1. Content

This is supplementary material of the paper **Stylized Face Sketch Extraction via Generative Prior with Limited Data.** In this document, we provide details of our experiments and results.

The table of contents are as follows:

2. Method Details

2.1. Sketch Generator

G_{sketch} consists of a total of 8 sequential deep fusion modules which accept 18 deep features as input. Deep features pass through a 1×1 convolution whose results are concatenated to be fused with previously fused features except for the first and the last deep features. The first deep feature is directly upsampled and fused while the last deep feature is directly concatenated to previously fused features followed by going through a convolution layer to output a sketch.

2.2. Discriminator

StyleSketch contains three types of discriminator, D_{full}, D_{in}, and D_{out}. All three discriminators share the same patch discriminator architecture with residual layers. All the discriminators are trained independently without sharing weights so that each discriminator learns to discriminate the inner face, hair, and full face correctly.

3. Dataset(SKSF-A)

In our SKSF-A dataset, we provide facial attributes. The attributes can increase the utility of the face dataset or make the dataset practically useful for real-world applications. The attributes can also help the researchers to find and cluster similar images for automatic handling of the data. Therefore, we record the median RGB value of the skin, lip, and eye colors for our SKSF-A dataset. See Fig. 2 for examples.

In addition, we also classified the image types as African, Asian, Caucasian, and Unknown based on human perception. Two different artists, who drew the sketch images, classified these categories solely based on visual information. If both participants did not agree on a specific category, the image was classified as Unknown. Please refer to Fig. 3 for examples.

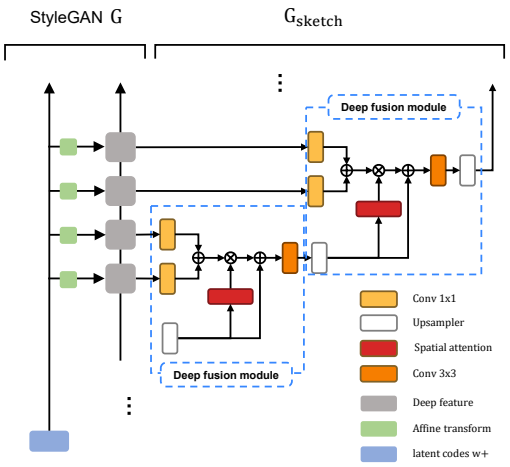
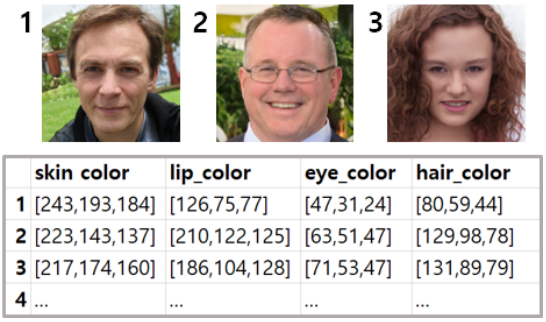


Figure 1: *G_{sketch} in detail. G_{sketch} consists of a total of 8 deep fusion modules.*



	skin color	lip_color	eye_color	hair_color
1	[243,193,184]	[126,75,77]	[47,31,24]	[80,59,44]
2	[223,143,137]	[210,122,125]	[63,51,47]	[129,98,78]
3	[217,174,160]	[186,104,128]	[71,53,47]	[131,89,79]
4	...	...	...	...

Figure 2: *The median RGB value of the skin, lip, and eye colors are recorded in attributes.*

The detailed explanation of each style is as follows:

- **Style 1**: Thin lines are used to visualize the face shape. Eyes and eyebrows are empty with no color filling. Details of hair and wrinkles are well expressed.
- **Style 2**: Very thick lines are used to visualize the face shape. While eyes are filled with black color, eyebrows are left empty and represented with a few lines. Details of hair and wrinkles are well expressed.
- **Style 3**: Thick lines are used to visualize the face shape. While eyes are filled with black color, eyebrows are left empty. Details of hair are expressed well, but wrinkles are represented with less details.
- **Style 4**: Thick lines are used to visualize the face shape. While eyes are filled with black color, eyebrows are left empty. Drawing style is similar to Japanese animation characters. For example, details are restrained and some

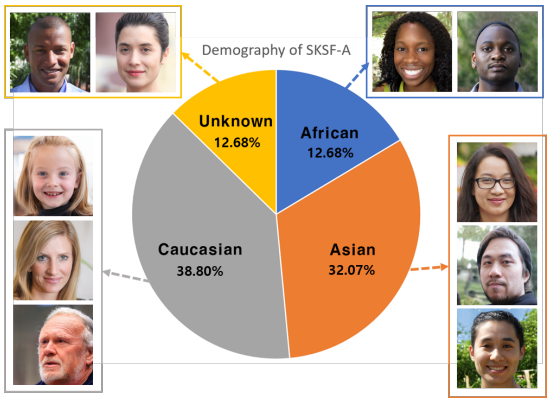


Figure 3: *The demography of SKSF-A was categorized into African, Asian, Caucasian, and Unknown groups. These labels were assigned by two artists based on visual information.*

symbolic parts of the face such as eyes and eyebrows are exaggerated.

- **Style 5**: High details are represented with clear lines. Eyes and eyebrows are filled with black color. Details of hair and wrinkles are expressed well.
- **Style 6**: Very thick lines and black color fillings are heavily used to express the shading style of drawing.
- **Style 7**: Very rough lines are drawn by a pencil brush tool. This style represents the most primitive sketching with no line simplification.

4. Experiment

4.1. Experiment Details

As mentioned in the main paper, the baseline methods and our sketch generator G_{sketch} were trained with SKSF-A, APDrawing [YXL*20], and CUFS [WT08a]. The settings used in this study were based on the official code provided by the authors and information obtained from their respective papers. Learn-to-draw [CDI22] was trained with 100 epochs with the batch size of four. Ref2sketch [ASK*22] was trained with 1500 epochs with the batch size of four. APdrawing++ [YXL*20] was trained with 200 epochs with the batch size of one. Mind the gap (MTG) [ZAFW21] was trained with 600 iterations, and JoJoGAN [CF22] was trained with 300 iterations. Few-shot image generation (Few-shot) [OLL*21] and RSSA [XLW*22] were both trained with 5000 iterations with the batch size of four. The comparison with the baseline methods and ablation study were evaluated with two different metrics LPIPS [ZIE*18] and FID [HRU*17] with the image resolution of 256×256. All experiments with a limited number of data were trained with the same train pairs while the experiments with full data were conducted with the training data which were provided in the original papers [YLLR19, WT08b].

Table 1: *Quantitative results of comparison with additional baselines.*

Method	LPIPS	FID
Ours	0.1910	87.67
ControlNet	0.4147	186.88
Dreambooth	0.3594	123.04
Textual Inversion	0.4426	241.44
DualstyleGAN	0.2628	112.17
Semi-ref2sketch	0.3081	206.08

4.2. Full Quantitative Results of Experiments

In this section, we present the complete table from our comparison with baseline models and our ablation study. Table 1 demonstrates that our method, trained with a limited number of data points (16), outperform most other methods under the same conditions and even some methods trained with a full dataset. Table 3 shows that *Ours* surpassed other methods in most styles. There were no instances where a comparative method achieved the best LPIPS score in more than one style, whereas *Ours* achieves the best LPIPS score four times. Additionally, *Ours* secured the best FID score on four occasions, while only *Ours* with the last half of the features attained the top score twice. The qualitative results are detailed in the main paper.

4.3. Additional Comparison

We conducted an additional comparison to other methods that generate sketch images by finetuning a network [YJLL22, RLJ*23, GAA*22] or training a network to match a target domain [SLC*22, ZA23].

ControlNet [ZA23] is a method that trains additional modules on top of Stable Diffusion [RBL*22] to translate a conditional image into the target image domain. Dreambooth [RLJ*23] finetunes a Stable Diffusion model to generate personalized images. Textual Inversion [GAA*22] optimizes an additional text embedding to generate a personalized concept for a style or object. DualstyleGAN [YJLL22] transfers the style of a face image by characterizing both the content and style of the face using an intrinsic style path. The extrinsic style path allows the model to modulate both color and complex structural styles hierarchically, enabling precise emulation of the style from the example. Semi-ref2sketch [SAN23] extracts a sketch from an image, which imitates the style of the reference sketch. The model is trained with unpaired data in a semi-supervised manner.

