IEEE International Conference on Computer Vision (ICCV) 2019

# Scoot: A Perceptual Metric for Facial Sketches

Deng-Ping Fan[1,2] ShengChuan Zhang[4] Yu-Huan Wu [1] Yun Liu [1]
Ming-Ming Cheng[1,*] Bo Ren[1] Paul L. Rosin[3] Rongrong Ji[4,5]
[1] TKLNDST, CS, Nankai University [2] Inception Institute of Artificial Intelligence (IIAI) [3] Cardiff University
[4] Department of Artificial Intelligence, School of Informatics, Xiamen University [5] Peng Cheng Lab
http://mmcheng.net/scoot/

## Abstract

*While it is trivial for humans to quickly assess the perceptual similarity between two images, the underlying mechanism is thought to be quite complex. Despite this, the most widely adopted perceptual metrics today, such as SSIM and FSIM, are simple, shallow functions, and fail to consider many factors of human perception. Recently, the facial modelling community has observed that the inclusion of both structure and texture has a significant positive benefit for face sketch synthesis (FSS). But how perceptual are these so-called "perceptual features"? Which elements are critical for their success? In this paper, we design a perceptual metric, called Structure Co-Occurrence Texture (**Scoot**), which simultaneously considers the block-level spatial structure and co-occurrence texture statistics. To test the quality of metrics, we propose three novel meta-measures based on various reliable properties. Extensive experiments verify that our Scoot metric exceeds the performance of prior work. Besides, we built the first large scale (152k judgments) human-perception-based sketch database that can evaluate how well a metric is consistent with human perception. Our results suggest that "spatial structure" and "co-occurrence texture" are two generally applicable perceptual features in face sketch synthesis.*

## 1. Introduction

The ability to compare data items is known to be a fundamental operation for all of the computing [80, 91], especially in the computer vision area [5, 8, 89]. For various end-user applications such as face sketch [49], image style transfer [27], image quality assessment [66], saliency detection [11–13, 90], segmentation [41–43] and disease classification [73], image denoising [71], the comparison can turn out to be evaluating a "perceptual distance", which assesses how similar two images are in a way that highly correlates with *human perception*.

*M.M. Cheng (cmm@nankai.edu.cn) is the corresponding author.

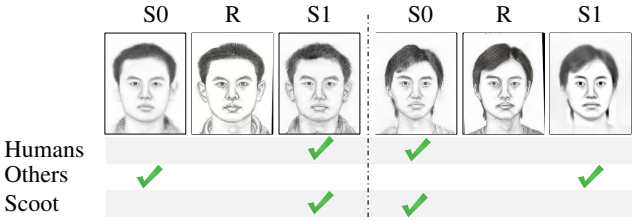



Figure 1: **Which sketch (left or right) is "closer" to the middle sketch in these examples?** For the right case, sketch 0 (S0) is more similar than sketch 1 (S1) *w.r.t.* reference (R) in terms of structure and texture. Sketch 1 almost completely destroys the texture of the hair. The widely-used (SSIM [66], FSIM [76]), classic (IFC [40], VIF [39]) and recently released (GMSD [72]) metrics disagree with humans. Only our Scoot metric agrees well with humans.

In this paper, we study facial sketch and show that human judgments are often different from current evaluation metrics, and as the first related attempt, we provide a novel perceptual distance for sketch according to human choice principles. As noticed in [80], human judgments of similarity depend on high-order image structure. Facial sketches are made up of a lot of textures, and there are many algorithms for synthesizing sketches, which is a good fit for this problem. However, designing a good perceptual metric should take into account human perception in facial sketch comparison, which should:

- closely match **human perception** so that good sketches can be directly used in various subjective applications, *e.g.*, law enforcement and entertainment.
- be **insensitive to slight mismatches** (*i.e.*, re-size, rotation) since real-world sketches drawn by artists do not precisely match each pixel to the original photos.
- be capable of **capturing holistic content**, that is, prefer the complete sketch to one that only contains strokes (*i.e.*, has lost some facial components).

To the best of our knowledge, no prior metric can satisfy all these properties simultaneously.

| No. | Model | Year'Pub | Sj. | Rr. | Ob. | No. | Model | Year'Pub | Sj. | Rr. | Ob. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **ST** [49] | 03'ICCV | | VRR | | 2 | **STM** [50] | 04'TCSVT | | VRR | |
| 3 | **LLE** [31] | 05'CVPR | | VRR | | 4 | **BTI** [32] | 07'IJCAI | | | RMSE |
| 5 | **E-HMMI** [22] | 08'NC | | VRR | UIQI | 6 | **EHMM** [21] | 08'TCSVT | | VRR | |
| 7 | **MRF** [64] | 08'PAMI | | VRR | | 8 | **SL** [70] | 10'NC | | VRR | UIQI |
| 9 | **RMRF** [86] | 10'ECCV | | VRR | | 10 | **SNS-SRE** [20] | 12'TCSVT | | VRR | |
| 11 | **MWF** [92] | 12'CVPR | | VRR | | 12 | **SCDL** [60] | 12'CVPR | | | PSNR |
| 13 | **Trans** [57] | 13'TNNLS | | VRR | | 14 | **SFS-SVR** [56] | 13'PRL | | VRR | VIF |
| 15 | **Survey** [58] | 14'IJCV | | | RMSE, UIQI, SSIM | 16 | **SSD** [45] | 14'ECCV | SV | VRR | |
| 17 | **SFS** [83] | 15'TIP | | VRR | FSIM, SSIM | 18 | **FCN** [75] | 15'ICMR | ES | VRR | |
| 19 | **RFSSS** [81] | 16'TIP | | VRR | FSIM, SSIM | 20 | **KD-Tree** [88] | 16'ECCV | | VRR | VIF, SSIM |
| 21 | **MrFSPS** [36] | 16'TNNLS | | VRR | FSIM, VIF, SSIM | 22 | **2DDCM** [51] | 16'TIP | | VRR | FSIM, SSIM |
| 23 | **RR** [59] | 17'NC | | VRR | VIF, SSIM | 24 | **Bayesian** [55] | 17'TIP | | VRR | VIF, SSIM |
| 25 | **RFSSS** [82] | 17'TCSVT | | VRR | FSIM, SSIM | 26 | **S-FSPS** [35] | 17'TCSVT | | VRR | FSIM, VIF, SSIM |
| 27 | **ArFSPS** [29] | 17'NC | | VRR | FSIM | 28 | **BFCN** [74] | 17'TIP | SV | VRR | |
| 29 | **DGFL** [93] | 17'IJCAI | | VRR | SSIM | 30 | **FreeH** [30] | 17'IJCV | SV | | |
| 31 | **Pix2pix** [27] | 17'CVPR | | | | 32 | **CA-GAN** [18] | 17'CVPR | | VRR | SSIM |
| 33 | **ESSFA** [14] | 17'TOG | | | | 34 | **PS$^2$MAN** [52] | 18'FG | | VRR | FSIM, SSIM |
| 35 | **NST** [38] | 17'NPAR | | | | 36 | **CMSG** [77] | 18'TC | SV | VRR | |
| 37 | **RSLCR** [53] | 18'PR | | VRR | SSIM | 38 | **MRNF** [78] | 18'IJCAI | | | VIF, SSIM |
| 39 | **$\rho$GAN** [84] | 18'IJCAI | | | FSIM | 40 | **FSSN** [28] | 18'PR | | | PSNR, SSIM |
| 41 | **MAL** [85] | 18'TNNLS | | | FSIM, SSIM | 42 | **MRNF** [79] | 18'AAAI | | VRR | VIF, SSIM |

Table 1: **Summarization of 42 representative FSS-based algorithms. Sj.**: Subjective metric. **Rr.**: Recognition rates. **Ob.**: Objective metric. SV = Subjective Voting. ES = Empirical Study. VRR = various recognition rate methods, such as, null-space LDA [4], Random Sampling LDA [62, 63], dual-space LDA [61], LPP [26], Sparse Representation and Classification [68]. Note that UIQI [65] is a special case of SSIM [66].

