Supplementary Materials for "Reference-Based Sketch Image Colorization using Augmented-Self Reference and Dense Semantic Correspondence"

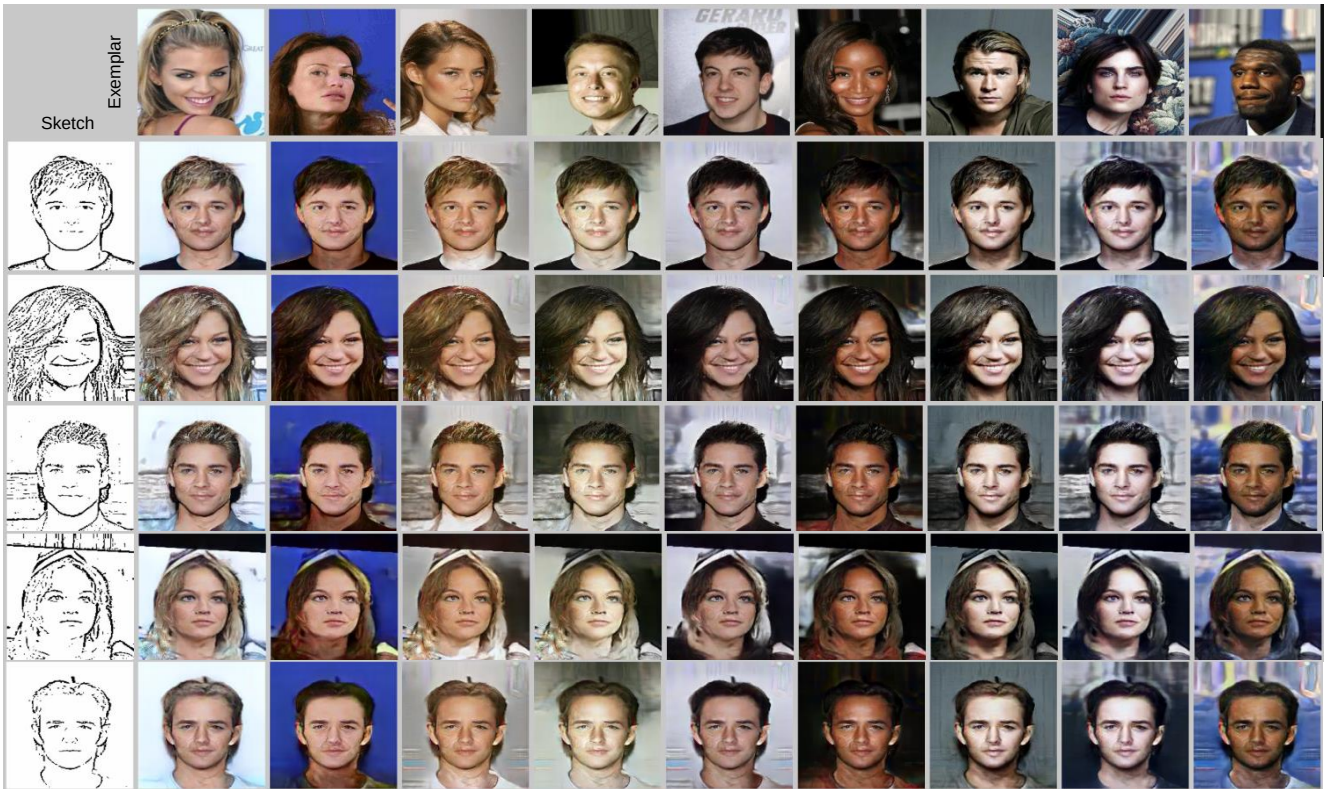

Figure 1: Qualitative results of our method on the CelebA [11] dataset. Each row has the same content while each column has the same reference.

6. Appendix

This supplementary document presents additional details of the paper. Section 6.1 discusses the effects of our spatially corresponding feature transfer mechanism with quantitative results. Section 6.2 demonstrates the human evaluation results that compare ours against baseline methods. Afterwards, Section 6.3 reports implementation details including network architectures, the processes of generating augmented-self reference images, and other training details. Comparisons to an existing study which shares similar network architectures are described in Section 6.4. Lastly, Section 6.5 addresses the case where a reference image does not exist. Qualitative results generated by our method are also shown throughout the document.

6.1. Effects of Aggregation Methods

The key assumption behind SCFT is that integrating spatially aligned reference features with content features would help reflect the exact color from the reference into corresponding positions. To prove this assumption, we compare our SCFT with two simple types of aggregation methods as shown in Fig. 2. Methods are as follows: (a) representations of the reference are simply added to the features of the content. (b) AdaIN [4] is utilized to transfer the style of reference by aligning the channel-wise mean and variance of content to match those of reference. (c) our SCFT module.

Qualitative comparison over three methods is shown in

	ImageNet			Human Face	Comics		Hand-drawn
Aggregation Method	Cat	Dog	Car	CelebA	Tag2pix	Yumi's Cells	Edges2Shoes
(a) Addition	78.47	103.73	55.80	51.94	47.72	47.67	117.15
(b) AdaIN	75.17	105.72	52.85	50.61	52.81	**45.36**	88.46
(c) SCFT (ours)	**74.12**	**102.83**	**52.23**	**47.15**	**45.34**	49.29	**78.32**

Table 1: FID scores [3] according to different aggregation methods.

