Scoot: A Perceptual Metric for Facial Sketches
Supplemental Material

Deng-Ping Fan[1,2] ShengChuan Zhang[3] Yu-Huan Wu [1] Yun Liu [1]
Ming-Ming Cheng[1,*] Bo Ren[1] Paul L. Rosin[4] Rongrong Ji[3]
[1] TKLNDST, CS, Nankai University [2] IIAI [3] Xiamen University [4] Cardiff University

Abstract

In this document, we include additional materials related to the benchmark, proposed dataset, and results. All of the benchmark results will be released in our website.

- ***Benchmark.*** *We systematically assess 9 representative measures [1–6, 8–10] over two (CUFS [7] and CUFSF [11]) widely-used datasets, making it the first one largest-scale measure benchmark in this field. Extensive experiments on this study verify that our Scoot measure exceeds performance of prior works.*
- ***Dataset.*** *We collect two new human-ranked datasets,* i.e.*, RCUFS and RCUFSF (3.8k face sketches in total), making it the largest-scale subjectively verified datasets.*
- ***Results.*** *We provide visualization results on the proposed 3 meta-measures and the proposed human-ranked datasets.*

1. Benchmark

As shown in Fig. 1, it has witnessed the dramatic development of face sketch modeling, while the community long-term suffered from the lack of a standard representative perceptual metric based benchmark. To the best of our knowledge, this work is the first and the largest-scale benchmark for this issue.

2. Dataset

In Fig. 2, we present the example of the proposed human-ranked dataset. We use these datasets to examine the ranking consistency between current measures and human judgments. The detail can be found in Section 5.3 (manuscript). We refer the reader to the accompanying attachment ("*Proposed Datasets*") for the final human-ranked datasets.

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

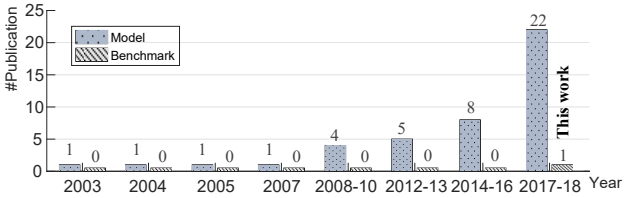


Figure 1: **Number of FSS papers** on top conferences (*i.e.*, CVPR, ICCV, ECCV), ACM TOG, and IEEE Trans. journals over the past 16 years.

3. Results

3.1. Meta-measure 1: Stability to Slight Resizing.

The first meta-measure specifies that the rankings of synthetic sketches should not change much with slight changes in the GT sketch. Therefore, we perform a minor 5 pixels downsizing of the GT by using nearest-neighbor interpolation.

Fig. 7 and Fig. 8 show the results: the narrower the band is, the more stable a measure is to slightly downsizing. We can see our Scoot measure achieves a significant improvement over the existing measures, such as GMSD [9], SSIM [8], FSIM [10], and VIF [4] measures in both the CUFS and CUFSF databases. Among the alternative texture measures, our measure is the best. These improvements are mainly because the proposed measure considers "block-level" statistics rather than "pixel-level".

One example shows in Fig. 3 which illustrates our Scoot measure is more robust than existing widely-used measures. For SSIM measure, it switches the ranking of LR and SSD when using GT or Resized-GT as reference. For other measures (VIF, FSIM), they also change the ranking in different ways. Only our measure keeps the original ranking.

3.2. Meta-measure 2: Rotation Sensitivity.

In real-wold situations, sketches drawn by artists may also have slight rotations compared to the original photographs. Thus, the proposed second meta-measure verifies the sensitivity of GT rotation for the evaluation measure. We did a slight counter-clockwise rotation (5^o) for each GT sketch.

Figure 2: Meta-measure 4. Sample images from our human ranked database. The first row is the GT sketch, followed by the first and second ranked synthesis result. We refer the reader to the accompanying attachment ("*Proposed Datasets*") for more details.