For example, in face sketch synthesis (FSS), the target is for the synthetic sketch to be indistinguishable from the reference by a human subject, although their pixel representations might be mismatched. Let us take a look at Fig. 1 in which there are three examples. Which one is closer to the middle reference? While this comparison task seems trivial for humans, to date the widely-used metrics disagree with human judgments. Not only are visual patterns very high-dimensional, but the very notion of visual similarity is often subjective [80].

Our contributions to the facial sketch community can be summarized in three points. Firstly, as described in Sec. 3, we propose a Structure Co-Occurrence Texture (**Scoot**) perceptual metric for FSS that provides a unified evaluation considering both structure and texture.

Secondly, as described in Sec. 4.2, we design three meta-measures based on the above three reliable properties. Extensive experiments on these meta-measures verify that our Scoot metric exceeds the performance of prior works. Our experiments indicate that "spatial structure" and "co-occurrence" texture are two generally applicable perceptual features in FSS.

Thirdly, we explore different ways of exploiting texture statistics (*e.g.*, Gabor, Sobel, and Canny, *etc.*). We find that simple texture features [16, 17] performs far better than the commonly used metrics in the literature [39, 40, 66, 72, 76]. Based on our findings, we construct the first large-scale human-perception-based sketch database that can evaluate how well a metric is in line with human perception.

Our three contributions presented above offer a complete metric benchmark suite, which provides a novel view and a practical tool (*e.g.*, metric, meta-measures and database) to analyze data similarity from the human perception direction.

## 2. Related Work

From Tab. 1, we observe that some works utilize recognition rates (Rr.) to evaluate the quality of synthetic sketches. However, Rr. cannot completely reflect the visual quality of synthetic sketches [54]. In the FSS area, the widely-used perceptual metrics, *e.g.*, SSIM [66], FSIM [76], and VIF [39] were initially designed for image quality assessment (IQA) which aims to evaluate image distortion such as Gaussian blur, jpeg, and jpeg 2000 compression. Directly introducing the IQA metric to FSS may be intractable (see Fig. 1) due to the different nature of their task.

Psychophysics [95] and prior work, *e.g.*, line drawings [15, 23] indicate that human perception of sketch similarity depends on two crucial factors, *i.e.*, image *structure* [66] and *texture* [54]. However, how perceptual are these so-called "perceptual features"? Which elements are critical for their success? How well do these "perceptual features" actually correspond to human visual perceptions? As noticed by Wang *et al*. [54], there is currently no reliable perceptual metric in FSS. We review the topics most pertinent to facial sketch within the constraints of space:

**Heuristic-based Metric.** The most widely used metric in FSS is SSIM proposed by Wang *et al*. [66]. SSIM computes structure similarity, and luminance and contrast comparison using a sliding window on the local patch. Sheikh and Bovik [39] proposed the VIF metric which evaluates

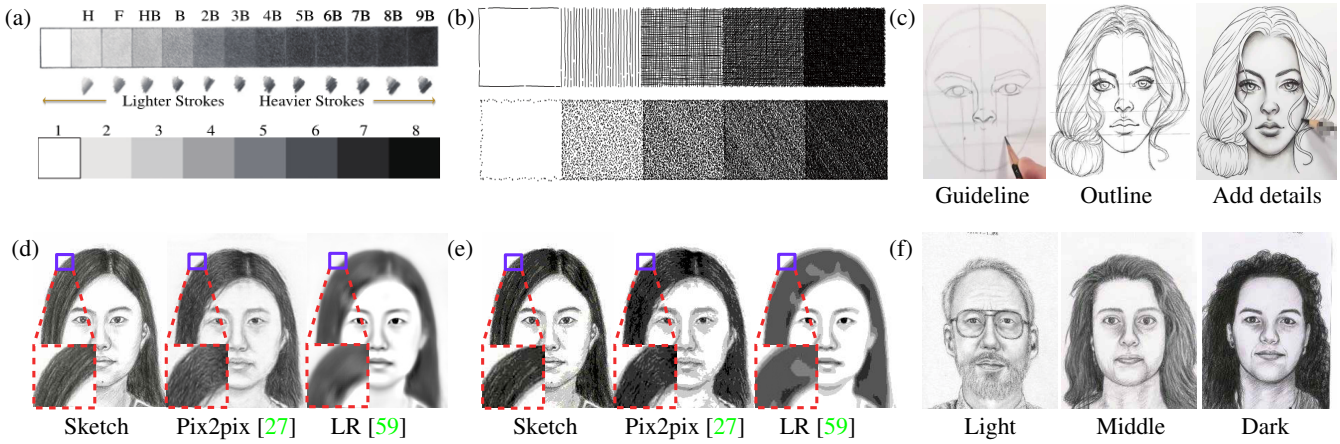



Figure 2: **Motivation of the proposed Scoot metric.** (a) Pencil grades and their strokes. (b) Using stroke tones to present texture. The stroke textures used, from top to bottom, are: "cross-hatching", "stippling". The stroke attributes, from left to right, are: spare to dense. Images are from [67]. (c) The artist draws the sketch from guideline to details. (d) The original sketches. (e) The quantized sketches. (f) Creating various tones of the stroke by applying different pressure (*e.g.*, light to dark) on the pencil tip.

the image quality by quantifying two kinds of information. One is obtained via the human visual system channel, with the input ground truth and the output reference image information. The other is achieved via the distortion channel, called distortion information, and the result is the ratio of these two types of information. Studies of the human vision system (HVS) found that the features perceived by human vision are consistent with the phase of the Fourier series at different frequencies. Therefore, Zhang *et al*. [76] chose phase congruency as the primary feature. Then they proposed a low-level feature similarity metric called FSIM.

Recently, Xue *et al*. [72] devised a simple metric named gradient magnitude similarity deviation (GMSD), where the pixel-wise gradient magnitude similarity is utilized to obtain image local quality. The standard deviation of the overall gradient magnitude similarity map is calculated as the final image quality index. Their metric achieves the state-of-the-art (SOTA) performance compared with the other metrics.

**Learning based Metric.** As well as the heuristic-based metric, there are numerous learning based metrics [7, 19, 48], for comparing images in a perceptual-based manner which have been used to evaluate image compression and many other imaging tasks. We refer readers to a recent survey [80] for a comprehensive review of various deep features adopted for perceptual metrics. This paper focuses on showing why face sketches require a specific perceptual distance metric that differs from or improves upon previously heuristic-based methods.

## 2.1. Motivation

We observed the basic principles of the sketch and noted that "graphite pencil grades" and "pencil's strokes" are the two fundamental elements in the sketch.

