Letter

A Domain-Guided Model for Facial Cartoonlization

Nan Yang, Bingjie Xia, Zhi Han, *Member, IEEE*, and Tianran Wang

Dear Editor,

This work investigates the issue of facial cartoonlization under the condition of lacking training data. We propose a domain-guided model (DGM) to realize facial cartoonlization for different kinds of faces. It includes two parts: 1) a domain-guided model that contains four different interface networks and can embed an image from a facial domain to a cartoon domain independently; and 2) a one-to-one tutoring strategy that uses a sub-model as a teacher to train other interface networks and can yield fine-grained cartoon faces. Extensive qualitative and quantitative experimental results validate our proposed method, and show that DGM can yield fine-translated results for different kinds of faces and outperforms the state of the art.

Introduction: Facial cartoonlization is a promising and interesting topic and can be used to image processing, secure computing, social media, the internet of things, and finance. CycleGAN [1] uses a cycle consistency loss to realize facial cartoonlization [2]–[7] based on generative adversarial networks [8], [9]. U-GAT-IT [10] proposes a module named class activation map (CAM) that guides a model to focus on key regions to realize facial cartoonlization. Photo2cartoon [11] is Minivision's open-source mini-program, which is built on U-GAT-IT and can be used to enhance the generative ability. These mentioned methods can only work well based on sufficient data, and can not realize a fine translation when lacking data. Besides, collecting different kinds of faces is very difficult and suffers from data imbalance, which leads to mode collapse in training phase and impedes facial cartoonlization. Therefore, two issues must be solved: 1) how to translate different kinds of faces in one training phase without affecting each other? 2) how to obtain fine-grained facial cartoonlization under insufficient or lack of data? This motivates us to develop a novel model that can: 1) realize facial cartoonlization for different kinds of faces with one model and without affecting each other; 2) yield fine-grained translated results under insufficient or lack of data.

After a rigorous statistical analysis of the dataset used by the above methods, we find that most of the existing data are for young women, not men, kids, and the elderly. In this work, we regard a young women dataset as a support set that is used to train a domain-guided model. We treat the model as a teacher for translating men, kids, and the elderly. It is reasonable since these faces are all in the facial domain despite they are different from those of men, kids, and the elderly. Therefore, we redesign a generative framework and detach four different interface networks for 1) young women, 2) men, 3) kids, and 4) the elderly, respectively. The interface networks aim to extract and translate different features in which their middle layers are shared with young women networks. Furthermore, we propose a one-to-one tutoring strategy to train these interface networks. Experimental results show that the proposed domain-guided model and strategy can well translate men, kids, and the elderly to cartoon faces without affecting each other. The creative contributions are:

1) We propose a domain-guided model based on the design idea of modularization, which consists of four different interface networks and describes the independence of networks. The model is flexible and can embed different kinds of faces into cartoons without affecting each other. It solves the issue of using only one detached model to realize facial cartoonlization for different kinds of faces.

2) We propose a one-to-one tutoring strategy, which uses the domain-guided model to train four different interface networks and describes the similarity of these networks. A well-trained women sub-model can guide the other sub-models to find an optimal initialized translation domain. It yields fine-grained results for men, kids, and the elderly as achieved for young women. It solves the issue of realizing fine-grained facial cartoonlization under the conditions of lacking data.

Experimental environment: The models involved in the experiments are trained on a workstation equipped with an Intel(R) Xeon (R) CPU @2.20GHz and two NVIDIA Tesla V100 dual-channel GPUs. All experiments are performed under the Pytorch 1.7 environment, Cuda 10.0.44, and cuDNN 10.0.20.

Proposed framework: The whole framework is shown in Fig. 1. It contains five modules, i.e., four different interface networks, hourglass blocks, a CAM module [12], a discriminator, and a classifier. The interface networks receive different input faces and output their corresponding translated cartoon faces. The sky-blue interface network receives young women's faces, which is regarded as a teacher to guide the other interface networks. The purple, yellow, and gray interface networks correspond to men (Student 1), kids (Student 2), and the elderly (Student 3), respectively. The hourglass blocks are fully considered to improve the extracted and translated performance in a progressive way. Notice that the sky-blue background denotes that the parameters of the teacher's hourglass blocks are shared with other students. The CAM module can be used to distinguish different facial domains by paying more attention to discriminative image regions. The discriminator is applied to distinguish whether a cartoon face is real or translated.