For the experiment, we trained the methods with a limited number of data (16 pairs). The training details are the same as the original setting described at the official code repository of the methods or in published papers. For Dreambooth [RLJ*23] and Textual Inversion [GAA*22], we used DDIM inversion [SME20] to invert the source image to the latent code of Stable Diffusion.

As shown in Fig. 4 and Table 1, these methods produced

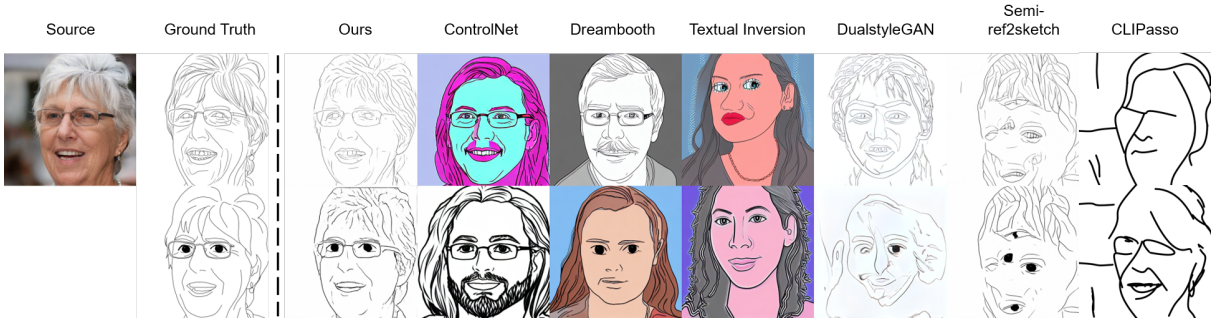


Figure 4: *Experimental results from additional baselines.*

Table 2: *Quantitative results of comparison with baselines. The best LPIPS scores are annotated in* ***bold*** *while the best FID scores are underlined.*

Sketch styles	SKSF-A(1)	SKSF-A(2)	SKSF-A(3)	SKSF-A(4)	SKSF-A(5)	SKSF-A(6)	SKSF-A(7)	APDrawing	CUFS
Methods	**LPIPS**↓/FID↓								
Ours with limited(16)	**0.1669**/62.24	0.1552/90.12	0.1665/76.53	**0.1405**/75.11	0.1956/175.68	**0.1025**/86.54	**0.1112**/149.01	0.2078/94.13	0.0513/91.54
RSSA with limited	0.2231/198.09	0.1876/168.23	0.2026/138.86	0.1789/129.10	0.2111/183.57	0.1615/122.60	0.1515/232.31	0.2531/133.87	0.1019/173.39
Few-shot with limited	0.2487/138.63	0.2389/157.19	0.2393/116.46	0.2125/116.49	0.2570/204.09	0.1854/114.73	0.1400/229.58	0.2697/154.13	0.0911/160.73
JoJoGAN with limited	0.1789/132.61	0.1611/125.93	0.1727/97.36	0.1466/87.30	0.1973/105.87	0.1046/123.45	0.1176/188.08	0.2340/111.18	0.0588/138.46
MTG with limited	0.2166/233.77	0.1920/192.30	0.1966/174.30	0.1724/128.35	0.2311/141.06	0.1315/117.53	0.1356/205.67	0.2252/102.32	0.0708/127.37
Learn-to-draw with limited	0.2212/260.91	0.2053/294.25	0.2218/311.29	0.1888/278.71	0.2217/353.55	0.1619/209.86	0.1436/277.72	0.2659/181.49	0.0577/78.36
APDrawing++ with limited	0.1945/125.09	0.1645/92.47	0.1782/116.77	0.1619/162.56	0.1838/189.84	0.1294/104.16	0.1220/182.46	0.2016/81.28	0.0518/82.62
JoJoGAN with one	0.1946/141.32	0.1781/104.86	0.1957/116.84	0.1634/85.82	0.2078/177.44	0.1188/107.34	0.1237/219.68	0.2239/114.42	0.0604/133.68
MTG with one	0.2150/212.81	0.1891/181.43	0.2048/154.94	0.1780/131.29	0.2072/246.72	0.1332/122.38	0.1401/231.39	0.2202/80.53	0.0591/134.41
Ours with four	0.1987/154.58	0.1733/202.58	0.1997/173.34	0.1534/118.85	0.2337/154.44	0.1161/118.03	0.1226/193.49	0.2181/123.55	0.0532/106.85
JoJoGAN with four	0.1859/115.96	0.1711/119.76	0.1847/97.51	0.1586/91.11	0.2053/113.51	0.1131/118.92	0.1206/196.41	0.2173/102.62	0.0585/128.90
MTG with four	0.2153/198.69	0.1916/192.98	0.2084/169.29	0.1722/130.97	0.2233/153.48	0.1260/112.76	0.1399/217.35	0.2393/98.55	0.0769/120.78
Ours with eight	0.1837/92.66	0.1596/121.56	0.1735/99.79	0.1551/85.09	0.2087/95.71	0.1125/94.84	0.1180/193.59	0.2165/131.66	0.0541/106.34
JoJoGAN with eight	0.1793/129.59	0.1623/125.49	0.1771/95.85	0.1521/80.62	0.1981/100.46	0.1085/115.88	0.1205/195.89	0.2152/105.99	0.0576/147.59
MTG with eight	0.2143/220.65	0.1916/200.23	0.2042/156.24	0.1801/138.63	0.2198/132.23	0.1208/101.97	0.1385/204.48	0.2283/96.75	0.0725/118.54
Ref2sketch with full	0.1998/129.96	0.1595/56.85	0.1828/80.58	0.1597/81.45	0.1836/141.48	0.1345/73.98	0.1423/143.15	0.2547/82.27	0.0735/106.83
Learn-to-draw with full	0.2082/166.18	0.1795/127.41	0.1988/143.14	0.1753/161.79	0.2059/271.74	0.1520/162.97	0.1362/245.96	0.2685/145.70	**0.0488**/71.46
APDrawing++ with full	0.1838/101.21	**0.1529**/70.15	**0.1646**/72.04	0.1543/85.34	**0.1696**/157.11	0.1177/77.73	0.1141/153.64	**0.1809**/62.41	0.0674/99.67

poor results which do not reflect the desired style or identity. ControlNet [ZZLZ15] generated face images that exhibit the same pose as the source image, but it failed to generate the desired style and identity represented by the source image. Dreambooth [RLJ*23] and Textual Inversion [GAA*22] could not generate images that imitate the original sketch style and the identity of the source image. While diffusion-based methods produced the results that mimicked the style of the sketch to a certain degree, many of the generated results were colorized artistic images that could not depict the original identity correctly. Both DualstyleGAN [YJLL22] and Semi-ref2sketch [SAN23] performed a source-to-target image domain translation, but the shapes of the generated images were severely distorted. This might be due to mode collapse occurred when using a limited number of data. Based on these qualitative examples, we chose not to consider these methods as our primary baseline for a precise performance comparison against our method.

In addition, we conducted a qualitative comparison with CLIPasso [VPB*22]. CLIPasso is a method that effectively sketches objects into different levels of abstraction with semantic guidance. Fig. 4 shows that because it utilizes vector representation, CLIPasso failed to generate dense sketches or control beyond its representation.