Fig. 3. The leftmost column contains sketch and reference, while next three columns contain colorized images from (a), (b) and (c), respectively. Method (a) tends not to perfectly locate the corresponding regions and results in colorizing car with overly yellowish color, which is mainly background color in the exemplar. Method (b) totally ignores the spatially varying color information, thus colorizing with dominant color from the reference. (c) is superior to other methods in terms of color transferability to the corresponding position.

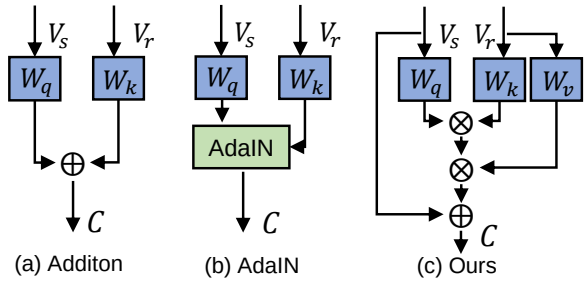

Figure 2: Diagram of three types of aggregation methods. (a) Addition block, (b) AdaIN [4] block, (c) Ours (SCFT)

Quantitative results comparing these methods are represented in Table 1. The network with SCFT module produces the most realistic results over most of the datasets. This is because the SCFT module properly aligns the corresponding local regions between the sketch and the reference image by using the attention matrix $\mathcal{A}$. On the other hand, the method (a) and (b) are not capable of aligning the local features of the reference with those of the sketch, resulting in low FID scores.

In Yumi's Cells [14] dataset, however, the SCFT module produced worse FID score than the others. The potential reason we assume is that the sketch and the reference we randomly pair for the inference time often contain different types of objects, e.g., Yumi (a human) and cells (non-human), which may have negatively impacted the colorization output.

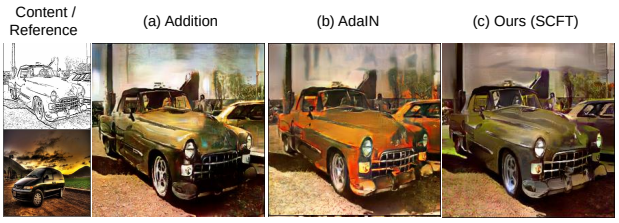

Figure 3: A qualitative example obtained from three different aggregation methods as shown in Figure 2.

6.2. User Study

We conduct two different human evaluation on the colorization outputs over various datasets. First, we randomly select ten sets of images per dataset, which contain the generated images from our method and other baselines. Second, we also randomly select ten sets of images for every dataset, and those contain the images obtained from the model trained with triplet loss, L_1-loss and no supervision for correspondence, respectively. For both cases, participants with no prior knowledge in this work are asked to rank them in terms of two types of questions sequentially as follows:

- **Overall Colorization Quality and Realism**
  How natural does the colorized image look? This question requires users to evaluate the overall quality of the generated colorization given an input sketch. The generated image should be perceptually realistic without any artifacts or color bleeding across sketches.

- **Detailed Reflection of Reference**
  How well is the colors of the reference image is reflected to a given sketch part by part? This question asks users to determine whether the particular color from a reference is injected into the corresponding regions in the sketch. For example, given an comics character image with green hair wearing a blue shirt as a reference, the generated output is expected to contain these colors at its corresponding hair and clothing part, respectively.

As seen in Fig. 15 and 16, superior measures indicate that our approach generates both more realistic and more faithfully colorized image than other methods. For both

question type 1 and 2, it can be observed that our approach achieves the rank 1 votings more than 50% over all the dataset we adopt for user study. When asked the first question on Comics domain dataset including Tag2pix [8] and Yumi's Cell [14], Style2Paints [12] perform realistic generation quality comparable to our method with a small gap in top 1 rate. This notable measure is obtained as Style2Paints [12] is a adept baseline especially on comic domain. However, the difference in top 1 rate increases as the users are asked to choose based on faithful colorization performance. The results demonstrate that our model utilizes the right color from the reference, which results in both realistic and exquisitely colorized output.

The results in Fig. 17 demonstrates that the model trained with triplet loss obtains more realistic and faithfully colorized outputs than with L_1-loss or no loss. Furthermore, along with the explanation of similarity-based triplet loss in Section 3.4 of the paper, these results support that the supervision for semantic correspondence with the L_1-loss leads to the inferior colorization performance even compared to the model without any supervision.

6.3. Implementation Details

This section provides the implementation details of our model, complementary to Section 3.5 of the paper.

Augmented-Self Reference Generation To automatically generate a sketch image from an original color image, We utilize a widely-used algorithm called XDoG [17]. The outputs, however, often involves superfluous edges, so in order to suppress them, we apply Gaussian blurring ($\sigma = 0.7$) to the original images before extracting sketches. The appearance transformation $a(\cdot)$ adds randomly sampled value from a uniform distribution on [-50, 50] to each of the RGB channels of the original image.

Encoder Our generator G contains two types of encoder, E_s and E_r. Both of them share the same architecture shown in Table 2, except for the number of input channels of the first layer, where E_s takes a single-channel, binarized sketch input while E_r takes a three-channel, RGB reference image. We utilize the an average pooling function for downsampling φ in Section 3.3 of the paper.

Resblocks We place four stacked residual blocks [2] with a kernel size of 3 and a stride of 1. Batch normalization [6] follows each convolutional block, and ReLU is used as the activation function.