## 2.2. Graphite Pencil Grades.

In the European system, "H" & "B" stand for "hard" & "soft" pencil, respectively. Fig. 2(a) illustrates the grade of graphite pencil. Sketch images are expressed through a limited medium (graphite pencil) which provides no color. Illustrator Sylwia Bomba [47] said that "if you put your hand closer to the end of the pencil, you have darker markings. Gripping further up the pencil will result in lighter markings." Besides, after a long period of practice, artists will form their fixed pressure (*e.g.*, from guideline to detail in Fig. 2(c)) style. In other words, the marking of the stroke can be varied (*e.g.*, light to dark in Fig. 2(f)) by changing the pressure on the pencil tip. Note that different pressures on the tip will result in various types of marking which is one of the quantifiable factors called gray tone.

*Gray Tone.* The quantification of gray tone should reduce the effect of slight amounts of noise and oversensitivity to subtle gray tone gradient changes in sketches. We introduce intensity quantization during the evaluation of gray tone similarity. Inspired by previous works [6], we can quantize the input sketch $I$ to $N_l$ different grades to reduce the number of intensities to be considered: $I' = \Omega(I)$. A typical example of such quantization is shown in Fig. 2(d, e). Humans will consistently rank Pix2pix higher than LR before (Fig. 2(d)) and after (Fig. 2(e)) quantizing the input sketches when evaluating the perceptual similarity. Although quantization may introduce artifacts, our experiments (Sec. 6) also show that this process can reduce sensitivity to minor intensity variations and balance the performance and computational complexity.

## 2.3. Pencil's Strokes.

Because all of the sketches are generated by moving a tip on the paper, different paths of the tip along the paper will create various stroke shapes. One example is

shown in Fig. 2(b), in which different spatial distributions of the stroke have produced various textures (*e.g.*, sparse or dense). Thus, the stroke tone is another quantifiable factor.

*Stroke Tone.* The stroke tone and grey tone are not independent concepts. The gray tone is based on the different strokes of gray-scale in a sketch image, while the stroke tone can be defined as the spatial distribution of gray tones.

An example is shown in Fig. 2(d). Intuitively, Pix2pix [27] is better than LR [59] since Pix2pix preserves the texture (or stroke tone) of the hair and details in the face. However, LR presents an overly smooth result and has lost much of the sketch style.

## 3. Proposed Algorithm

This section explains the proposed Scoot metric, which captures the *co-occurrence texture* statistics in the "block-level" *spatial structure*.

### 3.1. Co-Occurrence Texture

With the two quantifiable factors at hand, we start to describe the details. To simultaneously extract statistics about the "stroke tone" and their relationship to the surrounding "gray tone", we need to characterize their *spatial interrelationships*. Previous work in texture [25] verified that the *co-occurrence matrix* can efficiently capture the texture feature, due to the use of various powerful statistics. Since the sketches show a lot of similarities to textures, we use the co-occurrence matrix as our gray tone and stroke tone extractor. Specifically, this matrix $\mathcal{M}$ is defined as:

$$\mathcal{M}_{(i,j)|d} = \sum_{y=1}^{H}\sum_{x=1}^{W} \begin{cases} 1, \text{ if } I'_{x,y} = i \text{ and } I'_{(x,y)+d} = j \\ 0, \text{ otherwise} \end{cases} \tag{1}$$

where $i$ and $j$ denote the gray value; $d = (\Delta x, \Delta y)$ is the relative distance to $(x, y)$; $x$ and $y$ are the spatial positions in the given quantized sketch $I'$; $I'_{x,y}$ denotes the gray value of $I'$ at position $(x, y)$; $W$ and $H$ are the width and height of the sketch $I'$, respectively. To extract the perceptual features in a sketch, we test the three most widely used [24] statistics: Homogeneity ($\mathcal{H}$), Contrast ($\mathcal{C}$), and Energy ($\mathcal{E}$).

*Homogeneity* reflects how much the texture changes in local regions, it will be high if the gray tone of each pixel pair is similar. The homogeneity is defined as:

$$\mathcal{H} = \sum_{j=1}^{N_l}\sum_{i=1}^{N_l} \frac{\mathcal{M}_{(i,j)|d}}{1+|i-j|}, \tag{2}$$

*Contrast* represents the difference between a pixel in $I'$ and its neighbor summed over the whole sketch. This reflects that a low-contrast sketch is not characterized by low gray tones but rather by low spatial frequencies. The contrast is highly correlated with spatial frequencies. The contrast equals 0 for a constant tone sketch.

$$\mathcal{C} = \sum_{j=1}^{N_l}\sum_{i=1}^{N_l} |i-j|^2 \mathcal{M}_{(i,j)|d} \tag{3}$$

**Algorithm 1:** Structure Co-Occurrence Texture Measure

**Input:** Synthetic Sketch $X$, Ground Truth Sketch $Y$
Step 1: Quantize $X$ and $Y$ into $N_l$ grades
Step 2: Calculate the matrices $\mathcal{M}(X)$ and $\mathcal{M}(Y)$ according to Eq. 1
Step 3: Divide the whole sketch image into a $k \times k$ grid of $k^2$ blocks
Step 4: Extract the $\mathcal{CE}$ features according to Eq. 3 & 4 from each block and concatenate them together
Step 5: Compute the average feature of four orientations with Eq. 5
Step 6: Evaluate the similarity between $X$ and $Y$ according to Eq. 6
**Output:** Scoot score;

*Energy* measures textural uniformity. When only similar gray tones of pixels occur in a sketch ($I'$) patch, a few elements in $\mathcal{M}$ will be close to 1, while others will be close to 0. Energy will reach the maximum if there is only one gray tone in a sketch ($I'$) patch. Thus, high energy corresponds to the sketch's gray tone distribution having either a periodic or constant form.

$$\mathcal{E} = \sum_{j=1}^{N_l}\sum_{i=1}^{N_l} (\mathcal{M}_{(i,j)|d})^2 \tag{4}$$

### 3.2. Spatial Structure

To holistically represent the spatial structure, we follow the spatial envelope strategy [9, 34] to extract the statistics from the "block-level" *spatial structure* in the sketch. First, we divide the whole sketch image into a $k \times k$ grid of $k^2$ blocks. Our experiments demonstrate that the process can help to derive content information. Second, we compute the co-occurrence matrix $\mathcal{M}$ for all blocks and normalize each matrix such that the sum of its components is 1. Then, we concatenate $p$ statistics (*e.g.*, $\mathcal{H}, \mathcal{C}, \mathcal{E}$) of all the $k^2$ blocks into a vector $\overrightarrow{\Phi}(I'_s|d) \in \mathbb{R}^{p\times k\times k}$.

Note that each of the above statistics is based on a single direction (*e.g.*, $90^o$, that is $d$ = (0, 1)), since the direction of the spatial distribution is also very important to capture the style such as "hair direction", "the direction of shadowing strokes". To exploit this observation for efficiently extracting the *stroke direction* style, we compute the average feature $\overrightarrow{\Psi}(I'_s)$ of $T$ orientation vectors to capture more directional information:

$$\overrightarrow{\Psi}(I'_s) = \frac{1}{T}\sum_{i=1}^{T} \overrightarrow{\Phi}(I'_s|d_i), \tag{5}$$

where $d_i$ denotes the $i$th direction and $\overrightarrow{\Psi}(I'_s) \in \mathbb{R}^{p\times k\times k}$.