The sensitivity results are shown in Fig. 7 and Fig. 8. The thinner the band is, the better the measure performs. Our measure also significantly outperforms the current measures over the CUFS and CUFSF databases. We attribute the robust performance to the exploit of local region-level statistics rather than pixel-level statistics.

We show an example in Fig. 4. For SIM, VIF and FSIM, they change the ranking result. However, our measure correctly ranks these results and does not change their ranking when using different GT as reference.

3.3. Meta-measure 3: Content Capture Capability.

The third meta-measure describes that a good measure should assign a complete sketch generated by SOTA algorithm higher score than the sketches of only preserving incomplete strokes.

From Fig. 9 and Fig. 10, we observe a great improvement over the other measures in *CUFS* database. A slight improvement is also achieved for the *CUFSF* database.

Fig. 5 shows that our measure assigns the complete sketch generated by sate-of-the-art algorithm (*e.g.*, MRF [7] higher score than the sketch of only preserving light (*incomplete*) strokes.

3.4. Meta-measure 4: Human Judgment.

The fourth meta-measure (Jug) specifies that the ranking result according to an evaluation measure should agree with the human judgment.

For human judgment in Fig. 9 and Fig. 10, the proposed Scoot measure shows a great improvement over the best prior measure in *CUFS*. This improvement is due to our consideration of style similarity which human perception considers as an essential factor when evaluating sketches.

Various of examples in Fig. 6 demonstrate that our measure provide an reliable evaluation.

References

[1] J. Canny. A computational approach to edge detection. In *Readings in Computer Vision*, pages 184–203. Elsevier, 1987. 1, 3, 7, 8

[2] D. 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. 1, 3, 7, 8

[3] M. M. Galloway. Texture analysis using grey level run lengths. *NASA STI/Recon Technical Report N*, 75, 1974. 1, 3, 7, 8

[4] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE TIP*, 15(2):430–444, 2006. 1, 3, 4, 5, 6, 7

[5] H. R. Sheikh, A. C. Bovik, and G. De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE TIP*, 14(12):2117–2128, 2005. 1, 3

[6] I. Sobel. An isotropic 3× 3 image gradient operator. *Machine vision for three-dimensional scenes*, pages 376–379, 1990. 1, 3, 7, 8

Dataset	Meta-measure	Classical Measure IFC [5]	VIF [4]	GMSD [9]	FSIM [10]	SSIM [8]	**Scoot**	Improvement
CUFS	MM1↓	0.256	0.322	0.417	0.268	0.162	0.037	+77.2% = ((0.162 - 0.037)/0.162)
	ranking	3	5	6	4	2	1	
	MM2↓	0.189	0.236	0.210	0.123	0.086	0.025	+70.9% = ((0.086 - 0.025)/0.086)
	ranking	4	6	5	3	2	1	
	MM3↑	1.20%	43.5%	21.9%	14.2%	81.4%	95.9%	+14.5% = (95.9% - 81.4%)
	ranking	6	3	4	5	2	1	
	Jud↑	26.9%	44.1%	42.6%	50.0%	37.3%	76.3%	+26.3% = (76.3% - 50.0%)
	ranking	6	3	4	2	5	1	
CUFSF	MM1↓	0.089	0.111	0.259	0.151	0.073	0.012	+83.6% = ((0.073 - 0.012)/0.073)
	ranking	3	4	6	5	2	1	
	MM2↓	0.112	0.150	0.132	0.058	0.074	0.008	+86.2% = ((0.058 - 0.008)/0.058)
	ranking	4	6	5	2	3	1	
	MM3↑	3.07%	22.2%	63.6%	32.4%	97.4%	97.5%	+0.10% = (97.5% - 97.4%)
	ranking	6	5	3	4	2	1	
	Jud↑	25.4%	52.8%	58.6%	37.5%	36.8%	78.8%	+20.2% = (78.8% - 80.9%)
	ranking	6	3	2	4	5	1	
	overall ranking	6	5	4	3	2	1	