4.4. Comparison on Semantic Editing

We also conducted experiments to compare the semantic editing capabilities of StyleSketch and Instruct pix2pix [BHE23]. Instruct pix2pix is a method trained to edit images based on instructions by a large language model [BMR*20] and a prompt-based image editing model [HMT*22]. For the comparison, we used the pre-trained Instruct pix2pix model provided by the author with the prompt "Make him/her sad". In contrast, *Ours* was edited using the latent direction for less smile as described in the main paper. While *Ours*

Table 3: *Quantitative results of ablation study. The best LPIPS scores are annotated in bold while the best FID scores are underlined.*

Sketch styles	SKSF-A(1)	SKSF-A(2)	SKSF-A(3)	SKSF-A(4)	SKSF-A(5)	SKSF-A(6)	SKSF-A(7)	APDrawing	CUFS
Methods	**LPIPS**↓/FID↓								
Ours	**0.2155**/66.97	**0.1868**/93.46	0.2029/80.50	0.1742/65.69	**0.2341**/73.06	0.1524/88.24	0.1551/134.11	**0.2118**/87.05	0.0623/43.47
Ours W/O L_1 initialization	0.2300/80.85	0.1876/83.76	0.2068/85.86	**0.1658**/70.58	0.2442/84.66	0.1768/87.59	0.1612/211.53	0.2162/103.85	0.0648/46.12
Ours W L_1 until the end	0.2308/95.68	0.1913/122.91	0.2194/109.35	0.1836/104.26	0.2641/124.29	0.1729/110.68	**0.1524**/181.03	0.2190/126.98	0.0632/45.81
Ours W/O attention	0.2197/71.83	0.1879/86.52	0.2004/76.17	0.1785/73.19	0.2351/80.38	0.1599/90.18	0.1573/170.03	0.2233/141.53	**0.0621**/45.34
Ours W one adversarial loss	0.2329/80.75	0.1935/91.08	0.2048/79.80	0.1843/67.58	0.2444/93.23	0.1785/74.52	0.1552/204.05	0.2180/111.29	0.0632/44.23
Ours W/O L_{clip}	0.2296/89.17	0.1908/96.55	0.2010/71.54	0.1833/72.65	0.2524/78.88	0.1729/137.27	0.1605/173.08	0.2153/98.12	0.0636/46.80
Ours DFM W/O f_{i-1}	0.2933/159.33	0.1977/119.42	0.2275/106.43	0.2011/140.73	0.2824/112.27	0.1790/83.62	0.1830/233.41	0.2213/89.07	0.0727/37.07
Ours W first half features	0.3018/338.30	0.2373/208.25	0.3184/272.23	0.2656/210.17	0.3054/232.20	0.1942/108.88	0.1793/197.23	0.3058/125.20	0.0993/43.84
Ours W last half features	0.2686/177.37	0.1880/67.29	0.2012/70.56	0.1807/71.11	0.24722/78.79	**0.1438**/85.90	0.1552/187.16	0.2876/99.04	0.0960/48.39
Ours W middle ten features	0.2419/102.97	0.1892/83.63	**0.1997**/75.37	0.1777/78.47	0.2369/81.87	0.1591/93.36	0.1587/207.90	0.2780/85.76	0.0951/49.94

successfully edited the source sketch to reduce the smile, Instruct pix2pix either failed to modify the mouth (first row) or was unable to edit the source at all (second row) as shown in Fig. 5.

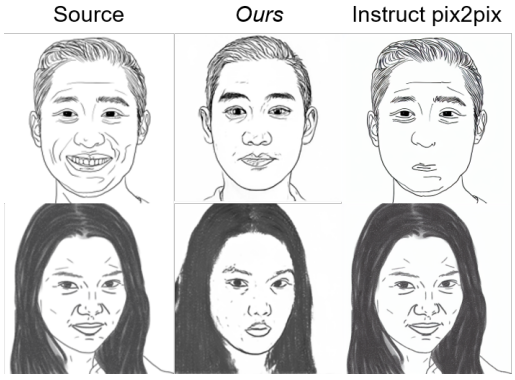


Figure 5: *Semantic editing comparison with Instruct pix2pix.*

5. Additional Results

We show additional sketch extraction results produced by StyleSketch trained with 16 paired data. Fig. 7 - 9, show sketches of all 9 styles extracted from the source images.

6. Examples in Perceptual Study

A total of 53 participants were asked to make 20 different comparisons and determine which sketch style appeared most similar to the target sketch. See examples of the comparisons in Figs. 10-14.

7. Analysis of Data Scaling

We conducted an additional study to investigate how the number of data used to train our model affects the performance. We varied the number of training data from 4 to 32. To evaluate the effect, we utilized LPIPS and FID scores

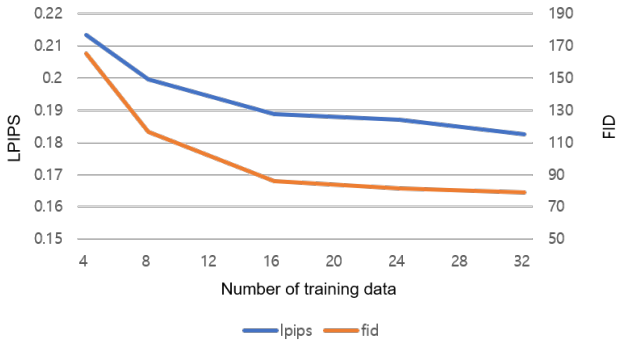


Figure 6: *The number of training data pairs used and their corresponding model performance.*

with the SKSF-A dataset as before. The results, as reported in Fig. 6, indicate that the number of data used up to 16 significantly improved the evaluation scores. However, increasing the number of data beyond 16 led to marginal changes. We determined that 16 pairs of data are sufficient for training our method. Here, we present the results of our experiments on scaling training data for seven different sketch styles. The experiment was conducted with 4, 8, 16, and 32 pairs of data. Styles 1-7 correspond to seven different styles from the SKSF-A dataset. Visual results are shown in Fig. 15. See Fig. 16 for the reported quantitative results.

References

[ASK*22] Ashtari A., Seo C. W., Kang C., Cha S., Noh J.: Reference based sketch extraction via attention mechanism. *ACM Transactions on Graphics (TOG) 41*, 6 (2022), 1–16.

[BHE23] Brooks T., Holynski A., Efros A. A.: Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 18392–18402.

[BMR*20] Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., et al.: Language models are few-shot learners. *Advances in neural information processing systems 33* (2020), 1877–1901.

[CDI22] Chan C., Durand F., Isola P.: Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 7915–7925.

[CF22] Chong M. J., Forsyth D.: Jojogan: One shot face stylization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI* (2022), Springer, pp. 128–152.

[GAA*22] Gal R., Alaluf Y., Atzmon Y., Patashnik O., Bermano A. H., Chechik G., Cohen-Or D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022).

[HMT*22] Hertz A., Mokady R., Tenenbaum J., Aberman K., Pritch Y., Cohen-Or D.: Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

[HRU*17] Heusel M., Ramsauer H., Unterthiner T., Nessler B., Hochreiter S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems 30* (2017).

[OLL*21] Ojha U., Li Y., Lu J., Efros A. A., Lee Y. J., Shechtman E., Zhang R.: Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 10743–10752.

[RBL*22] Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 10684–10695.

[RLJ*23] Ruiz N., Li Y., Jampani V., Pritch Y., Rubinstein M., Aberman K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 22500–22510.

[SAN23] Seo C. W., Ashtari A., Noh J.: Semi-supervised reference-based sketch extraction using a contrastive learning framework. *ACM Transactions on Graphics (TOG) 42*, 4 (2023), 1–12.

[SLC*22] Seo J., Lee G., Cho S., Lee J., Kim S.: Midms: Matching interleaved diffusion models for exemplar-based image translation. *arXiv preprint arXiv:2209.11047* (2022).

[SME20] Song J., Meng C., Ermon S.: Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[VPB*22] Vinker Y., Pajouheshgar E., Bo J. Y., Bachmann R. C., Bermano A. H., Cohen-Or D., Zamir A., Shamir A.: Clipasso: Semantically-aware object sketching. *ACM Trans. Graph. 41*, 4 (jul 2022). URL: https://doi.org/10.1145/3528223.3530068, https://doi.org/10.1145/3528223.3530068 doi:10.1145/3528223.3530068.

[WT08a] Wang X., Tang X.: Face photo-sketch synthesis and recognition. vol. 31, IEEE, pp. 1955–1967.

[WT08b] Wang X., Tang X.: Face photo-sketch synthesis and recognition. *IEEE transactions on pattern analysis and machine intelligence 31*, 11 (2008), 1955–1967.

[XLW*22] Xiao J., Li L., Wang C., Zha Z.-J., Huang Q.: Few shot generative model adaption via relaxed spatial structural alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 11204–11213.

[YJLL22] Yang S., Jiang L., Liu Z., Loy C. C.: Pastiche master: exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 7693–7702.

[YLLR19] Yi R., Liu Y.-J., Lai Y.-K., Rosin P. L.: Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 10743–10752.

[YXL*20] Yi R., Xia M., Liu Y.-J., Lai Y.-K., Rosin P. L.: Line drawings for face portraits from photos using global and local structure based gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence 43*, 10 (2020), 3462–3475.

[ZA23] Zhang L., Agrawala M.: Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).

