of CASD with those generated by DSDMMSE and DDSDMMSE—highly rated methods in user study (Fig. 6). The same eight test images are used, and the percentage of users preferring CASD over DSDMMSE is $108/160 = 67.5\%$, and that over DDSDMMSE is $89/160 = 55.6\%$, so on average CASD performs better than existing methods with single sampling pattern. In Fig. 16, we demonstrate the visual results of original image, DSDMMSE, DDSDMMSE, and proposed CASD. Since the edges are in vertical direction as shown in original image Fig. 16(a), DSDMMSE in Fig. 16(b) and DDSDMMSE in Fig. 16(c) have noticeable color fringing artifact, *e.g.* in the image patches marked out by the red box, color error occurs along the vertical edges. Therefore in this scenario, CASD automatically chooses the sampling patterns that avoid color artifact as shown in Fig. 16(d).

B. Subpixel Image Compression

Existing lossy image codecs usually significantly compress color information, under the assumption that human eyes are much less sensitive to chrominance. However this coding strategy is destroyable for subpixel images, since the color information is completely coupled with the extra apparent luminance resolution. In other words, compressing subpixel images requires extra preservation of the color information, so as to maintain the subpixel attributes after compression.

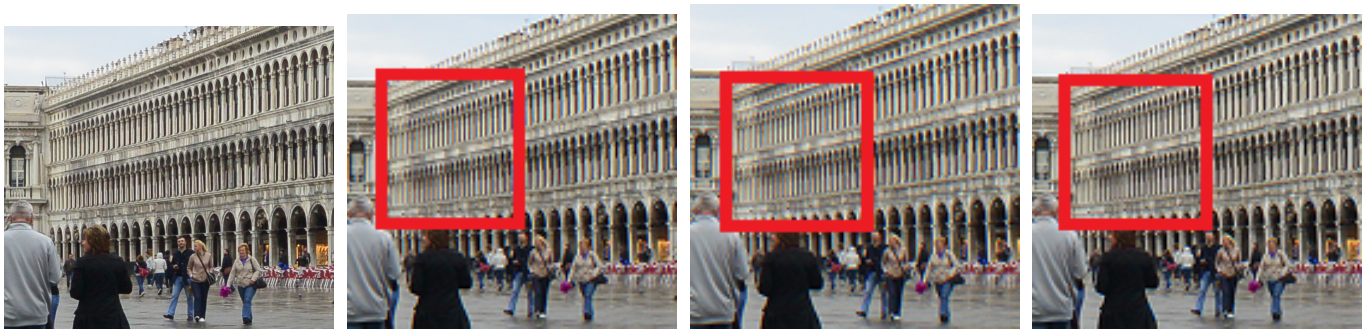
The widely used JPEG compression is integrated with SPA to investigate the *subpixel-metric-guided JPEG*. In general, the default sampling ratio for JPEG is 4:2:0, *i.e.*, Cb and Cr channels are downsampled by a factor of 2 both horizontally and vertically, which induces severe loss of high frequency details that are coupled with color information of subpixel images, making the merits of subpixel image forfeit. Instead of direct subsampling of Cb and Cr channels, we find the optimal compression quality settings for Y, Cb and Cr based on the SPA score. The procedure is as follows.

1) First of all, for a given quality value $q \in [1, 100]$, we compress the given subpixel image (generated with DSDMMSE) with JPEG default setting, *i.e.* Cb and Cr components are downsampled by a factor of 2.
2) The file size of default JPEG image, $b_{\max}$, serves as the *upper bound bitrate* for optimizing JPEG quality settings of Y, Cb and Cr. Specifically we generate a bunch of candidates, with a lower quality setting for Y, *i.e.* $q_y \leq q$, so as to allow room for tuning the quality setting for Cb and Cr (without downsampling), *i.e.* $q_c \in [1, 100]$. The tuning step is set to be 5 for q_y and q_c for computational complexity consideration. We calculate the score for each candidate with SPA, and choose the one with the highest score whose bitrate is under the upper bound $b_{\max}$. This gives the proposed method, named *metric-guided JPEG compression* of subpixel image.

For evaluation of the proposed metric-guided JPEG, we test the algorithm on 10 images and the quality values are chosen to be 95, 75, and 55 respectively. A user study is conducted to compare the default JPEG result with the metric-guided JPEG result. Each pair is evaluated by 20 users, so for a given quality value, 10 images will lead to 200 times of paired comparison. The percentage of paired comparisons in which the user prefers metric-guided JPEG is: $154/200 = 77\%$ for $q = 95$; $131/200 = 65.5\%$ for $q = 75$; $122/200 = 61\%$ for $q = 55$. On average, the metric-guided JPEG is preferred

(a) (b) (c) (d)

Fig. 15. Content adaptive subpixel-based downsampling results generated with different metrics: (a) PSNR; (b) SSIM; (c) S-CIELAB; (d) SPA.