Discriminator We adopt our discriminator architecture as PatchGAN [7]. We utilize the LSGAN [13] objective for the stable training.

Training Details For all the experiments, our network is trained using Adam optimizer [9] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We set an initial learning rate for the generator as 0.0001 and that for the discriminator as 0.0002. We train the model for the first 100 epochs using the same learning rate, and then we linearly decay it to zero until the 200 epochs. We set the margin value $\gamma = 12$ for our triplet loss (Eq. 5 in the paper). The batch size is set as 16. The parameters of all our models are initialized according to the normal distribution which has a mean as 0.0 and a standard deviation as 0.02.

Layer	Encoder
L1	Conv(I:C,O:16,K:3,P:1,S:1), Leaky ReLU:0.2
L2	Conv(I:16,O:16,K:3,P:1,S:1), Leaky ReLU:0.2
L3	Conv(I:16,O:32,K:3,P:1,S:2), Leaky ReLU:0.2
L4	Conv(I:32,O:32,K:3,P:1,S:1), Leaky ReLU:0.2
L5	Conv(I:32,O:64,K:3,P:1,S:2), Leaky ReLU:0.2
L6	Conv(I:64,O:64,K:3,P:1,S:1), Leaky ReLU:0.2
L7	Conv(I:64,O:128,K:3,P:1,S:2), Leaky ReLU:0.2
L8	Conv(I:128,O:128,K:3,P:1,S:1), Leaky ReLU:0.2
L9	Conv(I:128,O:256,K:3,P:1,S:2), Leaky ReLU:0.2
L10	Conv(I:256,O:256,K:3,P:1,S:1), Leaky ReLU:0.2

Table 2: The network architecture of Encoder E. Conv denotes a convolutional layer. I, O, K, P, and S denote the number of input channels, the number of output channels, a kernel size, a padding size, and a stride size, respectively.

Baselines We exploit Sun [16] and Style2Paints [12] as the sketch image colorization methods, Huang [2018] [5], and Lee [10] as the image translation methods and Huang [2017] [4] as the style transfer method as our baselines. For Style2Paints [12], we generate the images based on the publicly available Style2Paints V3 in a similar manner to Tag2pix [8]. For the other methods, we utilize the officially available codes to colorize images after training them on our datasets.

6.4. Comparison to Zhang *et al.* (2019) [18].

In this section, we discuss the detailed comparison between our method and Zhang *et al.*. These two works have similarity in that they both exploit geometric distortion for data augmentation and semantic correspondence module for color guidance. However the significant difference of our model against Zhang *et al.* lies in (1) direct supervision of semantic correspondence and (2) generalized attention module.

Direct supervision Our model directly supervises the attention module via a triplet loss, which enables the optimization of the attention module in an end-to-end manner. This fully trainable encoder encourages to generate plausible results over a wide range of datasets from real-world photos to comic images, as show in Fig. 4 of the paper and Fig. 9. In contrast, Zhang *et al.* requires a pre-trained, already reliable attention module, which is only indirectly supervised via a so-called contextual loss. According to Geirhos *et al.* (2019)

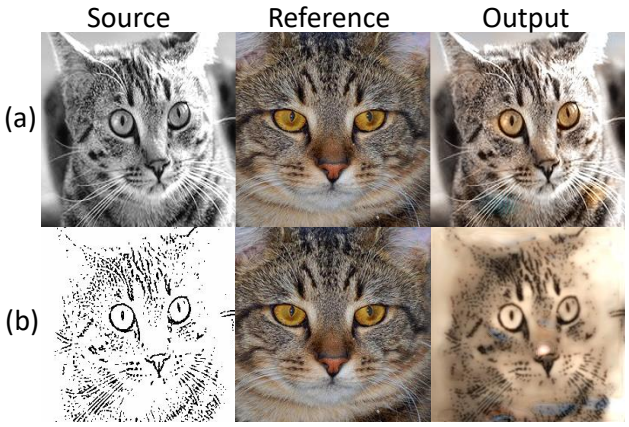

Figure 4: Qualitative results of our Zhang *et al.* [18] given gray-scale source image (row (a)) and sketch image (row (b)). In contrast to the output in the row (a), output in (b) fails to colorize the eyes with the color from the reference and spreads the yellow color over the face.

[1], the features extracted from the ImageNet pre-trained encoder may be severely degraded for a sketch image due to large domain shifts. In this sense, Zhang *et al.*'s work may not be easily applicable to sketch image colorization tasks, and the examples of failure case are shown in Fig. 4. We reimplemented the code of Zhang *et al.*, trained and tested the model over cat dataset. As this baseline exploits the ImageNet pre-trained encoder, row (a) shows that it produces the plausible colorized output given gray-scale source image. However, when given information scarce sketch image (row (b)), it fails to obtain the dense correspondence with the reference image, resulting in degraded output.

Generalized attention module Inspired by the self-attention module in the Transformer networks, our attention module involves different query, key, and value mappings for flexibility, while Zhang *et al.* use a relatively simple module. More importantly, in terms of value vectors, Zhang *et al.* uses only raw color values, but ours uses all the available low- to high- level semantic information extracted from multiple layers. In this respect, ours is capable of transferring significantly richer contextual information than just low-level color information.