Training strategy and loss function: We denote the interface networks for the encoder and decoder as $E^i_{f\to c}$ and $G^i_{f\to c}$, where $i = 0$, 1, 2, and 3 corresponding to teacher, Students 1, 2, and 3. The shared hourglass blocks for the encoder and decoder are denoted as $\tilde{E}_{f\to c}$ and $\tilde{G}_{f\to c}$. For a face x, the translator T aims to map x to a cartoon face from facial domain f to cartoon domain c. It can be written as

$$T^i_{f\to c} = G^i_{f\to c}\left(\tilde{G}_{f\to c}\left(\tilde{E}_{f\to c}\left(E^i_{f\to c}(x)\right)\right)\right),\ i = 0,1,2,3. \quad (1)$$

The training process comprises the following steps: 1) train a fine translator $T^0_{f\to c}$ as a domain-guided model, which can map young women face to cartoon face well; 2) initialize shared parameters $\tilde{E}_{f\to c}$ and $\tilde{G}_{f\to c}$ for each student. $\tilde{E}_{f\to c}$ and $\tilde{G}_{f\to c}$ are sufficient to capture the common styles for different facial domains. It is owing to that $T^0_{f\to c}$ is trained on a large number of samples; and 3) update the parameters for different interface networks, i.e., $E^i_{f\to c}$ and $G^i_{f\to c}$, where $i = 1$, 2, and 3.

Notice that the parameters of each interface network are updated independently and under the supervision of $T^0_{f\to c}$. Thus, it belongs to a one-to-one tutoring strategy.

Denote the facial domain distribution as X_f, cartoon domain distribution as X_c, the discriminator for the cartoon domain as D_c. The training process for $f \to c$ is formulated as follows:

1) Adversarial loss: It is introduced to match the distribution of translated images to X_c:

Nan Yang and Bingjie Xia contributed equally to this work. Corresponding author: Nan Yang.

Citation: N. Yang, B. J. Xia, Z. Han, and T. R. Wang, "A domain-guided model for facial cartoonlization," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 10, pp. 1886–1888, Oct. 2022.

The authors are with Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, and also with University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yangnansia@gmail.com; xia bingjie@sia.cn; hanzhi@sia.cn; wtr@sia.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2022.105887

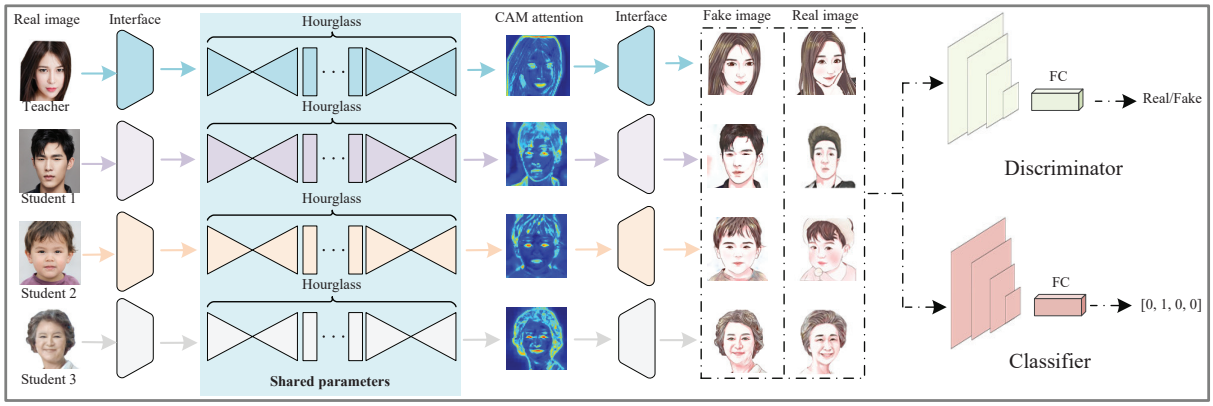

Fig. 1. The framework of DGM, which receives four different kinds of faces and can translate these faces to cartoons well. Details are described in proposed framework.

$$L_A^{f\to c}=\left(\mathbf{E}_{x\sim X_c}\left[(D_c(x))^2\right]+\mathbf{E}_{x\sim X_f}\left[\left(1-D_c\left(T^i_{f\to c}(x)\right)\right)^2\right]\right). \quad (2)$$