[ZAFW21] Zhu P., Abdal R., Femiani J., Wonka P.: Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. *arXiv preprint arXiv:2110.08398* (2021).

[ZIE*18] Zhang R., Isola P., Efros A. A., Shechtman E., Wang O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 586–595.

[ZZLZ15] Zhang K., Zhang L., Lam K.-M., Zhang D.: A level set approach to image segmentation with intensity inhomogeneity. *IEEE transactions on cybernetics 46*, 2 (2015), 546–557.

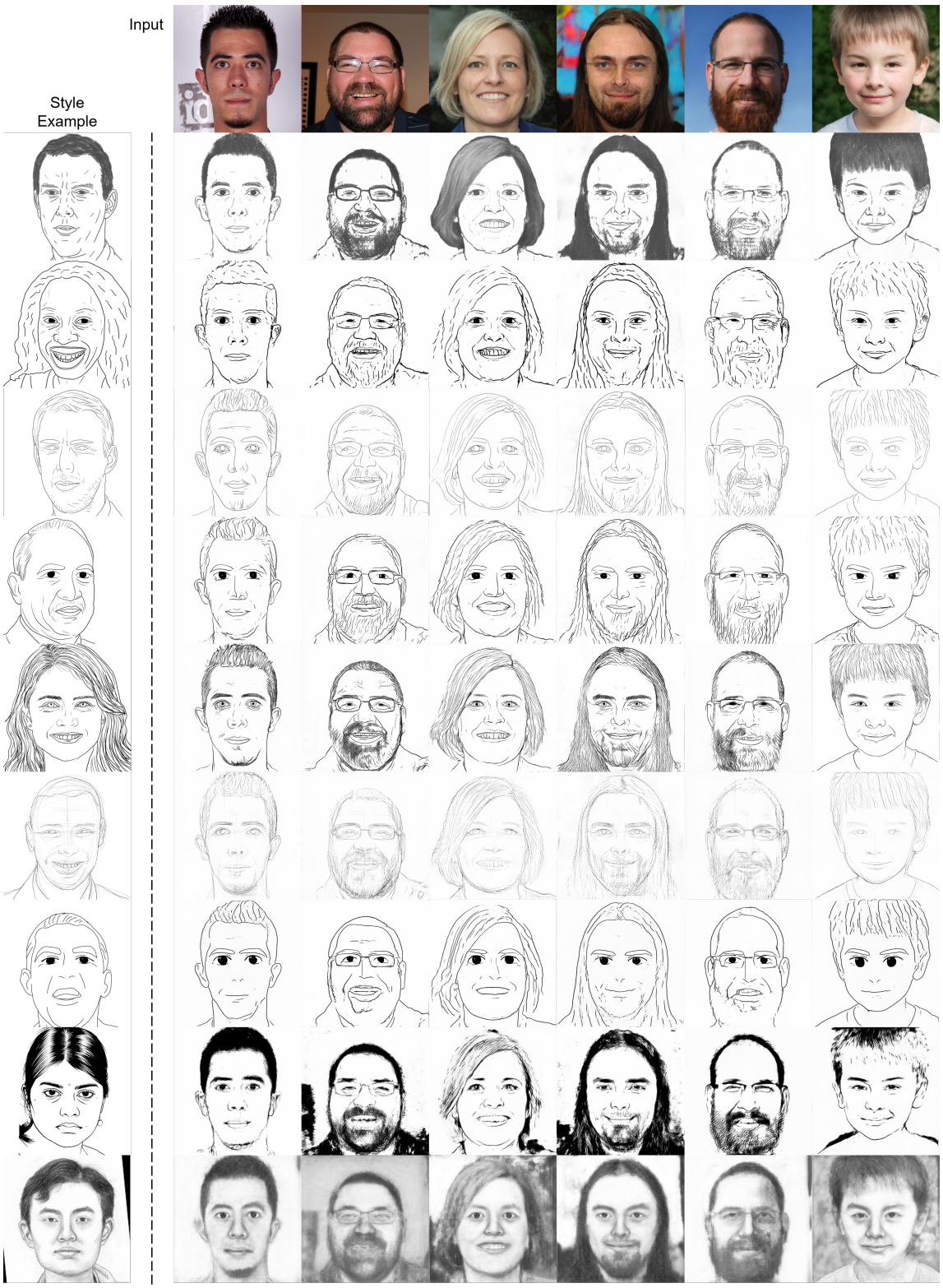


Figure 7: *Sketches extracted in nine different styles from the input images. Each style is trained with 16 pairs of data.*

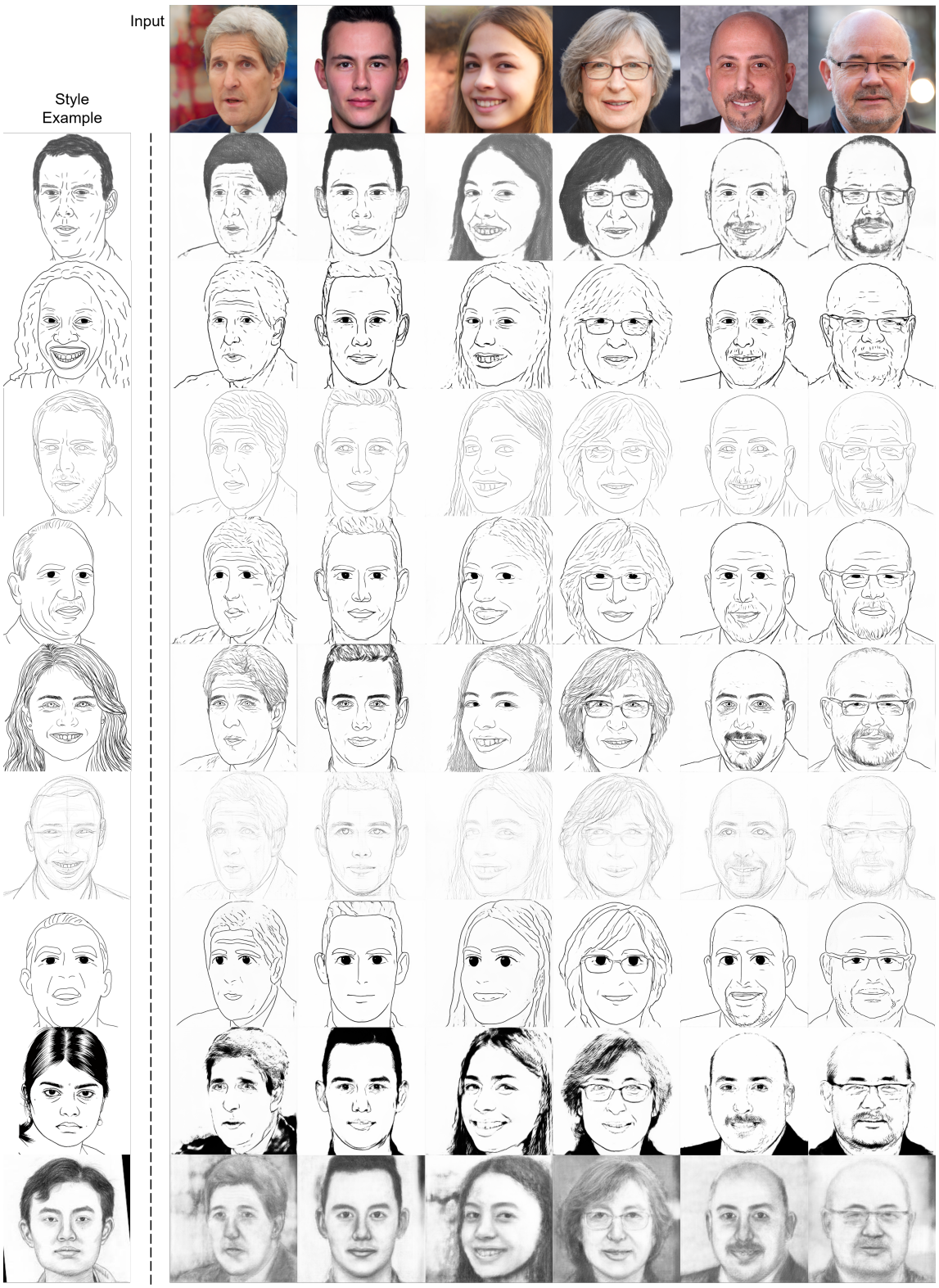


Figure 8: *Sketches extracted in nine different styles from the input images. Each style is trained with 16 pairs of data.*