6.5. Colorization without reference.

Our main scope is focused on the colorization task with a reference available, but we can easily extend our method for no-reference cases by occasionally providing a zero-filled image as a reference to the networks during the training time. We feed the zero-filled image to our model as a reference with a ratio of 9:1 at the training time. As shown in Fig. 5, we confirm that our network still generates a reasonable quality of colorization output at test time. In this

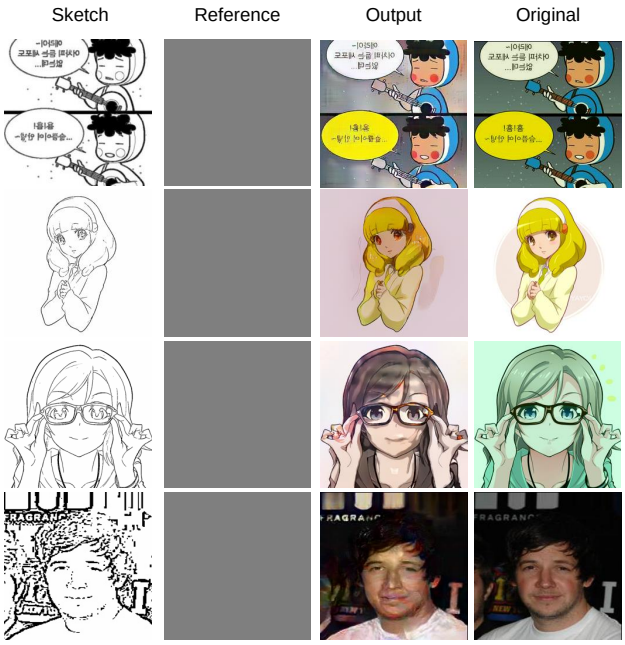

Figure 5: A qualitative example when there is no reference image. Our model takes the first column image (sketch) as a target and the second column image (zero-filled reference) to synthesize the third column image (output). The results of first row, second-to-third rows, last row are obtained from our model trained for Yumi's Cells [14], Tag2pix [8], and CelebA [11], respectively.

case, the zero-filled reference image does not have any information to guide. Therefore, the model is encouraged to synthesize an output image with colors that often appear in trainset conditioned on the sketch image. We recall that the main goal of this work is not restricted to generating the original image.

References

[1] Geirhos et al. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ICLR*, 2019. 4

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pages 6626–6637, 2017. 2

[4] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 1, 2, 3

Figure 6: Qualitative results of our method on the Tag2pix [8] dataset.

[5] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018. 3

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, page 448–456, 2015. 3

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 3, 6

[8] Hyunsu Kim, Ho Young Jhoo, Eunhyeok Park, and Sungjoo Yoo. Tag2pix: Line art colorization using text tag with secat and changing loss. In *ICCV*, pages 9056–9065, 2019. 3, 4, 5

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3

[10] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 2020. 3

[11] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 1, 4

[12] lllyasviel. style2paints. https://github.com/lllyasviel/style2paints, 2018. [Online; accessed 22-03-2018]. 3

[13] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017. 3

[14] NaverWebtoon. Yumi's cells. https://comic.naver.com/webtoon/list.nhn?titleId=651673, 2019. [Online; accessed 22-11-2019]. 2, 3, 4, 11

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 7, 10

[16] Tsai-Ho Sun, Chien-Hsun Lai, Sai-Keung Wong, and Yu-Shuen Wang. Adversarial colorization of icons based on contour and color conditions. In *MM*, pages 683–691, 2019. 3

[17] Holger WinnemöLler, Jan Eric Kyprianidis, and Sven C Olsen. Xdog: an extended difference-of-gaussians compendium including advanced image stylization. *Computers & Graphics*, 36(6):740–753, 2012. 3

[18] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *CVPR*, pages 8052–8061, 2019. 3, 4

Figure 7: Qualitative results of our method on the Edges→Shoes [7] dataset.

Figure 8: Qualitative results of our method on the ImageNet [15] dataset.

Figure 9: Qualitative comparisons with the baselines on the Tag2pix dataset.

Figure 10: Qualitative results of our method on the Edges→Shoes dataset.