### 3.3. Scoot Metric

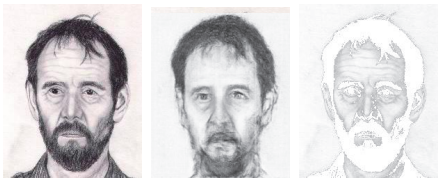
After obtaining the perceptual feature vectors of the reference sketch $Y$ and synthetic sketch $X$, a function is needed to evaluate their similarity. We have tested various forms of functions such as Euclidean distance or exponential functions, *etc*., but have found that the simple Euclidean distance is a simple and effective function and works best in our experiments. Thus, the proposed perceptual similarity

(a) Photo (b) Reference (c) Difference (d) Downsized (e) Reference (f) Re-sized (g) Pix2pix (h) LR

Figure 3: **Meta-measure 1: Stability to Slight Re-sizing.**

(a) Reference (b) R-Reference (c) Pix2pix (d) MWF

Figure 4: **Meta-measure 2: Rotation Sensitivity.**



(a) Reference (b) Synthetic (c) Light

Figure 5: **Meta-measure 3: Content Capture Capability.**

Scoot metric can be defined as:

$$E_s = \frac{1}{1 + \left\| \vec{\Psi}(X'_s) - \vec{\Psi}(Y'_s) \right\|_2}. \quad (6)$$

where $\|\cdot\|_2$ denotes the $l_2$-norm. $X'_s$, $Y'_s$ denote the quanitized $X_s$, $Y_s$, respectively. $E_s = 1$ represents identical style.

# 4. Experiments

## 4.1. Implementation Details

The size of spatial structure $k$ in Sec. 3.2 is set to 4 to achieve the best performance. The quantization parameter $N_l$ in Eq. (6) is set to 6 grades. We have demonstrated that $p = 2$ (*e.g.*, $\mathcal{C}$ in Eq. 3 combined with $\mathcal{E}$ in Eq. 4) achieve the best performance (see Sec. 6). Due to the symmetry of the co-occurrence matrix $\mathcal{M}(i, j)$, the statistical features in 4 orientations are actually equivalent to the 8 neighbor directions at 1 distance. Empirically, we set $T = 4$ orientations $d_i \in \{(0, 1), (-1, 1), (-1, 0), (-1, -1)\}$ to achieve the robust performance.

## 4.2. Meta-measures

As described in [33], one of the most challenging tasks in designing a metric is proving its performance. Following [37], we use the *meta-measure* methodology, which is a measure that assesses a metric. Inspired by [9, 10, 33], we further propose three meta-measures based on the 3 properties described in Sec. 1.

**Meta-measure 1: Stability to Slight Resizing.** The first meta-measure (MM1) specifies that the rankings of synthetic sketches should not change much with slight changes in the reference sketch. Therefore, we perform a minor 5 pixels downsizing of the reference by using nearest-neighbor interpolation. Fig. 3 gives an example. The hair of the reference in (b) drawn by the artist has a slight size discrepancy compared to the photo (a). We observe that about 5 pixels deviation (Fig. 3(c)) in the boundary is common. Although the two sketches (e) & (f) are almost identical, widely-used metrics, *e.g.*, SSIM [66], VIF [39], and GMSD [72] switched the ranking of the two synthetic sketches (g, h) when using (e) or (f) as the reference. However, the proposed Scoot metric consistently ranked (g) higher than (h).

For this meta-measure, we applied the $\theta = 1 - \rho$ [2] measure to test the metric ranking stability before and after the reference downsizing was performed. The value of $\theta$ falls in the range [0, 2].

Tab. 2 shows the results: the lower the result is, the more stable a metric is to slight downsizing. We can see a significant ($\approx 77\%$ and $83\%$) improvement over the existing SSIM, FSIM, GMSD, and VIF metrics in both the CUFS and CUFSF databases. These improvements are mainly because the proposed metric considers "block-level" statistics rather than "pixel-level".

**Meta-measure 2: Rotation Sensitivity.** In real-world situations, sketches drawn by artists may also have slight rotations compared to the original photographs. Thus, the proposed second meta-measure (MM2) verifies the sensitivity of reference rotation for the evaluation metric. We did a slight counter-clockwise rotation ($5^o$) for each reference. Fig. 4 shows an example. When the reference (a) is switched to the slightly rotated reference (b), the ranking results should not change much. In MM2, we got the ranking results for each metric by applying reference sketches and slightly rotated reference sketches (R-Reference) separately. We utilized the same measure ($\theta$) as meta-measure 1 to evaluate the rotation sensitivity.

The sensitivity results are shown in Tab. 2. It is worth noting that MM2 and MM1 are two aspects of the expected property described in Sec. 4.2. Our metric again significantly outperforms the current metrics over the CUFS and CUFSF databases.

| Metic | *CUFS* [64] MM1↓ resize | MM2↓ rotation | MM3↑ content | *RCUFS* Jud↑ judgment | *CUFSF* [87] MM1↓ resize | MM2↓ rotation | MM3↑ content | *RCUFSF* Jud↑ judgment |
|---|---|---|---|---|---|---|---|---|
| Classical & Widely Used | | | | | | | | |
| IFC [40] | 0.256 | 0.189 | 1.20% | 26.9% | 0.089 | 0.112 | 3.07% | 25.4% |
| SSIM [66] | 0.162 | 0.086 | 81.4% | 37.3% | 0.073 | 0.074 | 97.4% | 36.8% |
| FSIM [76] | 0.268 | 0.123 | 14.2% | 50.0% | 0.151 | 0.058 | 32.4% | 37.5% |
| VIF [39] | 0.322 | 0.236 | 43.5% | 44.1% | 0.111 | 0.150 | 22.2% | 52.8% |
| GMSD [72] | 0.417 | 0.210 | 21.9% | 42.6% | 0.259 | 0.132 | 63.6% | 58.6% |
| **Scoot (Ours)** | **0.037** | **0.025** | **95.9%** | **76.3%** | **0.012** | **0.008** | **97.5%** | **78.8%** |
| Texture-based & Edge-based | | | | | | | | |
| Canny [3] | 0.086 | 0.078 | 33.7% | 27.8% | 0.138 | 0.146 | 0.00% | 0.10% |
| Sobel [44] | 0.040 | 0.037 | 0.00% | 32.8% | 0.048 | 0.044 | 0.00% | 52.6% |
| GLRLM [17] | 0.111 | 0.111 | 18.6% | 73.7% | 0.125 | 0.079 | 64.6% | 68.0% |
| Gabor [16] | 0.062 | 0.055 | 0.00% | 72.2% | 0.089 | 0.043 | 19.3% | **80.9%** |
| **Scoot (Ours)** | **0.037** | **0.025** | **95.9%** | **76.3%** | **0.012** | **0.008** | **97.5%** | 78.8% |
| Feature Combination | | | | | | | | |
| $\mathcal{HEC}$ | 0.034 | 0.024 | 95.9% | 76.3% | 0.011 | 0.008 | 97.4% | 78.7% |
| $\mathcal{H}$ | 0.007 | 0.005 | 61.5% | 77.5% | 0.003 | 0.003 | 79.1% | 77.8% |
| $\mathcal{E}$ | 0.200 | 0.104 | 98.5% | 73.1% | 0.044 | 0.026 | 99.2% | 77.4% |
| $\mathcal{C}$ | 0.010 | 0.007 | 54.4% | 74.6% | 0.009 | 0.006 | 64.7% | 73.4% |
| $\mathcal{HC}$ | 0.011 | 0.007 | 60.1% | 74.6% | 0.007 | 0.005 | 78.1% | 73.7% |
| $\mathcal{HE}$ | 0.156 | 0.088 | 97.9% | 75.7% | 0.030 | 0.017 | 98.8% | 80.3% |
| $\mathcal{CE}$ **(Scoot)** | 0.037 | 0.025 | 95.9% | 76.3% | 0.012 | 0.008 | 97.5% | 78.8% |

Table 2: **Benchmark results of classical and alternative texture/edge based metrics.** The best result is highlighted in **bold**, and these differences are all statistically significant at the $\alpha < 0.05$ level. The ↑ indicates that the higher the score is, the better the metric performs, and vice verse (↓).