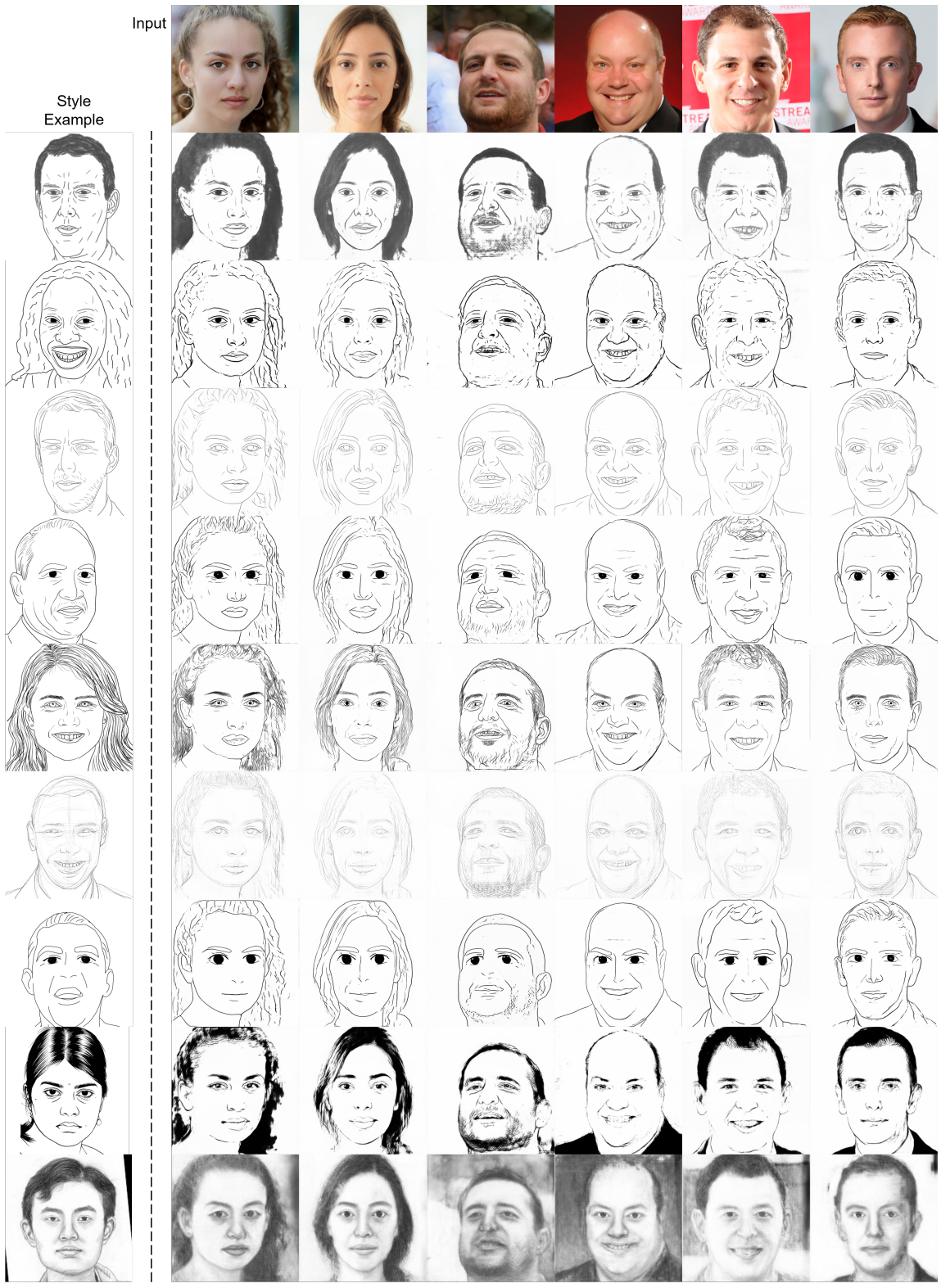


Figure 9: *Sketches extracted in nine different styles from the input images. Each style is trained with 16 pairs of data.*

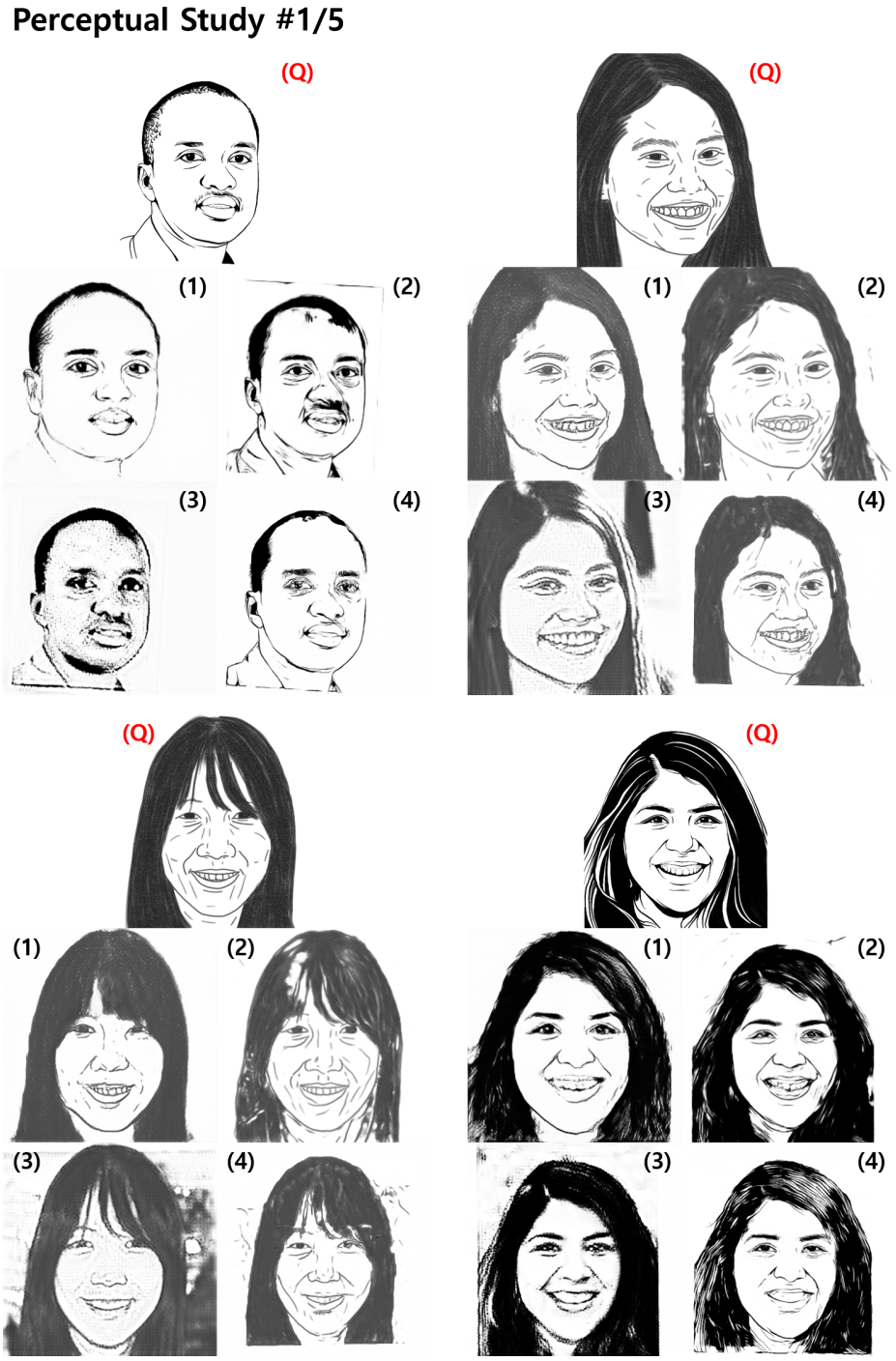


Figure 10: *Examples of the perceptual study. The sketch with red font (Q) is the ground truth target sketch and the others with black fonts (1)-(4) are the results from Ours and baseline methods. (1) Ours, (2) Ref2sketch [ASK*22], (3) Learn-to-draw [CDI22], (4) APdrawing++[YXL*20]. The display order of the sketches was chosen randomly for each comparison.*