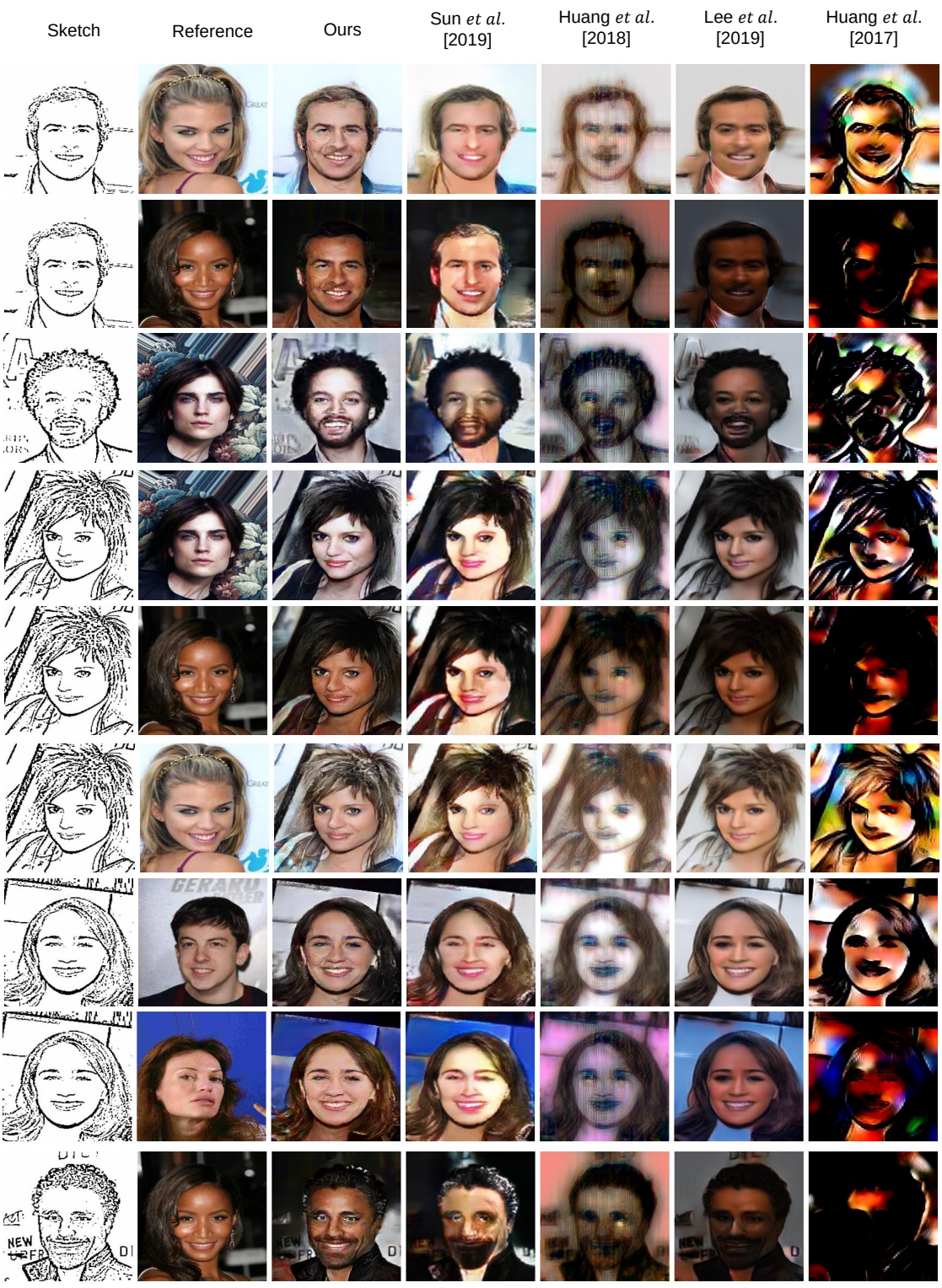

Figure 11: Qualitative comparisons with baselines on the CelebA dataset.

Figure 12: Qualitative comparisons with baselines on the ImageNet [15] dataset.