Dataset	Meta-measure	Alternative Texture Measure Canny [1]	GLRLM [3]	Sobel [6]	Gabor [2]	**Scoot**	Improvement
CUFS	MM1↓	0.086	0.111	0.040	0.062	0.037	+7.50% = ((0.04 - 0.037)/0.04)
	ranking	4	5	2	3	1	
	MM2↓	0.078	0.111	0.037	0.055	0.025	+32.4% = ((0.037 - 0.025)/0.037)
	ranking	4	5	2	3	1	
	MM3↑	33.7%	18.6%	0.00%	0.00%	95.9%	+62.2% = (95.9% - 33.7%)
	ranking	2	3	4	4	1	
	Jud↑	27.8%	73.7%	32.8%	72.2%	76.3%	+2.60% = (76.3% - 73.7%)
	ranking	5	2	4	3	1	
CUFSF	MM1↓	0.138	0.125	0.048	0.089	0.012	+75.0% = ((0.048 - 0.012)/0.048)
	ranking	5	4	2	3	1	
	MM2↓	0.146	0.079	0.044	0.043	0.008	+81.4% = ((0.043 - 0.008)/0.043)
	ranking	5	4	3	2	1	
	MM3↑	0.00%	64.6%	0.00%	19.3%	97.5%	+32.9% = (97.5% - 64.6%)
	ranking	4	2	4	3	1	
	Jud↑	0.10%	68.0%	52.6%	80.9%	78.8%	-2.10% = (78.8% - 80.9%)
	ranking	5	3	4	1	2	
	overall ranking	5	4	3	2	1	

Table 1: **Benchmarking results of 5 classical measures and 4 alternative textures on the *CUFS* and *CUFSF* datasets.** Darker color indicates better performance. These differences are all statistically significant at the $\alpha < 0.05$ level.

[7] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE TPAMI*, 31(11):1955–1967, 2009. 1, 2, 6

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 1, 3, 4, 5, 6, 7

[9] W. Xue, L. Zhang, X. Mou, and A. C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE TIP*, 23(2):684–695, 2014. 1, 3, 7

[10] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE TIP*, 20(8):2378–2386, 2011. 1, 3, 4, 5, 6, 7

[11] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *IEEE CVPR*, pages 513–520, 2011. 1