**Meta-measure 3: Content Capture Capability.** The third meta-measure (MM3) describes that a good metric should assign a complete sketch generated by a SOTA algorithm a higher score than any sketches that only preserve incomplete strokes. Fig. 5 presents an example. We expect that a metric should prefer the SOTA synthetic result (b) over the *light strokes*[1] result (c). For MM3, we compute the *mean score* of 10 SOTA [27, 31, 45, 53, 55, 59, 64, 75, 92, 93] face sketch synthesis algorithms. The mean score is robust to situations in which a certain model generates a poor result. We recorded the number of times the mean score of SOTA synthetic algorithms is higher than a light stroke's score.

For the case shown in Fig. 5, the current widely-used metrics (SSIM, FSIM, VIF) are all in favor of the light sketch. Only the proposed Scoot metric gives the correct order. In terms of pixel-level matching, it is obvious that the regions where dark strokes are removed are different from the corresponding parts in (a). But at other positions, the pixels are identical to the reference. Previous metrics only consider "pixel-level" matching and will rank the light strokes sketch higher. However, the synthetic sketch (b) is better than the light one (c) in terms of both style and content. From Tab. 2, we observe a great ($\geq$14%) improvement over the other metrics in *CUFS* database. A slight improvement is also achieved for the CUFSF database.

[1] To test the third meta-measure, we use a simple threshold of grayscale (*e.g.* 170) to separate the sketch (Fig. 5 Reference) into darker strokes & lighter strokes. The image with lighter strokes loses the main texture features of the face (*e.g.* hair, eye, beard), resulting in an incomplete sketch.

## 5. Proposed Perceptual Similarity Dataset

To evaluate the performance of different perceptual metrics, we built a large-scale highly diverse dataset of perceptual judgments using the 2 alternative forced-choice (2AFC) scheme [80]. These judgments are derived from a wide space of distortions and real algorithm synthesises. Because the true test of a synthetic sketch assessment metric is on real problems and real algorithms, we gather perceptual judgments using such outputs.

### 5.1. Distortions

**Source Images.** Data on real algorithms is more limited, as each synthesis model will have its own unique properties. To obtain more distortion data, we collect 338 pairs (*CUFS*) and 944 pairs (*CUFSF*) of test set images as source images following the split scheme of [53].

**Distortion Types.** We simulate diverse possible distortions introduced by traditional and CNN-based synthetic methods, to more closely simulate the space of artifacts that can arise from real algorithms [27, 31, 45, 53, 55, 59, 64, 75, 92, 93]. Our goal of each selected algorithms is not to address the task *per se*, but rather to explore common artifacts that plague the outputs of traditional/deep-based methods. As shown in Fig. 6, we introduce 10 distortion types, such

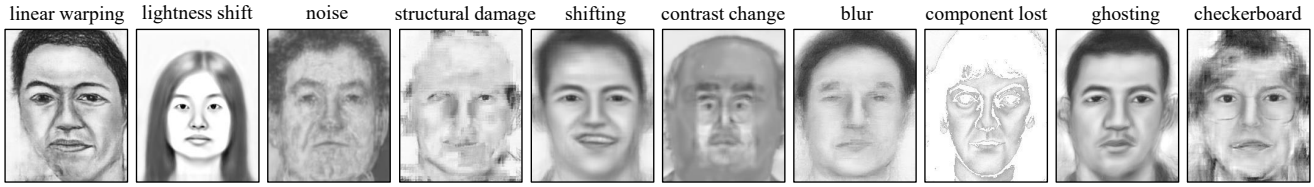



Figure 6: **Our distortions.** These distortions are generated by various real synthesis algorithms [27, 31, 45, 53, 55, 59, 64, 75, 92, 93].

as lightness shift, foreground noise, shifting, linear warping, structural damage, contrast change, blur, component lost, ghosting, and checkerboard artifact.

### 5.2. Psychophysical Similarity Measurements

**Data selection.** 21 viewers, who were pre-trained with 50 pairs of ranking, are asked to rank the synthetic sketch result based on two criteria: texture similarity and content similarity. To minimize the ambiguity of human ranking, we follow the voting strategy [54] to conduct this experiment (∼152K judgments) through the following stages:

- We let the first group of viewers (7 subjects) select four out of ten sketches for each photo. The 4 sketches should consist of two good and two bad ones. Thus, we are left with 1352 ($4 \times 338$), and 3776 ($4 \times 944$) sketches for *CUFS* and *CUFSF*, respectively.
- For the selected four sketches in each photo, the second group of viewers (seven people) is further asked to choose three sketches for which they can rank them easily. Based on the voting results of viewers, we pick out the 3 most frequently selected sketches.
- Sketches that are too similar will make it difficult for viewers to judge which sketch is better, potentially causing them to give random decisions. To avoid this random selection, we ask the last group of viewers (seven) to pick out the pair of sketches that are most obvious to rank.

**2AFC similarity judgments.** For each image, we have a reference sketch $r$ drawn by artists and two distortions $s_0, s_1$. We ask the viewer which is closer to the reference $s$, and record the response $q \in \{0, 1\}$. On average, viewers spent about 2 seconds per judgment. Let $\gamma = (s, s_0, s_1, q)$ denote our dataset of image triplets. Note that we have 5 volunteers involved in the whole process for cross-checking the ranking. For example, if there $\geq 4$ viewer preferences for $s_0$ and $\leq 1$ for $s_1$, the final ranking will be $s_0 > s_1 \& q = 1$. All triplets with a clear majority will be preserved and the other triplets are discarded. Finally, we establish two new human-ranked[2] datasets: ***RCUFS*** and ***RCUFSF***. Please refer to our website for complete datasets.

[2] The two datasets include 1014 (3×338 triplets), and 2832 (3×944 triplets) human-ranked images, respectively. Recent works [46, 69] show that the scale of a dataset is important. To our best knowledge, this is the first large-scale publicly available human judgment dataset in FSS.

### 5.3. Human Judgments

Here, we evaluate how well our Scoot and other compared metrics. The *RCUFS* and *RCUFSF* contain 338 and 944 judged triplets, respectively. To increase an inherently noisy process, we compute the agreement of a metric with each triplet and adopt the *average* statistics among the dataset as the final performance.

**How well do classical metrics and our Scoot perform?** Tab. 2 shows the performance of various classical metrics (*e.g.*, IFC, SSIM, FSIM, VIF, and GMSD). Interestingly, these metrics perform at about the same low level (*e.g.*, ≤59%). Despite its common use in FSS, these metrics were not designed for situations where pixel mismatching is a large factor. However, the proposed Scoot metric shows a significant (∼26.3%) improvement over the best prior metric in *RCUFS*. This improvement is due to our consideration of structure and texture similarity which human perception considers as two essential factors when evaluating sketches.

## 6. Discussion

**Which elements are critical for their success?** In Sec. 3.1, we considered 3 widely-used statistics: Homogeneity ($\mathcal{H}$), Contrast ($\mathcal{C}$), and Energy ($\mathcal{E}$). To achieve the best performance, we need to explore the best combination of these statistical features. We have applied our three meta-measures as well as human judgments to test the performance of the Scoot metric using each single feature, each feature pair and the combination of all three features.