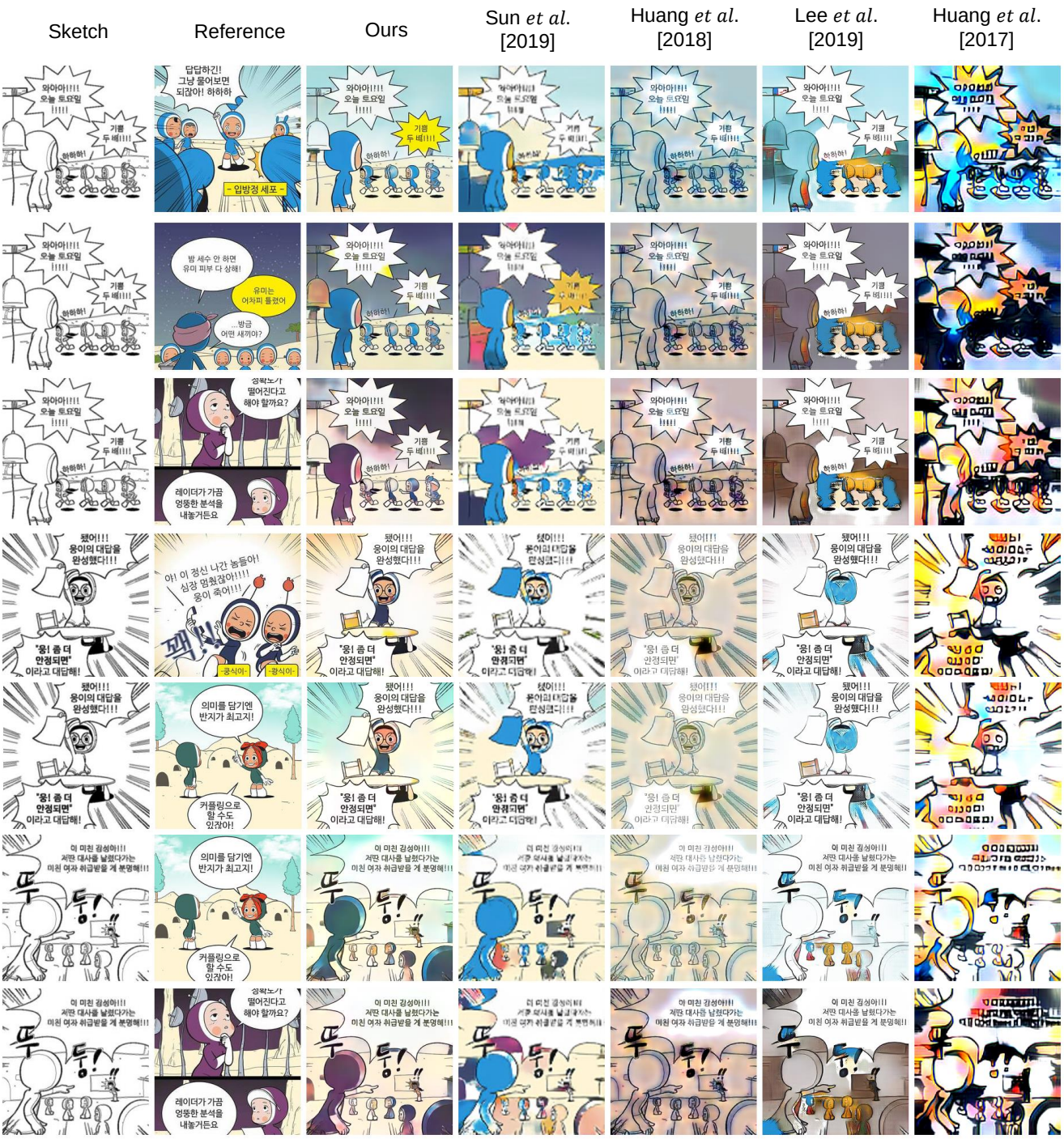

Figure 13: Qualitative comparisons with baselines on the Yumi's Cells [14] dataset.

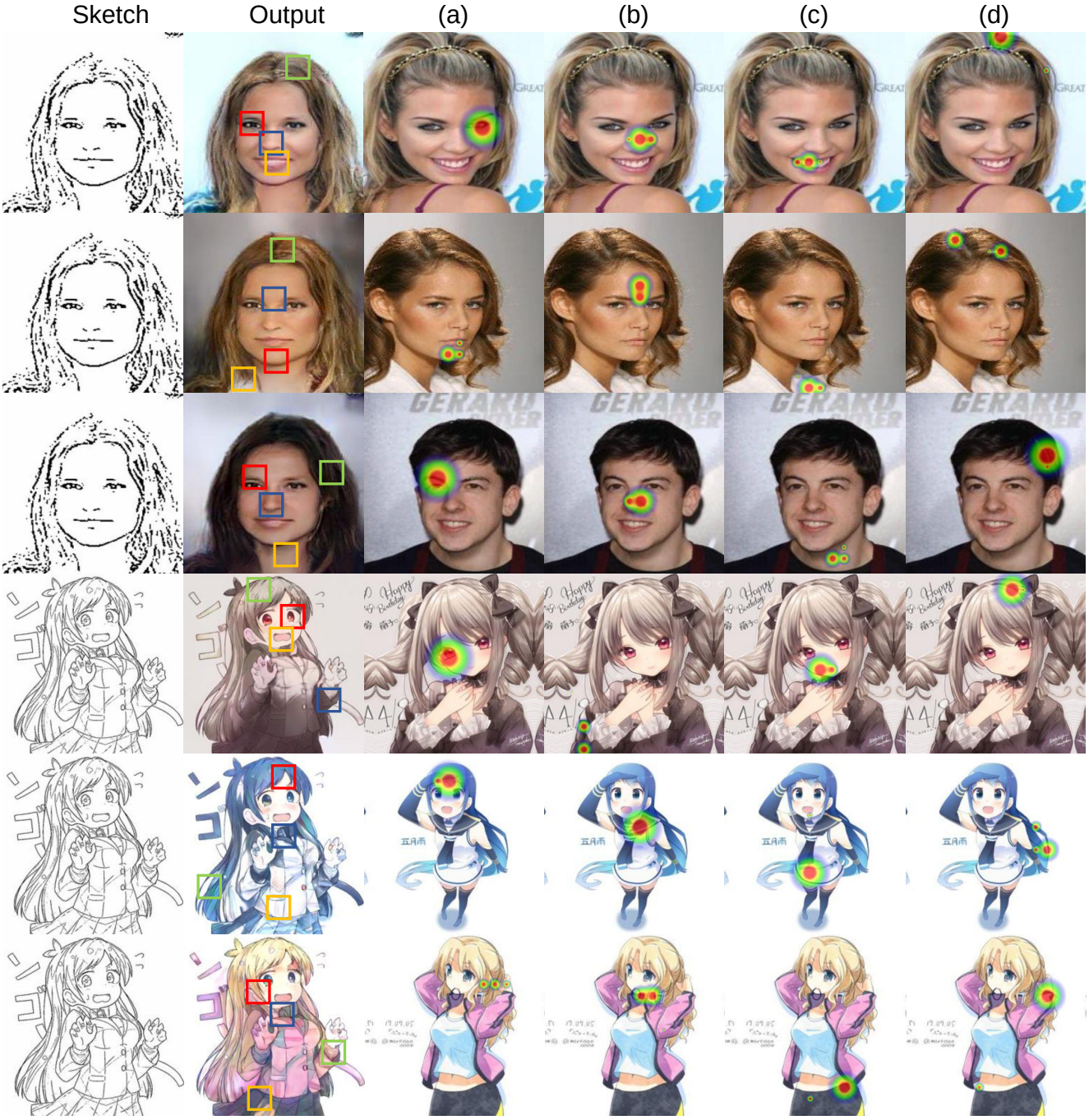

Figure 14: The visualization of attention maps on CelebA and Tag2pix dataset. The colored squares on the second column indicate the query region and corresponding key regions are highlighted in the next four columns. The different color of square means the different query region, and each red, blue, yellow, and green corresponds with the column (a), (b), (c), and (d), respectively.