2) Cycle loss: It is employed to alleviate the mode collapse problem [1]

$$L_C^{f\to c}=\mathbf{E}_{x\sim X_f}\left[\left\|x-T^i_{c\to f}\left(T^i_{f\to c}(x)\right)\right\|_1\right]. \quad (3)$$

3) Identity loss: The cosine distance [13] based on a pre-trained face recognition model F is fully considered, which ensures the identity distributions of input and output images are similar

$$\begin{aligned}L_I^{f\to c}=&\mathbf{E}_{x\sim X_f}\left[1-\cos\left(F(x),F\left(T^i_{f\to c}(x)\right)\right)\right]\\&+\mathbf{E}_{x\sim X_c}\left[1-\cos\left(F(x),F\left(T^i_{c\to f}(x)\right)\right)\right].\end{aligned} \quad (4)$$

4) CAM loss: It is used to capture what makes the most difference between two domains [12] in the current state by using the auxiliary classifiers η_f and η_{D_c}.

$$\begin{aligned}L_{\mathrm{M}}^{f\to c}&=-\left(\mathbf{E}_{x\sim X_f}\left[\log\left(\eta_f(x)\right)\right]+\mathbf{E}_{x\sim X_c}\left[\log\left(1-\eta_{D_c}(x)\right)\right]\right)\\L_{\mathrm{M}}^{D_c}&=\mathbf{E}_{x\sim X_c}\left[\eta_{D_c}(x)^2\right]+\mathbf{E}_{x\sim X_f}\left[\left(1-\eta_{D_c}\left(T^i_{f\to c}(x)\right)\right)^2\right].\end{aligned} \quad (5)$$

5) Domain-guided loss: We detach another classifier $\tilde{D}_c$ on the top of D_c, which can be used to classify the different interface networks. The loss contains two items: a) a loss of real facial images used to optimize D_c; b) a loss of fake images used to optimize $E^i_{f\to c}$ and $G^i_{f\to c}$

$$L_T=\mathbf{E}_{x\sim X_c}\left[-\log\tilde{D}_c(i\mid x)\right]+\mathbf{E}_{x\sim X_f}\left[-\log\tilde{D}_c\left(i\mid T^i_{f\to c}(x)\right)\right] \quad (6)$$

where $\tilde{D}_c(i\mid x)$ represents a probability distribution for different interface networks computed by D_c.

Comparison: We compare DGM with CycleGAN, U-GAT-IT, Photo2cartoon, and group-based method (GP) [13], these methods are the most recent and the best ones. The last one proposes a group-based method for few-shot cartoon face generation. The results are shown in Fig. 2. We have three observations: 1) DGM can well translate real faces to cartoons for men, kids, and the elderly, despite lack of training data; 2) DGM can well preserve the facial identity, outline, and local details, i.e., lipstick and hair color; 3) The translated results for different kinds of faces without affecting each other. The favorable translation is owing to our detached domain-guided model and one-to-one tutoring strategy.

The detached domain-guided model contains four different interface networks, which ensures that other sub-models are trained with supervision from the sub-model of women in one training process. The one-to-one tutoring strategy is an online one, which can adjust the model's parameters in real-time.

Table 1 shows the experimental results on spatial and computational complexity: 1) Trainable model parameters (TMP); 2) Model size (MS); 3) Floating-point of operations (FLOPs); 4) Amount of multiply-adds (MAdd); 5) Memory usage (MemR); and 6) Inference time (IT).

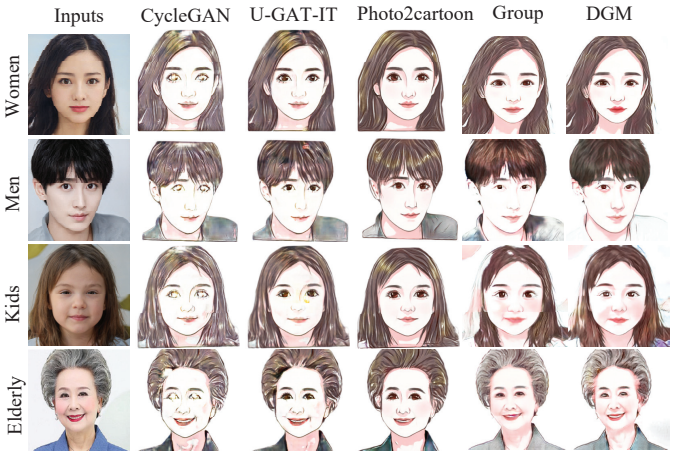

Fig. 2. The qualitative comparison results with CycleGAN [1], U-GAT-IT [12], Photo2cartoon [11], and Group [13].