The results are shown in Tab. 2. All possibilities ($\mathcal{H}$, $\mathcal{E}$, $\mathcal{C}$, $\mathcal{CE}$, $\mathcal{HE}$, $\mathcal{CH}$, $\mathcal{HEC}$) perform well in Jud (human judgment). $\mathcal{H}$ and $\mathcal{C}$ are insensitive to re-sizing (MM1) and rotation (MM2), while they are not good at content capture (MM3). $\mathcal{E}$ is the opposite compared to $\mathcal{H}$ and $\mathcal{C}$. Thus, using a single feature is not good. The results of combining two features show that if $\mathcal{H}$ is combined with $\mathcal{E}$, the sensitivity to re-sizing and rotating will still be high, while partially overcoming the weakness of $\mathcal{E}$. The performance of $\mathcal{H} + \mathcal{E} + \mathcal{C}$ shows no improvement compared to the combination of "$\mathcal{CE}$" features. Previous work in [1] also found the energy and contrast to be the most efficient features for discriminating textural patterns. Thus, we choose "$\mathcal{CE}$" feature as our final combination to extract the perceptual features.

**How well do these "perceptual features" actually correspond to human visual perceptions?** As described in Sec. 3.1, sketches are quite close to textures. There are many other texture & edge-based features (*e.g.*

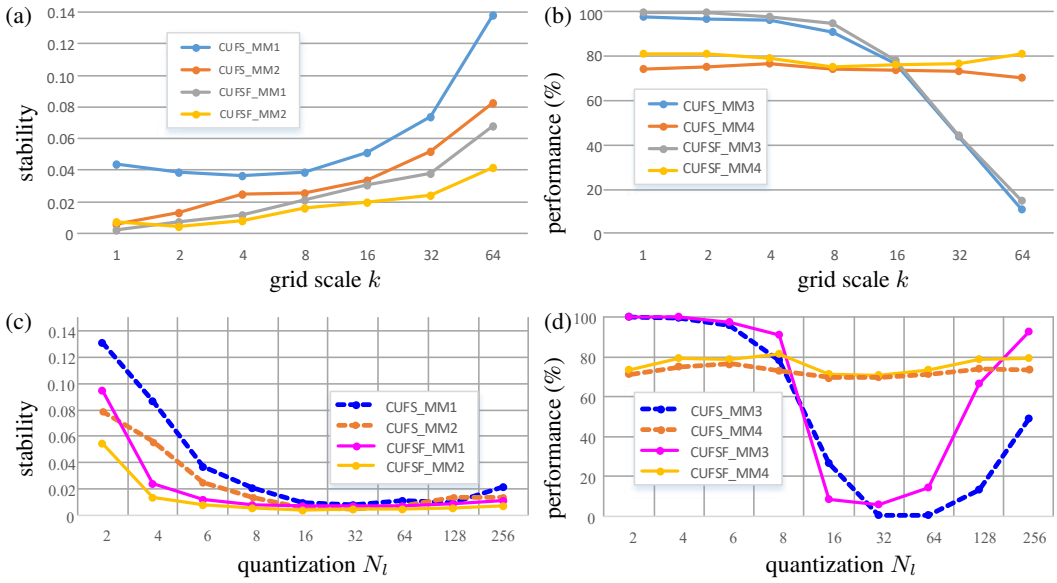



Figure 7: **Sensitivity experiments of the spatial structure (top) and quantization (bottom)**. For MM1 & MM2, the lower the better. For MM3 & MM4, the higher the better.

GLRLM [17], Gabor [16], Canny [3], Sobel [44]). Here, we select the most wide-used features as candidate alternatives to replace our "$\mathcal{CE}$" feature. For GLRLM, we select all five statistics mentioned in the original version. Results are shown in Tab. 2. Gabor and GLRLM are texture features, while the other two are edge-based. All the texture features (GLRLM, Gabor) and the proposed Scoot metric provide a good (*e.g.*, $\geq$68%) consistency with human ranking (Jud). Among all the texture features, the proposed metric provides a consistently high average performance with human ranking (Jud). GLRLM performs well according to MM1 & 2 & 3. Gabor is reasonable in terms of MM1 & 2, but not good at MM3. For edge-based features, Canny fails according to all meta-measures. Sobel is very stable to slight re-sizing (MM1) or rotating (MM2), but cannot capture content (MM3) and is not consistent with human judgment (Jud). Interestingly, Canny, Sobel, and Gabor assigned the incomplete stroke a higher score than the sketch generated by the SOTA algorithm. In other words, the metric has completely reversed the ranking results for all the tested cases. In terms of overall results, we conclude that our "$\mathcal{CE}$" feature is more robust than other competitors.

**What is the Sensitivity to the Spatial Structure?** To analyze the effect of spatial structure, we derive seven variants, each of which divides the sketch with a different sized grid, *i.e.*, $k$ is set to 1, 2, 4, $\cdots$, 64. The results of MM3 & 4 in Fig. 7(b) show that $k$ = 1 achieves the best performance. However, the weakness of this version is that it only captures the "image-level" statistics, and the structure of the sketch is ignored. That is, a sketch made up of an arbitrary arrangement can also achieve a high score. The experiment of MM1 in Fig. 7(a), clearly shows that $k$ = 4 achieves the best performance for the CUFS dataset. Based on the two experiments, $k = 4$ gains the most robust performance.

**What is the Sensitivity to Quantization?** To determine which quantization parameter $N_l$ (baseline: $N_l = \{2, 4, 6, 8, 16, 32, 64, 128\}$) produces the best performance we perform a further sensitivity test. From Fig. 7(c)&(d), we observe that quantizing the input sketch to 32 grey levels achieves an excellent result. However, for the experiments of MM3 & MM4, it gains the worst performance. Considering overall experiments, $N_l = 6$ achieves a more robust result.

## 7. Conclusion

In this work, we explore the human perception problem, *e.g.*, what is the difference between human choice and metrics. A tool used to analyze the above question are facial sketches. We provide a specific metric, called Scoot (Structure Co-Occurrence Texture), that captures human perception, and is analyzed by the proposed three meta-measures. Finally, we built the first human-perception-based sketch database that can evaluate how well a metric is in line with human perception. We systematically evaluate different texture-based/edge-based features on our Scoot architecture and compare them with classic metrics. Our results show that "spatial structure" and "co-occurrence" texture are two generally applicable perceptual features in facial sketches. In the future, we will continue to develop and apply Scoot in order to further push the frontiers of research, *e.g.*, for evaluation of background subtraction [94].

**Acknowledgments.** This research was supported by NSFC (61572264, 61620106008, 61802324, 61772443), the national youth talent support program, and Tianjin Natural Science Foundation (17JCJQJC43700, 18ZXZNGX00110).

## References


[1] Andrea Baraldi and Flavio Parmiggiani. An investigation of the textural characteristics associated with gray level cooccurrence matrix statistical parameters. *IEEE T Geosci. Remote.*, 33(2):293–304, 1995.

[2] DJ Best and DE Roberts. Algorithm AS 89: the upper tail probabilities of Spearman's rho. *J R STAT SOC C-APPL*, 24(3):377–379, 1975.

[3] John Canny. A computational approach to edge detection. *IEEE TPAMI*, 8:679–698, 1986.