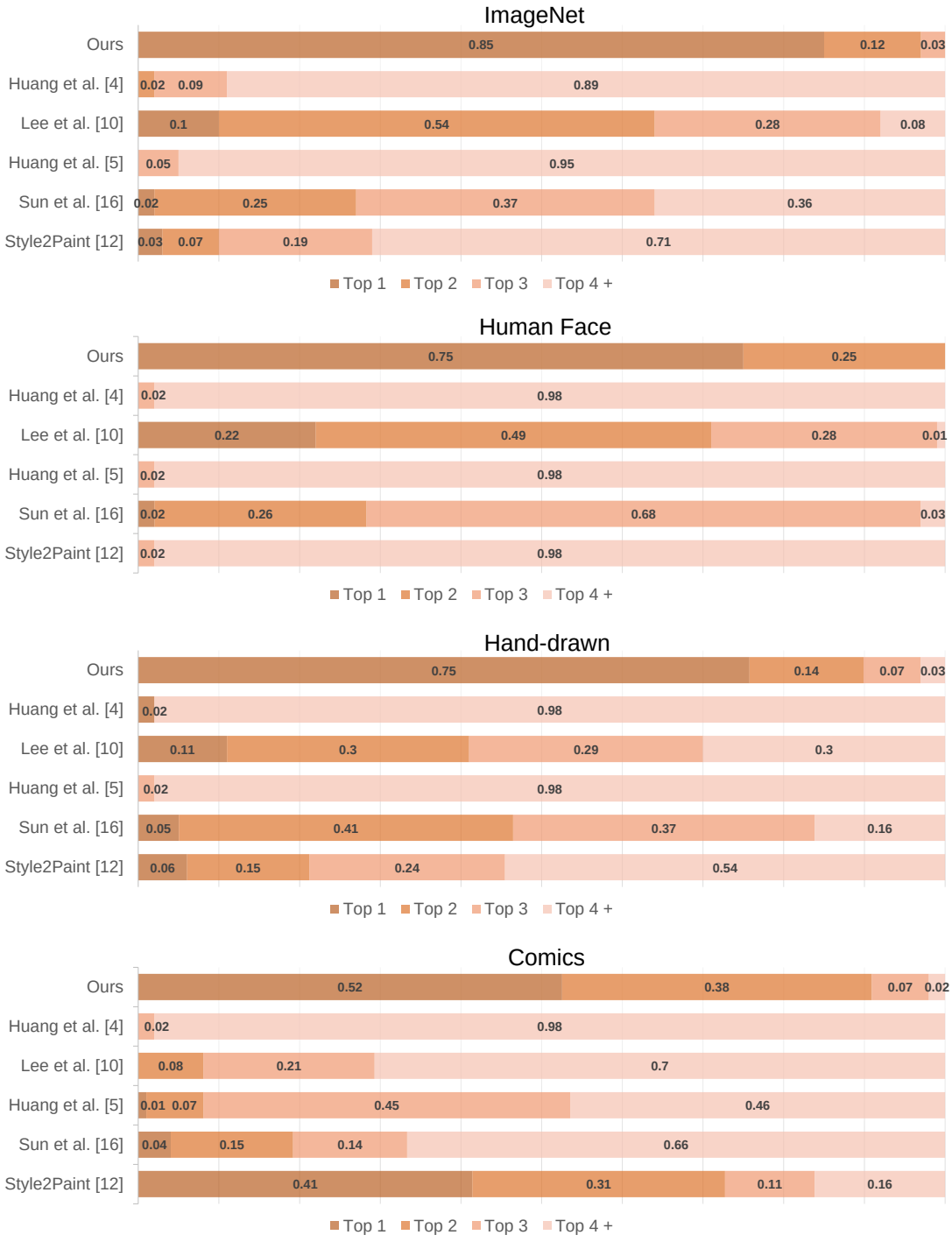

Figure 15: The results of the user study for comparison between our model and existing baselines. Question type 1: Overall Colorization Quality and Realism.