(a) (b) (c) (d)

Fig. 16. Selected image patch in: (a) original image; (b) DSDMMSE result; (c) DDSDMMSE result; (d) proposed CASD.

over default JPEG for subpixel images, and for higher quality value, the preference rate is higher. When the quality value is low, the bitrate is too low to leave room for quality parameter tuning because reducing bitrate for Y channel will introduce ringing artifacts that affect the overall image quality severely.

Visual comparisons between default JPEG and metric-guided JPEG images are shown in Fig. 17. We can see that by transferring some amount of the luminance channel bitrate to the chrominance channel, the color information of the original image is preserved more vividly especially in the part of red steel frames (highlighted by the blue boundary), and the metric-guided JPEG has higher contrast and sharper edges due to the support of color information.

VI. Conclusion

The paper demonstrates the design of a perceptually-validated half-reference metric called SPA for evaluating the quality of subpixel-based rendered images. Extensive user studies are conducted to collect users' evaluations/preferences on the visual quality of subpixel images. Meanwhile the characteristics of subpixel images are decomposed into several fundamental local subpixel features and global pixel features, with properly designed metric for each of them. Finally SPA is obtained as a mapping from the feature space to the scalar score that matches the users' evaluation. The performance of SPA is evaluated, and the applications incorporating SPA further justify the effectiveness of the proposed work.

References

[1] S.Gibson, "Sub-pixel font rendering technology," 2010-05-28. Retrieved: 2014-05-01. [Online]. Available: https://www.grc.com/ct/ctwhat.htm

[2] M. A. Klompenhouwer, G. Haan, and R. A. Beuker, "13.4: Subpixel image scaling for color matrix displays," in *SID Symposium Digest of Technical Papers*, vol. 33, no. 1, 2002, pp. 176–179.

[3] Microsoft, "Cleartype information," 2010. Retrieved: 2014-05-10. [Online]. Available: http://research.microsoft.com/en-us/projects/cleartype

[4] L. Fang, O. C. Au, K. Tang, and A. K. Katsaggelos, "Antialiasing filter design for subpixel downsampling via frequency-domain analysis," *Image Processing, IEEE Transactions on*, vol. 21, no. 3, pp. 1391–1405, 2012.

[5] L. Fang, O. C. Au, K. Tang, and X. Wen, "Increasing image resolution on portable displays by subpixel rendering–a systematic overview," *APSIPA Transactions on Signal and Information Processing*, vol. 1, p. e1, 2012.

[6] J. Zeng, O. C. Au, Y. Guo, J. Pang, K. Tang, and Y. Ling, "Analysis of sampling pattern and luma-chroma filter design for subpixel-based image downsampling," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2014, pp. 5834–5838.

[7] L. Fang, O. C. Au, K. Tang, X. Wen, and H. Wang, "Novel 2-d mmse subpixel-based image down-sampling," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 5, pp. 740–753, 2012.

[8] J. C. Platt, "Optimal filtering for patterned displays," *IEEE Signal Processing Letters*, vol. 7, no. 7, pp. 179–181, 2000.

[9] X. Zhang and B. A. Wandell, "A spatial extension of cielab for digital color-image reproduction," *Journal of the Society for Information Display*, vol. 5, no. 1, pp. 61–63, 1997.

[10] Apple, "Quartz 2d information," 2014. Retrieved: 2016-07-17. [Online]. Available: https://developer.apple.com/library/mac/documentation/GraphicsImaging/Conceptual/drawingwithquartz2d

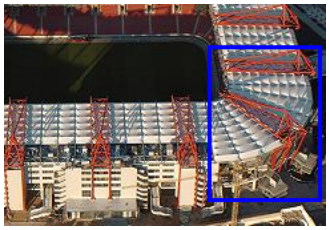
[11] T. Giannattasio, "The ails of typographic anti-aliasing," 2009. Retrieved: 2016-07-17. [Online]. Available: https://www.smashingmagazine.com/2009/11/the-ails-of-typographic-anti-aliasing

[12] D. Turner, "The freetype project," 2016. Retrieved: 2016-07-17. [Online]. Available: https://www.freetype.org