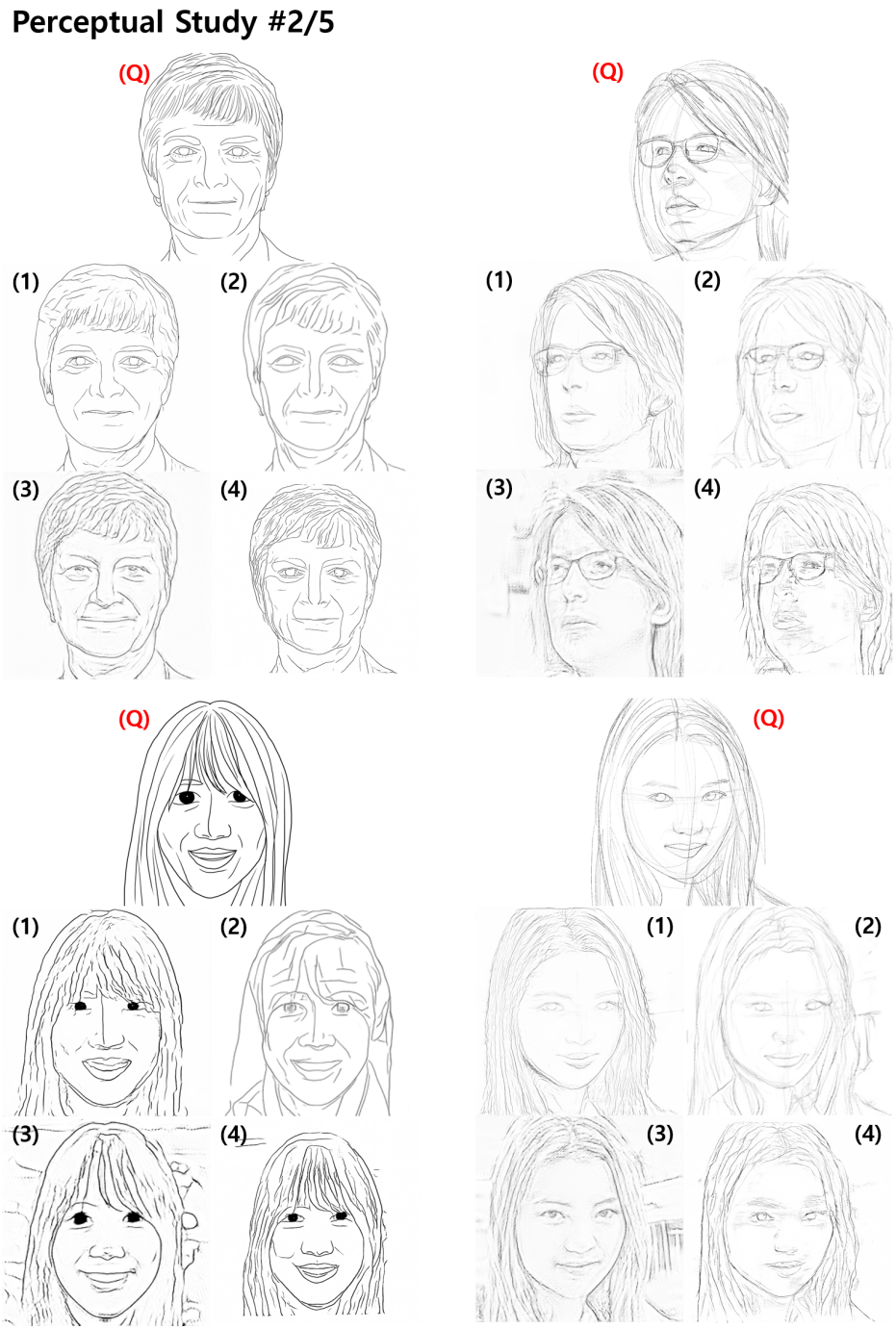


Figure 11: *Examples of the perceptual study. The sketch with red font (Q) is the ground truth target sketch and the others with black fonts (1)-(4) are the results from Ours and baseline methods. (1) Ours, (2) Ref2sketch [ASK*22], (3) Learn-to-draw [CDI22], (4) APdrawing++[YXL*20]. The display order of the sketches was chosen randomly for each comparison.*

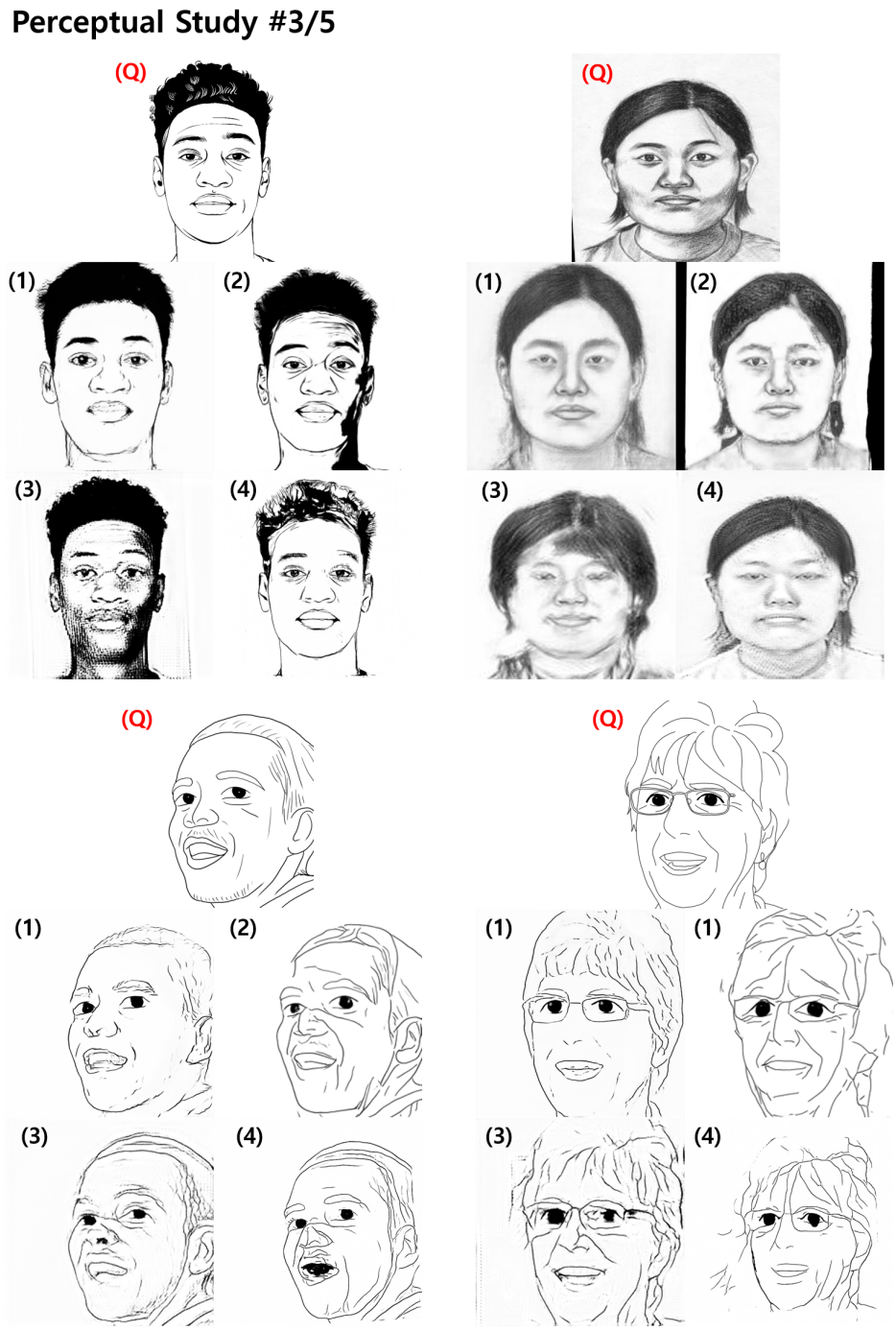


Figure 12: *Examples of the perceptual study. The sketch with red font (Q) is the ground truth target sketch and the others with black fonts (1)-(4) are the results from Ours and baseline methods. (1) Ours, (2) Ref2sketch [ASK*22], (3) Learn-to-draw [CDI22], (4) APdrawing++[YXL*20]. The display order of the sketches was chosen randomly for each comparison.*