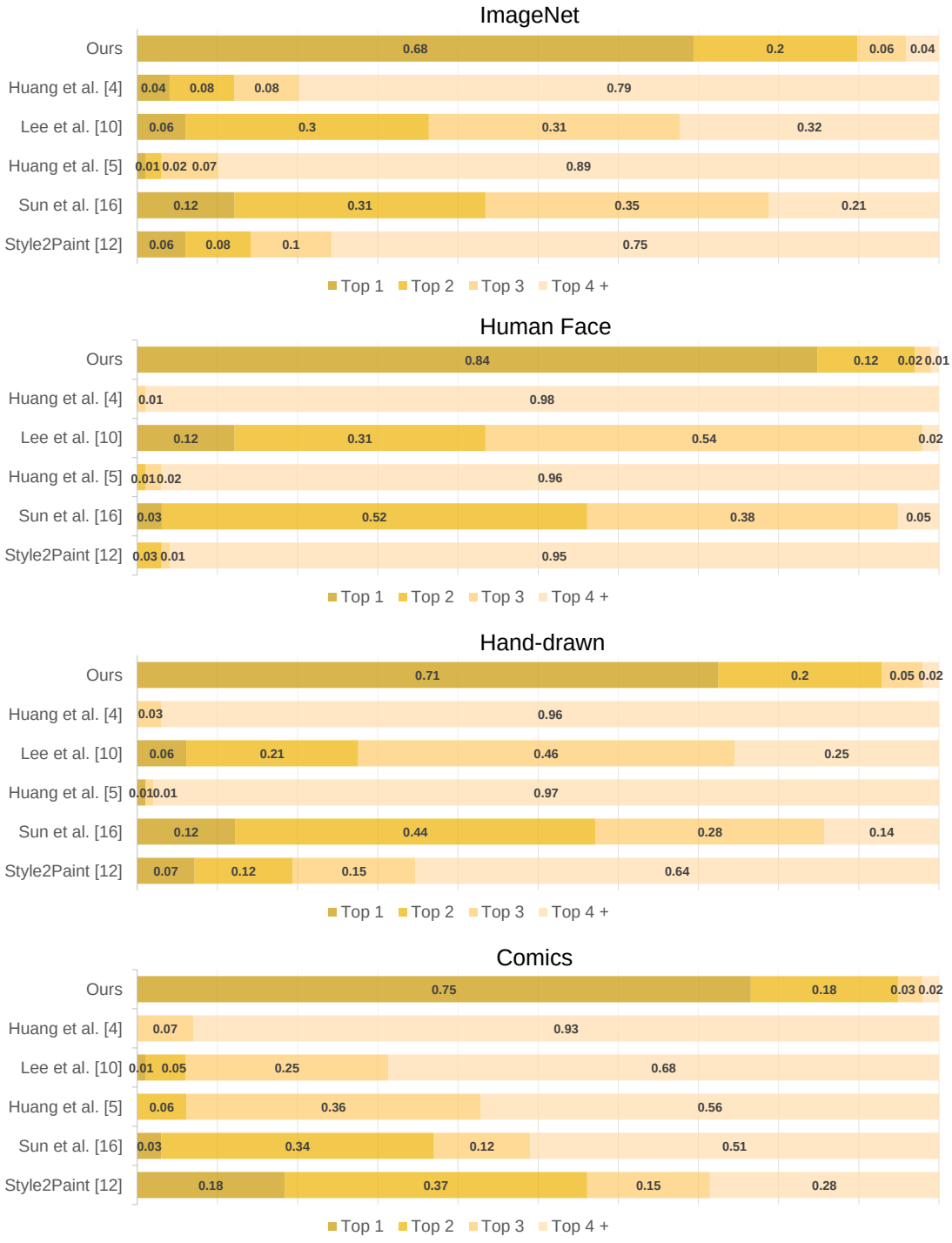

Figure 16: The results of the user study for comparison between our model and existing baselines. Question type 2: Detailed Reflection of Reference.

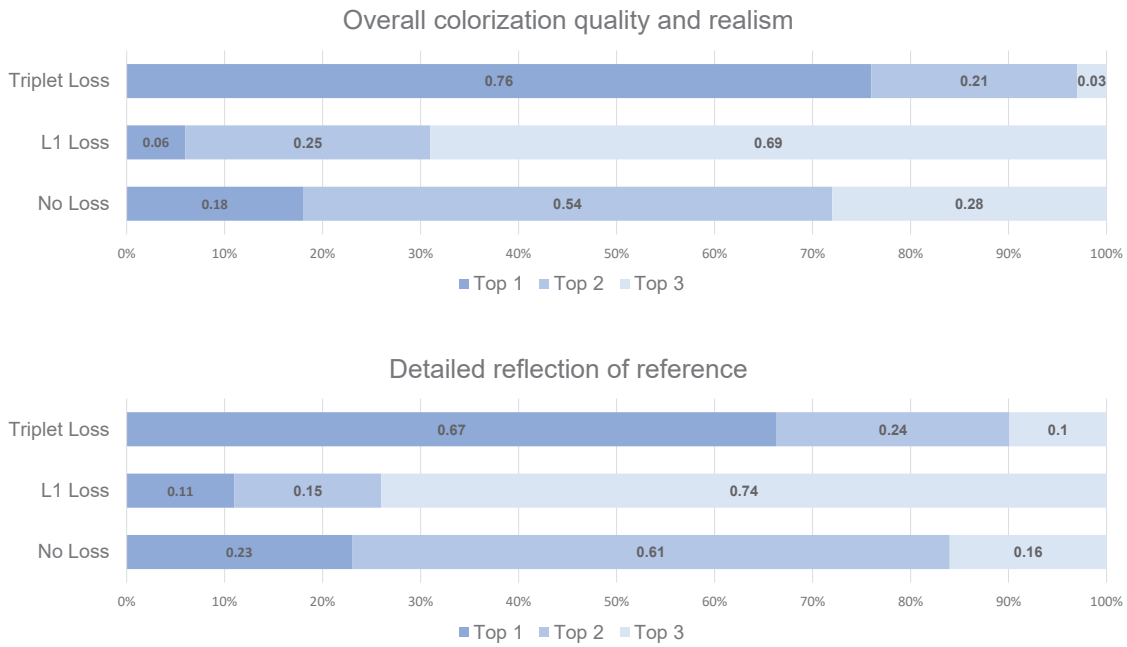

Figure 17: The results of the user study for comparison between model with triplet loss, L_1-loss and no loss. The percentages are averaged over all the datasets.