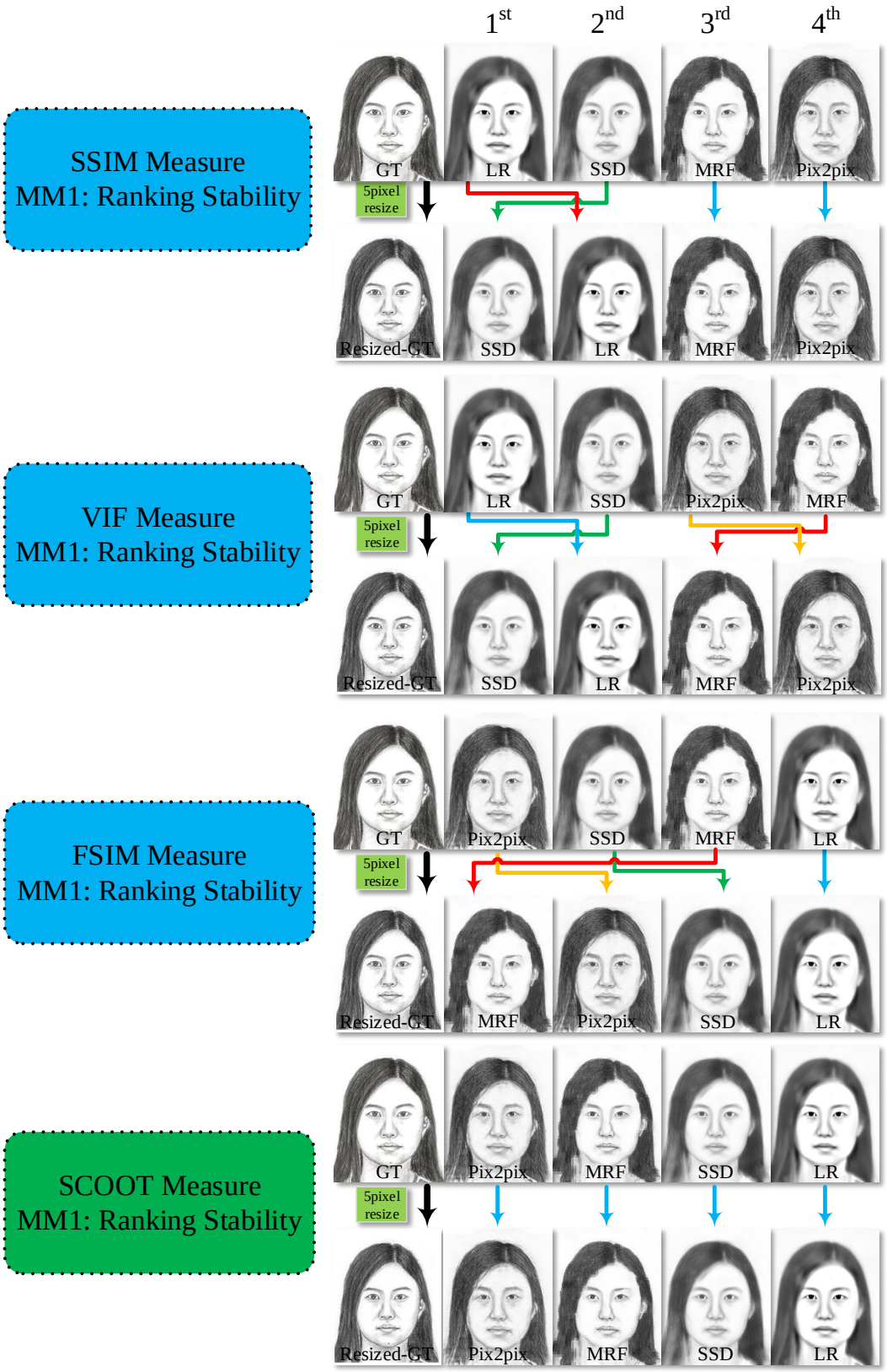


Figure 3: **Visual comparison of existing widely-used FSS measures (SSIM [8], FSIM [10], and VIF [4]) on meta-measure 1.** The experiment clearly shows that the proposed SCOOT measure is more stable to slightly resizing.