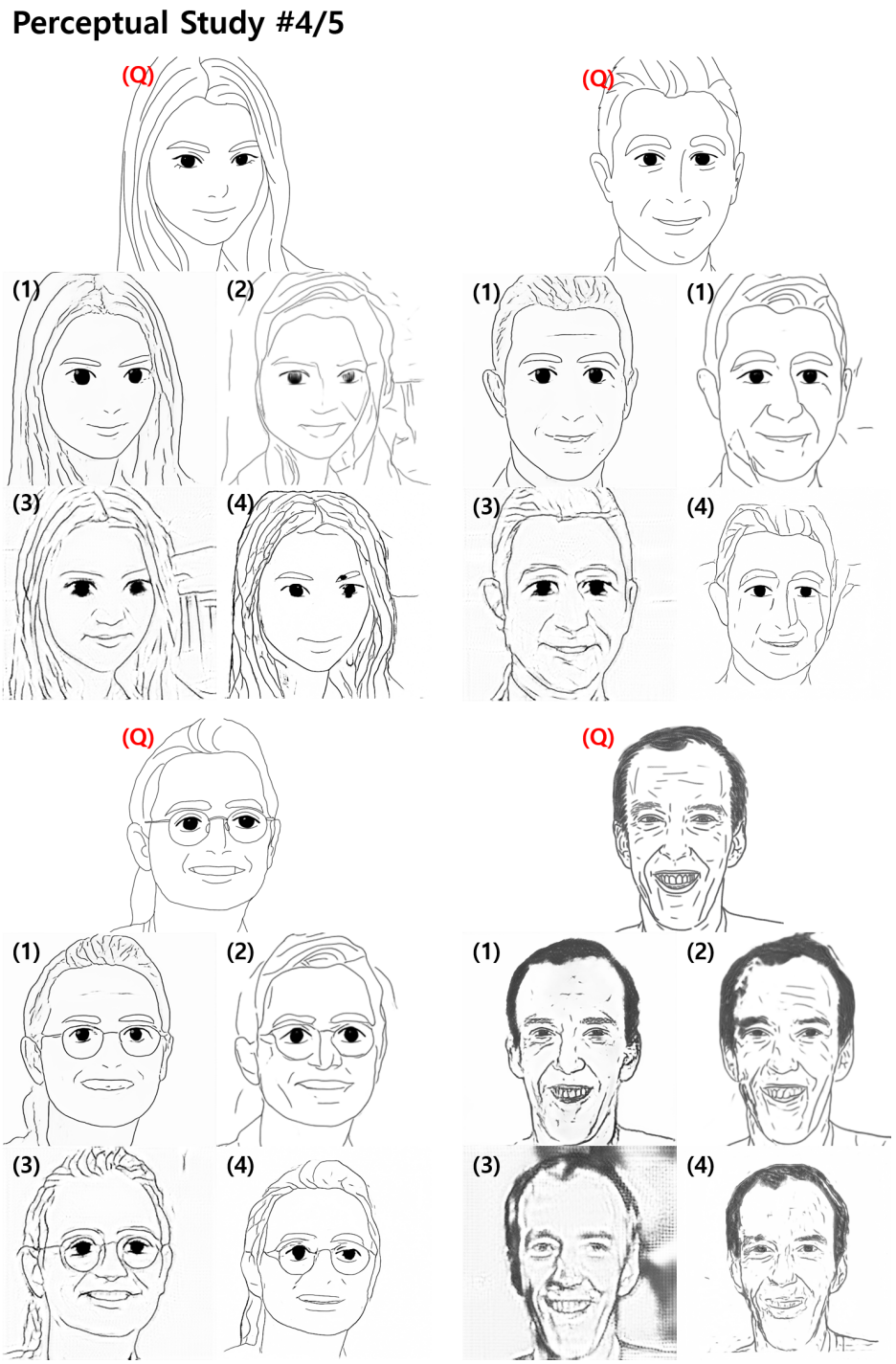


Figure 13: *Examples of the perceptual study. The sketch with red font (Q) is the ground truth target sketch and the others with black fonts (1)-(4) are the results from Ours and baseline methods. (1) Ours, (2) Ref2sketch [ASK*22], (3) Learn-to-draw [CDI22], (4) APdrawing++[YXL*20]. The display order of the sketches was chosen randomly for each comparison.*

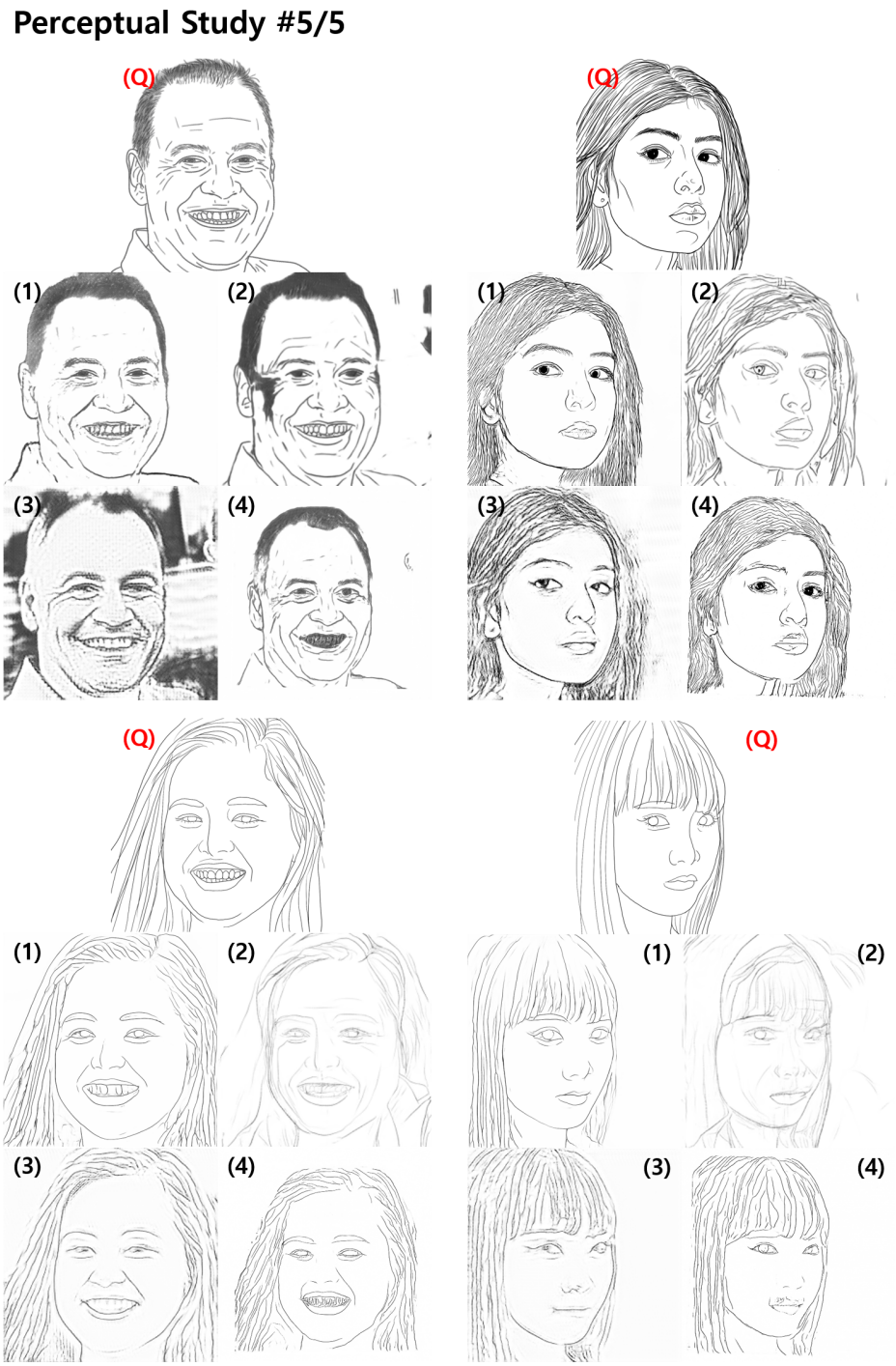


Figure 14: *Examples of the perceptual study. The sketch with red font (Q) is the ground truth target sketch and the others with black fonts (1)-(4) are the results from Ours and baseline methods. (1) Ours, (2) Ref2sketch [ASK*22], (3) Learn-to-draw [CDI22], (4) APdrawing++ [YXL*20]. The display order of the sketches was chosen randomly for each comparison.*