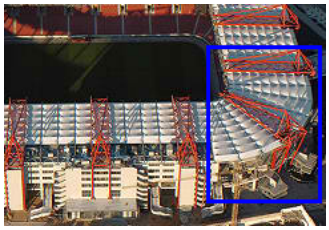
(a) Original image

(b) 2.254 bpp, $q = 95$

(c) 2.196 bpp, $q_y = 90$, $q_c = 80$

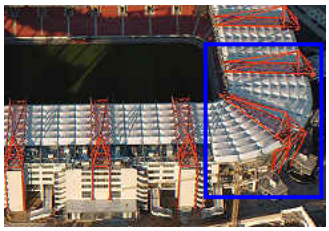


(d) 1.033 bpp, $q = 75$



(e) 1.038 bpp, $q_y = 70$, $q_c = 35$

(f) 0.733 bpp, $q = 55$



(g) 0.734 bpp, $q_y = 40$, $q_c = 35$

Fig. 17. Comparison between default JPEG image and metric-guided JPEG image: (a) original image; (b)(d)(f) default JPEG results with quality value q; (c)(e)(g) metric-guided JPEG results with quality value q_y for Y channel and q_c for Cb and Cr channels.

[13] M. Shemanarev, "The agg project: Texts rasterization exposureas," 2007. Retrieved: 2016-07-17. [Online]. Available: http://www.antigrain.com/research/font_rasterization/

[14] J.-S. Kim and C.-S. Kim, "A filter design algorithm for subpixel rendering on matrix displays," in *Proc. 15th European Signal Processing Conf (EUSIPCO)*, 2007, pp. 1487–1491.

[15] T. Engelhardt, T.-W. Schmidt, J. Kautz, and C. Dachsbacher, "Low-cost subpixel rendering for diverse displays," in *Computer Graphics Forum*, vol. 33, no. 1, 2014, pp. 199–209.

[16] J. Pang, L. Fang, J. Zeng, Y. Guo, and K. Tang, "Subpixel-based image scaling for grid-like subpixel arrangements: A generalized continuous-domain analysis model," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1017–1032, 2016.

[17] K. Tang, O. C. Au, L. Fang, J. Pang, Y. Guo, and J. Li, "Fast algorithm of arbitrary factor subpixel downsampling based on frequency analysis," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2014, pp. 1–6.

[18] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[20] Y.-K. Lee, "Comparison of cielab δe* and ciede2000 color-differences after polymerization and thermocycling of resin composites," *Dental Materials*, vol. 21, no. 7, pp. 678–682, 2005.

[21] J. Farrell, S. Eldar, K. Larson, T. Matskewich, and B. Wandell, "Optimizing subpixel rendering using a perceptual metric," *Journal of the Society for Information Display*, vol. 19, no. 8, pp. 513–519, 2011.

[22] W3schools, "Browser display statistics," 2014. Retrieved: 2014-08-01. [Online]. Available: http://www.w3schools.com/browsers

[23] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "A crowdsourceable qoe evaluation framework for multimedia content," in *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009, pp. 491–500.

[24] Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, and Y. Yao, "Hodgerank on random graphs for subjective video quality assessment," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 844–857, 2012.

[25] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.

[26] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, "Statistical ranking and combinatorial hodge theory," *Mathematical Programming*, vol. 127, no. 1, pp. 203–244, 2011.

[27] F. Caron and A. Doucet, "Efficient bayesian inference for generalized bradley–terry models," *Journal of Computational and Graphical Statistics*, vol. 21, no. 1, pp. 174–196, 2012.

[28] S. Daly, "47.3: Analysis of subtriad addressing algorithms by visual system models," in *SID Symposium Digest of Technical Papers*, vol. 32, no. 1. Wiley Online Library, 2001, pp. 1200–1203.

[29] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[30] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[31] I. Omer and M. Werman, "Color lines: Image specific color representation," in *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on*, vol. 2, no. 2, 2004, pp. 946–953.

[32] L. Kaufman and P. Rousseeuw, *Clustering by means of medoids*. North-Holland, 1987.

[33] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[34] R. C. Gonzalez and R. E. Woods, "Digital image processing," *Nueva Jersey*, 2008.

[35] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," in *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, 2007, p. 70.

[36] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on*, vol. 1, 2005, pp. 886–893.

[37] C. Spearman, "The proof and measurement of association between two things," *The American journal of psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[38] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, pp. 81–93, 1938.

[39] A. J. Dobson and A. Barnett, *An introduction to generalized linear models*. CRC press, 2008.

[40] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Communications in Statistics-theory and Methods*, vol. 6, no. 9, pp. 813–827, 1977.

[41] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2, 1995, pp. 1137–1145.