[4] Li-Fen Chen, Hong-Yuan Mark Liao, Ming-Tat Ko, Ja-Chen Lin, and Gwo-Jong Yu. A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33(10):1713–1726, 2000.

[5] Ming-Ming Cheng, Yun Liu, Wen-Yan Lin, Ziming Zhang, Paul L Rosin, and Philip HS Torr. BING: Binarized normed gradients for objectness estimation at 300fps. *Computational Visual Media*, 5(1):3–20, 2019.

[6] David A Clausi. An analysis of co-occurrence texture statistics as a function of grey level quantization. *Can. J Remote. Sens.*, 28(1):45–62, 2002.

[7] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *NIPS*, pages 658–666, 2016.

[8] Mohamed Elhoseiny, Yizhe Zhu, Han Zhang, and Ahmed Elgammal. Link the head to the beak: Zero shot learning from noisy text description at part precision. In *IEEE CVPR*, 2017.

[9] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A New Way to Evaluate Foreground Maps. In *IEEE ICCV*, pages 4548–4557, 2017.

[10] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *IJCAI*, pages 698–704, 2018.

[11] Deng-Ping Fan, Zheng Lin, Jia-Xing Zhao, Yun Liu, Zhao Zhang, Qibin Hou, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. *arXiv preprint arXiv:1907.06781*, 2019.

[12] Deng-Ping Fan, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, Ali Borji, and Ming-Ming Cheng. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, pages 1597–1604. Springer, 2018.

[13] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *IEEE CVPR*, pages 8554–8564, 2019.

[14] Jakub Fišer, Ondřej Jamriška, David Simons, Eli Shechtman, Jingwan Lu, Paul Asente, Michal Lukáč, and Daniel Sỳkora. Example-based synthesis of stylized facial animations. *ACM TOG*, 36(4):155, 2017.

[15] William T Freeman, Joshua B Tenenbaum, and Egon C Pasztor. Learning style translation for the lines of a drawing. *ACM TOG*, 22(1):33–46, 2003.

[16] Dennis Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):429–441, 1946.

[17] Mary M Galloway. Texture analysis using grey level run lengths. *NASA STI/Recon Technical Report N*, 75, 1974.

[18] Fei Gao, Shengjie Shi, Jun Yu, and Qingming Huang. Composition-aided sketch-realistic portrait generation. *arXiv preprint arXiv:1712.00899*, 2017.

[19] Fei Gao, Yi Wang, Panpeng Li, Min Tan, Jun Yu, and Yani Zhu. Deepsim: Deep similarity for image quality assessment. *Neurocomputing*, 257:104–114, 2017.

[20] Xinbo Gao, Nannan Wang, Dacheng Tao, and Xuelong Li. Face sketch–photo synthesis and retrieval using sparse representation. *IEEE TCSVT*, 22(8):1213–1226, 2012.

[21] Xinbo Gao, Juanjuan Zhong, Jie Li, and Chunna Tian. Face sketch synthesis algorithm based on E-HMM and selective ensemble. *IEEE TCSVT*, 18(4):487–496, 2008.

[22] Xinbo Gao, Juanjuan Zhong, Dacheng Tao, and Xuelong Li. Local face sketch synthesis learning. *Neurocomputing*, 71(10-12):1921–1930, 2008.

[23] Stéphane Grabli, Emmanuel Turquin, Frédo Durand, and François X Sillion. Programmable style for NPR line drawing. *Rendering Techniques (Eurographics Symposium on Rendering)*, 2004.

[24] Robert M Haralick et al. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[25] Robert M Haralick, Karthikeyan Shanmugam, et al. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 610–621, 1973.

[26] Xiaofei He and Partha Niyogi. Locality preserving projections. In *NIPS*, pages 153–160, 2004.

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE CVPR*, pages 1125–1134, 2017.

[28] Licheng Jiao, Sibo Zhang, Lingling Li, Fang Liu, and Wenping Ma. A modified convolutional neural network for face sketch synthesis. *PR*, 76:125–136, 2018.

[29] Jie Li, Xinye Yu, Chunlei Peng, and Nannan Wang. Adaptive representation-based face sketch-photo synthesis. *Neurocomputing*, 269:152–159, 2017.

[30] Yi Li, Yi-Zhe Song, Timothy M Hospedales, and Shaogang Gong. Free-hand sketch synthesis with deformable stroke models. *IJCV*, 122(1):169–190, 2017.

[31] Qingshan Liu, Xiaoou Tang, Hongliang Jin, Hanqing Lu, and Songde Ma. A nonlinear approach for face sketch synthesis and recognition. In *IEEE CVPR*, volume 1, pages 1005–1010, 2005.

[32] Wei Liu, Xiaoou Tang, and Jianzhuang Liu. Bayesian Tensor Inference for Sketch-Based Facial Photo Hallucination. In *IJCAI*, pages 2141–2146, 2007.

[33] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *IEEE CVPR*, pages 248–255, 2014.

[34] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[35] Chunlei Peng, Xinbo Gao, Nannan Wang, and Jie Li. Superpixel-based face sketch–photo synthesis. *IEEE TCSVT*, 27(2):288–299, 2017.

[36] Chunlei Peng, Xinbo Gao, Nannan Wang, Dacheng Tao, Xuelong Li, and Jie Li. Multiple representations-based face sketch–photo synthesis. *IEEE TNNLS*, 27(11):2201–2215, 2016.

[37] Jordi Pont-Tuset and Ferran Marques. Measures and meta-measures for the supervised evaluation of image segmentation. In *IEEE CVPR*, pages 2131–2138, 2013.

[38] Amir Semmo, Tobias Isenberg, and Jürgen Döllner. Neural style transfer: a paradigm shift for image-based artistic rendering? In *ACM NPAR*, page 5, 2017.

[39] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE TIP*, 15(2):430–444, 2006.

[40] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE TIP*, 14(12):2117–2128, 2005.

[41] Jianbing Shen, Yunfan Du, Wenguan Wang, and Xuelong Li. Lazy random walks for superpixel segmentation. *IEEE TIP*, 23(4):1451–1462, 2014.

[42] Jianbing Shen, Xiaopeng Hao, Zhiyuan Liang, Yu Liu, Wenguan Wang, and Ling Shao. Real-time superpixel segmentation by DBSCAN clustering algorithm. *IEEE TIP*, 25(12):5933–5942, 2016.

[43] Jianbing Shen, Jianteng Peng, and Ling Shao. Submodular trajectories for better motion segmentation in videos. *IEEE TIP*, 27(6):2688–2700, 2018.

[44] Irvin Sobel. An isotropic 3×3 image gradient operator. *Machine Vision for Three-dimensional Scenes*, pages 376–379, 1990.

[45] Yibing Song, Linchao Bao, Qingxiong Yang, and Ming-Hsuan Yang. Real-time exemplar-based face sketch synthesis. In *ECCV*, pages 800–813. Springer, 2014.

[46] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *ECCV*, pages 206–222. Springer, 2016.

[47] Bomba Sylwia, Cai Rovina, Croes Brun, Gerard Justin, and Lewis Marisa. *Beginner's Guide to Sketching*. 3dtotal Publishing, 2015.

[48] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE TIP*, 27(8):3998–4011, 2018.

[49] Xiaoou Tang and Xiaogang Wang. Face sketch synthesis and recognition. In *IEEE CVPR*, pages 687–694, 2003.

[50] Xiaoou Tang and Xiaogang Wang. Face sketch recognition. *IEEE TCSVT*, 14(1):50–57, 2004.