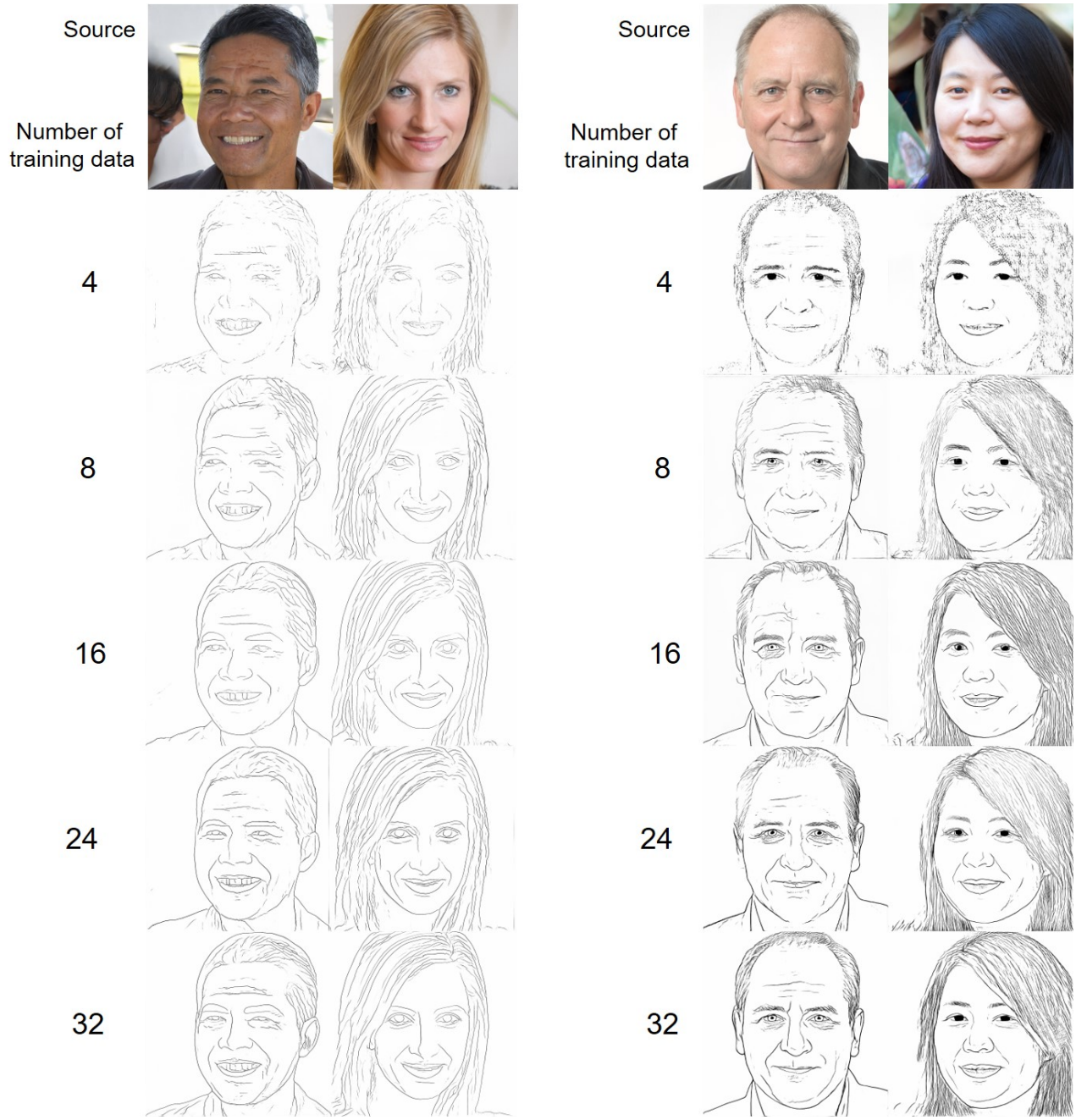


Figure 15: *The number of training data pairs used and their corresponding visual results.*

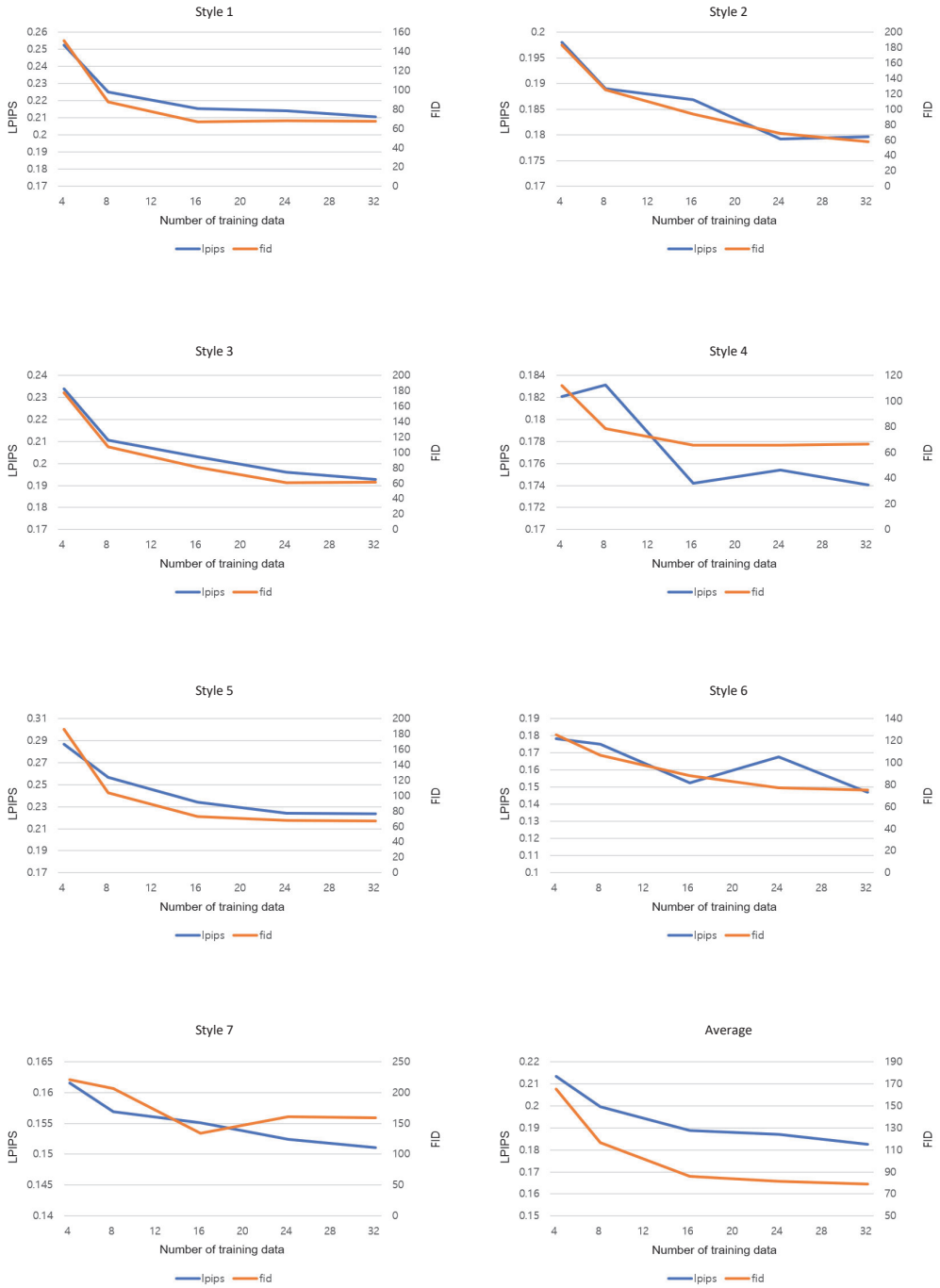


Figure 16: *The number of training data pairs used and their corresponding model performance. Styles 1-7 denote seven different styles in the SKSF-A dataset.*