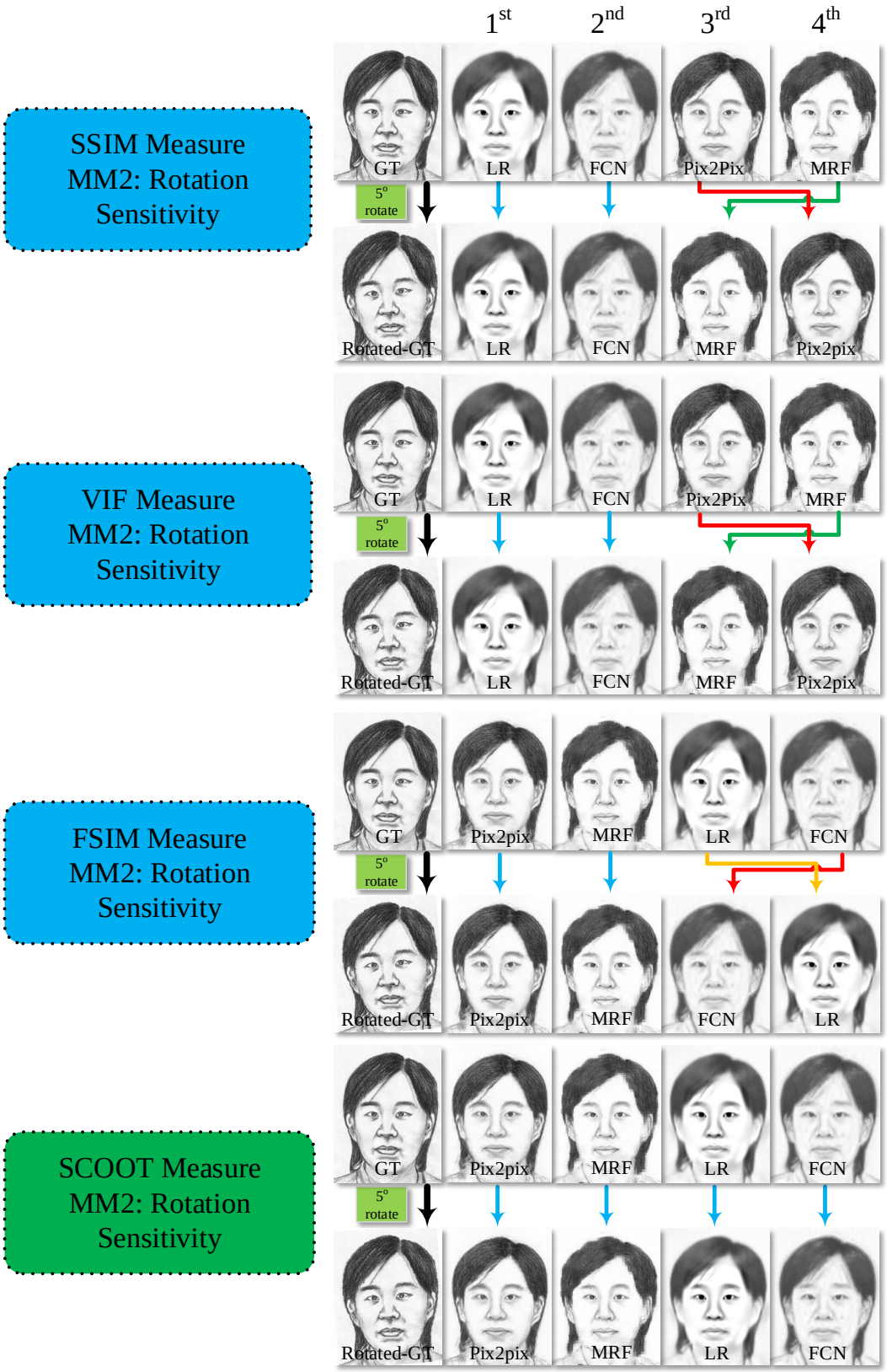


Figure 4: **Visual comparison of existing widely-used FSS measures (SSIM [8], FSIM [10], and VIF [4]) on meta-measure 2.** The experiment clearly demonstrates that the proposed SCOOT measure is less sensitive to minor rotation.