Table 1. The Results on Model Complexity. M/u Denotes a Metric M and its Corresponding Unit

TMP (million)	MS/MB	Flops (G)	MAdd (G)	MemR (GB)	IT (s)
3.58	616.06	15.65	31.21	645.27	0.053

Ablation studies: We expect to evaluate DGM in two aspects: 1) its effectiveness; and 2) whether the one-to-one tutoring strategy improves the performance. Therefore, we set two configurations: DGM with and without the one-to-one tutoring strategy. The results show that DGM with the strategy eliminates image artifacts and further improves the translated quality for facial cartoonlization. Besides, we fully evaluate the translated faces for domains $f\to c$ and $c\to f$ based on many full-reference metrics [14]. The quantitative results are shown in Tables 2 and 3. DGM outperforms all the compared methods on SSIM, PSNR, MSSSIM, VSI, VIF, FSIM, GMSD, LPIPS, and DISTS [8], [14]–[19].

Discussion and conclusion: In this letter, we propose: 1) a detached domain-guided model that translates different kinds of faces in one model and without affecting each other; 2) a one-to-one tutoring strategy that realizes fine-grained facial cartoonlization under the conditions of lacking data. However, our method may have some failure cases, e.g., it has obvious distortions, especially in the eyes. We also do not fully consider edge detection. Besides, the evaluation metrics are subjective since cartoon faces have no specific ground

Table 2. Quantitative Comparison Results for Domain $f \to c$. Methods With the Best and Runner-Up Performances are Colored With Red and Blue, Respectively. The Higher the Metric Value, the Better the Performance

$f \to c$:	Women					Men					Kids					Elderly				
Configure	Cycle	GAT	P2C	GP	DGM	Cycle	GAT	P2C	GP	DGM	Cycle	GAT	P2C	GP	DGM	Cycle	GAT	P2C	GP	DGM
SSIM	0.23	0.25	0.24	0.93	0.95	0.36	0.30	0.25	0.91	0.93	0.3	0.29	0.30	0.90	0.92	0.28	0.34	0.25	0.76	0.91
PSNR	8.74	8.29	8.33	30.0	31.9	8.26	7.85	7.47	26.6	29.2	8.91	8.52	8.85	26.7	28.5	7.78	8.47	7.56	14.7	28.2
MS-SSIM	0.05	0.10	0.09	0.98	0.99	0.09	0.06	0.05	0.97	0.98	0.14	0.13	0.17	0.95	0.96	0.09	0.1	0.03	0.75	0.97
VSI	0.82	0.84	0.82	0.99	0.99	0.84	0.84	0.84	0.98	0.99	0.86	0.86	0.86	0.97	0.98	0.87	0.87	0.85	0.91	0.98
VIF	0.02	0.22	0.01	0.67	0.71	0.01	0.01	0.01	0.51	0.58	0.02	0.01	0.02	0.53	0.55	0.02	0.01	0.01	0.25	0.51
FSIM	0.63	0.63	0.62	0.96	0.97	0.66	0.66	0.64	0.94	0.96	0.65	0.64	0.67	0.93	0.94	0.67	0.68	0.64	0.80	0.94
GMSD	0.74	0.74	0.75	0.96	0.97	0.74	0.75	0.75	0.93	0.96	0.75	0.75	0.76	0.94	0.95	0.76	0.77	0.76	0.81	0.95
LPIPS	0.33	0.31	0.26	0.95	0.96	0.35	0.33	0.27	0.92	0.94	0.33	0.30	0.32	0.90	0.91	0.32	0.34	0.28	0.79	0.90
DISTS	0.5	0.49	0.49	0.90	0.91	0.47	0.50	0.48	0.88	0.89	0.50	0.46	0.51	0.84	0.86	0.68	0.47	0.48	0.77	0.86