[51] Ching-Ting Tu, Yu-Hsien Chan, and Yi-Chung Chen. Facial Sketch Synthesis Using 2D Direct Combined Model-Based Face-Specific Markov Network. *IEEE TIP*, 25(8):3546–3561, 2016.

[52] Lidan Wang, Vishwanath Sindagi, and Vishal Patel. High-quality facial photo-sketch synthesis using multi-adversarial networks. In *IEEE FG*, pages 83–90, 2018.

[53] Nannan Wang, Xinbo Gao, and Jie Li. Random sampling for fast face sketch synthesis. *Pattern Recognition*, 76:215–227, 2018.

[54] Nannan Wang, Xinbo Gao, Jie Li, Bin Song, and Zan Li. Evaluation on synthesized face sketches. *Neurocomputing*, 214:991–1000, 2016.

[55] Nannan Wang, Xinbo Gao, Leiyu Sun, and Jie Li. Bayesian face sketch synthesis. *IEEE TIP*, 26(3):1264–1274, 2017.

[56] Nannan Wang, Jie Li, Dacheng Tao, Xuelong Li, and Xinbo Gao. Heterogeneous image transformation. *PRL*, 34(1):77–84, 2013.

[57] Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. Transductive face sketch-photo synthesis. *IEEE TNNLS*, 24(9):1364–1376, 2013.

[58] Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. A comprehensive survey to face hallucination. *IJCV*, 106(1):9–30, 2014.

[59] Nannan Wang, Mingrui Zhu, Jie Li, Bin Song, and Zan Li. Data-driven vs. model-driven: Fast face sketch synthesis. *Neurocomputing*, 2017.

[60] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *IEEE CVPR*, pages 2216–2223, 2012.

[61] Xiaogang Wang and Xiaoou Tang. Dual-space linear discriminant analysis for face recognition. In *IEEE CVPR*, volume 2, pages II–II, 2004.

[62] Xiaogang Wang and Xiaoou Tang. Random sampling lda for face recognition. In *IEEE CVPR*, pages 259–265, 2004.

[63] Xiaogang Wang and Xiaoou Tang. Random sampling for subspace face recognition. *IJCV*, 70(1):91–104, 2006.

[64] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE TPAMI*, 31(11):1955–1967, 2009.

[65] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE SPL*, 9(3):81–84, 2002.

[66] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004.

[67] Georges Winkenbach and David H Salesin. Computer-generated pen-and-ink illustration. In *ACM SIGGRAPH*, pages 91–100, 1994.

[68] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE TPAMI*, 31(2):210–227, 2009.

[69] Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. Ip102: A large-scale benchmark dataset for insect pest recognition. In *IEEE CVPR*, pages 8787–8796, 2019.

[70] Bing Xiao, Xinbo Gao, Dacheng Tao, Yuan Yuan, and Jie Li. Photo-sketch synthesis and recognition based on subspace learning. *Neurocomputing*, 73(4-6):840–852, 2010.

[71] Jun Xu, Lei Zhang, David Zhang, and Xiangchu Feng. Multi-channel weighted nuclear norm minimization for real color image denoising. In *IEEE ICCV*, 2017.

[72] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE TIP*, 23(2):684–695, 2014.

[73] Jufeng Yang, Xiaoxiao Sun, Jie Liang, and Paul L Rosin. Clinical skin lesion diagnosis using representations inspired by dermatologist criteria. In *IEEE CVPR*, pages 1258–1266, 2018.

[74] Dongyu Zhang, Liang Lin, Tianshui Chen, Xian Wu, Wenwei Tan, and Ebroul Izquierdo. Content-adaptive sketch portrait generation by decompositional representation learning. *IEEE TIP*, 26(1):328–339, 2017.

[75] Liliang Zhang, Liang Lin, Xian Wu, Shengyong Ding, and Lei Zhang. End-to-end photo-sketch generation via fully convolutional representation learning. In *ACM ICMR*, pages 627–634, 2015.

[76] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE TIP*, 20(8):2378–2386, 2011.

[77] Mingjin Zhang, Jie Li, Nannan Wang, and Xinbo Gao. Compositional model-based sketch generator in facial entertainment. *IEEE TOC*, 48(3):904–915, 2018.

[78] Mingjin Zhang, Nannan Wang, Xinbo Gao, and Yunsong Li. Markov random neural fields for face sketch synthesis. In *IJCAI*, pages 1142–1148, 2018.

[79] Mingjin Zhang, Nannan Wang, Yunsong Li, Ruxin Wang, and Xinbo Gao. Face sketch synthesis from coarse to fine. In *AAAI*, 2018.

[80] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE CVPR*, pages 586–595, 2018.

[81] Shengchuan Zhang, Xinbo Gao, Nannan Wang, and Jie Li. Robust face sketch style synthesis. *IEEE TIP*, 25(1):220–232, 2016.

[82] Shengchuan Zhang, Xinbo Gao, Nannan Wang, and Jie Li. Face sketch synthesis from a single photo–sketch pair. *IEEE TCSVT*, 27(2):275–287, 2017.

[83] Shengchuan Zhang, Xinbo Gao, Nannan Wang, Jie Li, and Mingjin Zhang. Face sketch synthesis via sparse representation-based greedy search. *IEEE TIP*, 24(8):2466–2477, 2015.

[84] Shengchuan Zhang, Rongrong Ji, Jie Hu, Yue Gao, and Chia-Wen Lin. Robust face sketch synthesis via generative adversarial fusion of priors and parametric sigmoid. In *IJCAI*, pages 1163–1169, 2018.

[85] Shengchuan Zhang, Rongrong Ji, Jie Hu, Xiaoqiang Lu, and Xuelong Li. Face sketch synthesis by multidomain adversarial learning. *IEEE TNNLS*, 2018.

[86] Wei Zhang, Xiaogang Wang, and Xiaoou Tang. Lighting and pose robust face sketch synthesis. In *ECCV*, pages 420–433. Springer, 2010.

[87] Wei Zhang, Xiaogang Wang, and Xiaoou Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *IEEE CVPR*, pages 513–520, 2011.

[88] Yuqian Zhang, Nannan Wang, Shengchuan Zhang, Jie Li, and Xinbo Gao. Fast face sketch synthesis via kd-tree search. In *ECCV*, pages 64–77. Springer, 2016.

[89] Jiaxing Zhao, Ren Bo, Qibin Hou, Ming-Ming Cheng, and Paul L. Rosin. FLIC: fast linear iterative clustering with active search. *Computational Visual Media*, 4(4):333–348, 2018.

[90] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for RGBD salient object detection. In *IEEE CVPR*, 2019.

[91] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Jufeng Yang, and Ming-Ming Cheng. Edge-based network for salient object detection. In *IEEE ICCV*, 2019.

[92] Hao Zhou, Zhanghui Kuang, and Kwan-Yee K Wong. Markov weight fields for face sketch synthesis. In *IEEE CVPR*, pages 1091–1097, 2012.

[93] Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Li. Deep graphical feature learning for face sketch synthesis. In *IJCAI*, pages 3574–3580, 2017.

[94] Yizhe Zhu and Ahmed Elgammal. A multilayer-based framework for online background subtraction with freely moving cameras. In *IEEE ICCV*, 2017.

[95] Steven W Zucker, Allan Dobbins, and Lee Iverson. Two stages of curve detection suggest two styles of visual computation. *Neural computation*, 1(1):68–81, 1989.