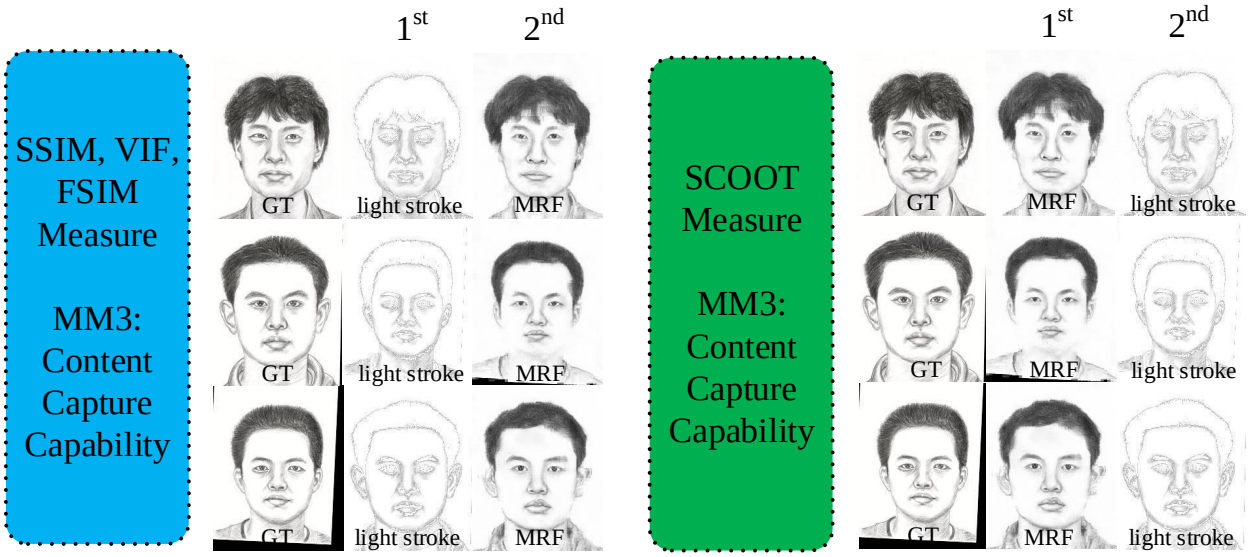


Figure 5: **Visual comparison of existing widely-used FSS measures (SSIM [8], FSIM [10], and VIF [4]) on meta-measure 3.** We conduct this experiment to verify the content capture capability of the measures. Our measure assigns the complete sketch generated by sate-of-the-art algorithm (*e.g.*, MRF [7] higher score than the sketch of only preserving light (*incomplete*) strokes.