Table 3. Quantitative Comparison Results for Domain $c \to f$

$c \to f$:	Women					Men					Kids					Elderly				
Configure	Cycle	GAT	P2C	GP	DGM	Cycle	GAT	P2C	GP	DGM	Cycle	GAT	P2C	GP	DGM	Cycle	GAT	P2C	GP	DGM
SSIM	0.66	0.63	0.63	0.95	0.96	0.77	0.71	0.71	0.95	0.96	0.75	0.71	0.65	0.93	0.96	0.77	0.73	0.73	0.93	0.94
PSNR	15.0	13.5	12.1	29.8	31.0	16.9	14.5	12.1	28.9	30.7	16.5	15.6	12.0	28.1	30.9	15.7	14.9	13.6	28.8	30.0
MSSSIM	0.75	0.74	0.67	0.98	0.99	0.83	0.77	0.72	0.98	0.98	0.80	0.77	0.66	0.96	0.98	0.79	0.76	0.73	0.97	0.97
VSI	0.89	0.89	0.88	0.98	0.99	0.92	0.91	0.89	0.98	0.98	0.93	0.92	0.89	0.97	0.98	0.91	0.90	0.89	0.97	0.97
VIF	0.19	0.16	0.14	0.59	0.62	0.31	0.19	0.17	0.56	0.60	0.32	0.27	0.16	0.51	0.59	0.29	0.24	0.21	0.50	0.53
FSIM	0.74	0.74	0.73	0.95	0.96	0.81	0.79	0.81	0.95	0.96	0.80	0.79	0.76	0.92	0.95	0.79	0.77	0.80	0.93	0.94
GMSD	0.76	0.77	0.76	0.96	0.97	0.78	0.77	0.79	0.94	0.96	0.80	0.80	0.78	0.93	0.95	0.79	0.78	0.82	0.94	0.95
LPIPS	0.77	0.76	0.74	0.97	0.98	0.85	0.81	0.80	0.96	0.97	0.83	0.81	0.74	0.95	0.97	0.82	0.80	0.78	0.95	0.96
DISTS	0.72	0.72	0.72	0.90	0.91	0.74	0.73	0.73	0.88	0.90	0.73	0.73	0.72	0.88	0.91	0.73	0.73	0.76	0.88	0.89

truth. We plan to study a novel attention module by considering edge detection to improve translated performance.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (61903358).

References

[1] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.

[2] S. Chen, Y. Tian, F. Wen, Y.-Q. Xu, and X. Tang, "Easytoon: An easy and quick tool to personalize a cartoon storyboard using family photo album," in *Proc. 16th ACM Int. Conf. Multimed.*, 2008, pp. 499–508.

[3] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "CartoonGAN: Generative adversarial networks for photo cartoonization," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2018, pp. 9465–9474.

[4] H. Li, G. Liu, and K. N. Ngan, "Guided face cartoon synthesis," *IEEE Trans. Multimed.*, vol. 13, no. 6, pp. 1230–1239, 2011.

[5] Y. Zhang, W. Dong, O. Deussen, F. Huang, K. Li, and B.-G. Hu, "Data-driven face cartoon stylization," in *Proc. SIGGRAPH Asia Tech. Briefs*, 2014, pp. 1–4.

[6] M. Yang, S. Lin, P. Luo, L. Lin, and H. Chao, "Semantics-driven portrait cartoon stylization," in *Proc. IEEE Int. Conf. Image Process*, 2010, pp. 1805–1808.

[7] J. Gong, Y. Hold-Geoffroy, and J. Lu, "Autotoon: Automatic geometric warping for face cartoon generation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 360–369.

[8] K. Liu, Z. Ye, H. Guo, D. Cao, L. Chen, and F. Y. Wang, "FISS Gan: A generative adversarial network for foggy image semantic segmentation," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 8, pp. 1428–1439, 2021.

[9] J. Li, Y. Tao, and T. Cai, "Predicting lung cancers using epidemiological data: A generative-discriminative framework," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 5, pp. 1067–1078, 2021.

[10] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," arXiv preprint arXiv: 1907. 10830, 2019.

[11] Minivision-AI, "Photo2cartoon," [Online]. Available: https://github.com/minivision-ai/photo2cartoon. Accessed: Jun. 2022.

[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[13] N. Zhuang and C. Yang, "Few-shot knowledge transfer for fine-grained cartoon face generation," in *Proc. IEEE Int. Conf. Multimed. Expo*, 2021, pp. 1–6.

[14] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022.

[15] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.

[16] B. Zhang, P. V. Sander, and A. Bermak, "Gradient magnitude similarity deviation on multiple scales for color image quality assessment," in *Proc. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process*, 2017, pp. 1253–1257.

[17] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2013, vol. 2, pp. 1398–1402.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.