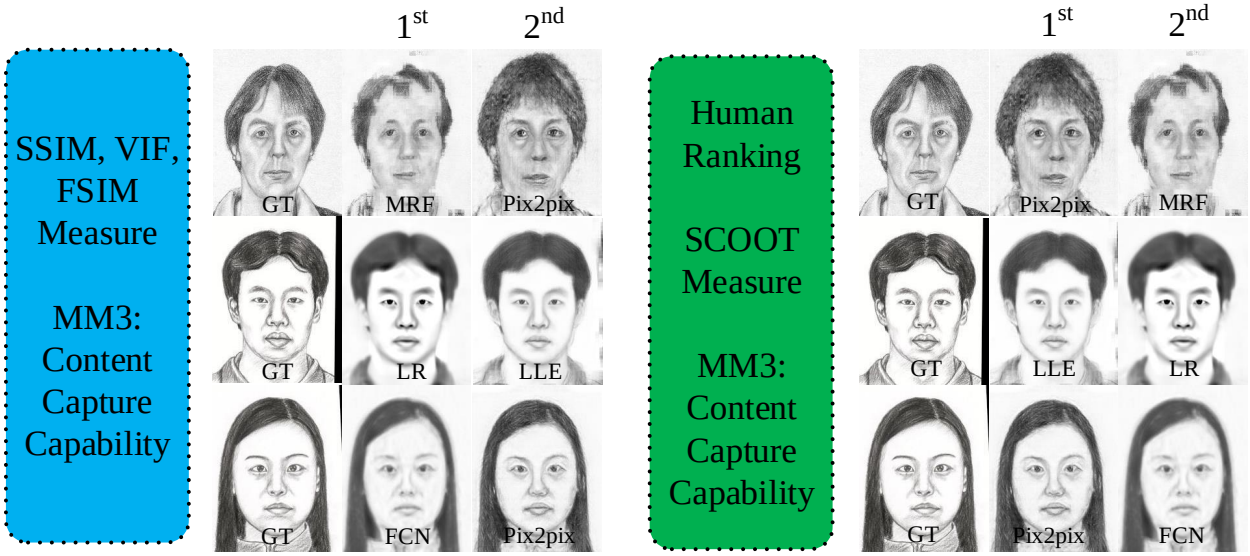
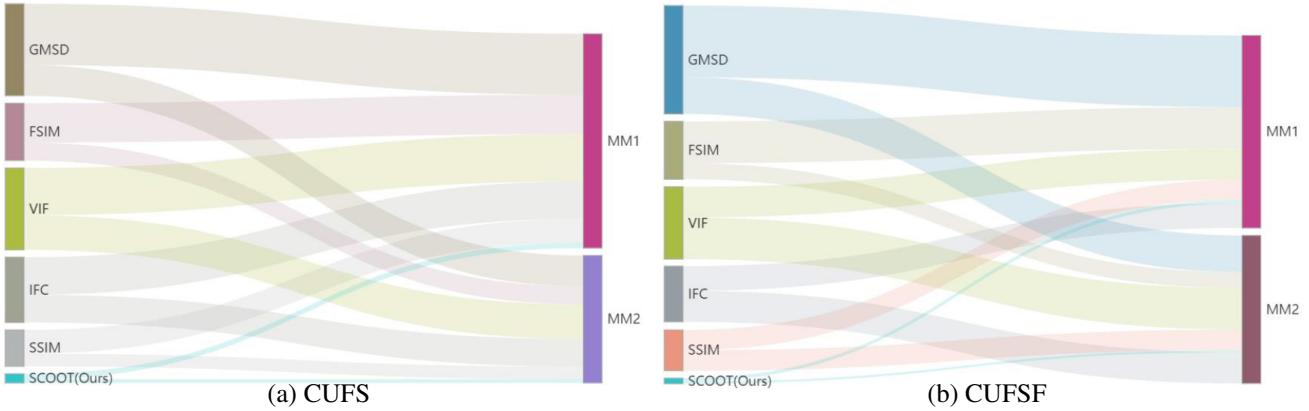
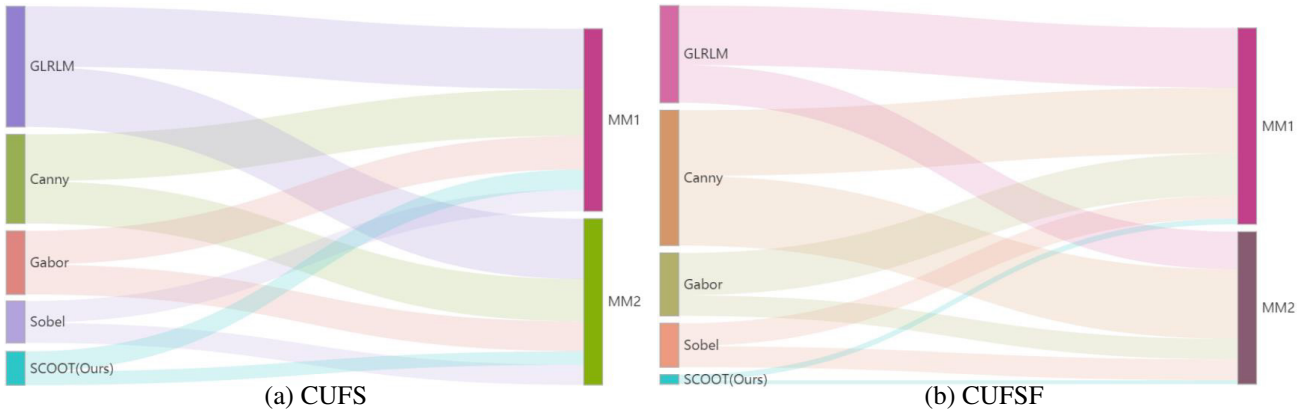


Figure 6: **Visual comparison of existing popular FSS measures (SSIM [8], FSIM [10], and VIF [4]) on meta-measure 4.** The proposed Scoot measure shows higher consistency with human judgment than previous methods on our collected datasets. Zoom-in for details.



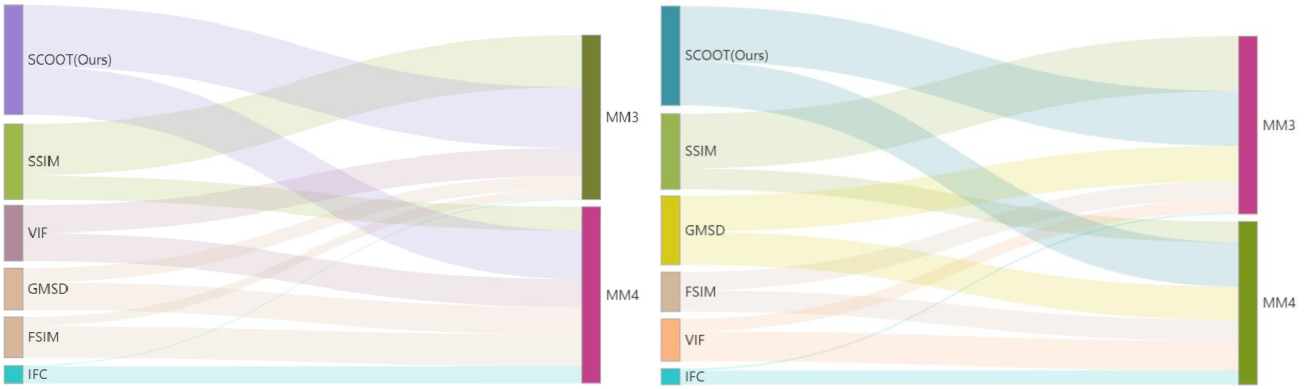
(a) CUFS (b) CUFSF

Figure 7: **Quantitative comparison of different measure (GMSD [9], SSIM [8], FSIM [10], VIF [4], and the proposed Scoot) on our meta-measure 1 & 2.** The narrower the color band is, the better the performance is. The proposed Scoot measure (in the last row) achieves the best performance. Zoom-in for details.



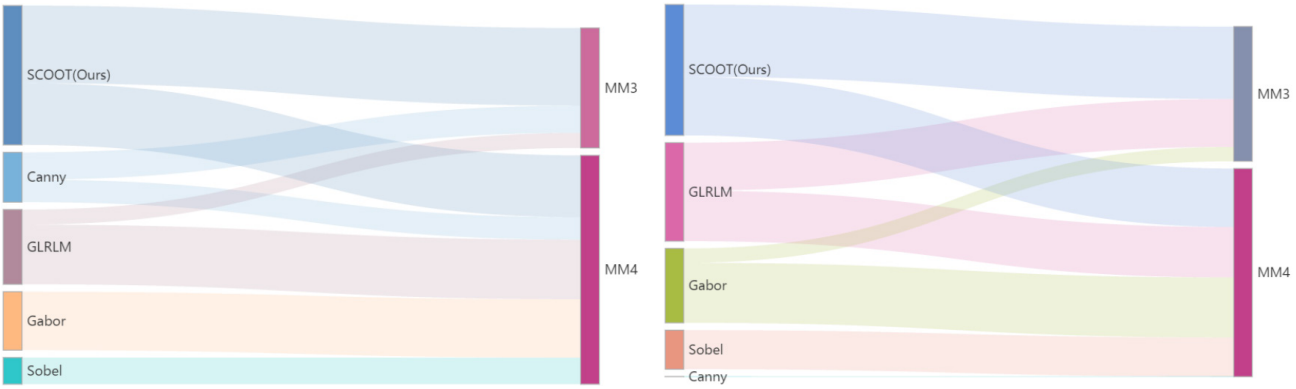
(a) CUFS (b) CUFSF

Figure 8: **Quantitative comparison of alternative texture (Canny [1], Sobel [6], GLRLM [3], Gabor [2] and the proposed Scoot) on our meta-measure 1 & 2.** The narrower the color band is, the better the performance is. Our Scoot measure (in the last row) achieves the preferable performance.



(a) CUFS (MM3) & RCUFS (MM4) (b) CUFSF (MM3) & RCUFSF (MM4)

Figure 9: **Quantitative comparison of different measure (GMSD [9], SSIM [8], FSIM [10], VIF [4], and the proposed Scoot) on our meta-measure 3 & 4.** The wider the color band is, the better the performance is. Our Scoot measure (1^{st} row) outperforms prior works. Zoom-in for details.



(a) CUFS (MM3) & RCUFS (MM4) (b) CUFSF (MM3) & RCUFSF (MM4)

Figure 10: **Quantitative comparison of alternative texture (Canny [1], Sobel [6], GLRLM [3], Gabor [2] and the proposed Scoot) on our meta-measure 3 & 4.** The wider the color band is, the better the performance is. Our measure present in first row provides the best performance in both MM3 and MM4.