

Supplementary Materials: Latent Representation Reorganization for Face Privacy Protection

Anonymous Authors

ABSTRACT

In this supplementary material, we present more details about our proposed LReOrg framework, including the network structure, methodology comparisons and more experimental results.

1 THE DETAILS OF CDFNET

In this part, we first introduce the details of our CDFNet which consists of a forward process and a backward process. The two processes share a same set of network structure and parameters. For clarity, we introduce their network structure separately, where the forward process is realized by stacking the conditional disentanglement block (CDB) and the backward process is realized by stacking the conditional fusion block (CFB).

1.1 The CDB and CFB Blocks

In Table 1, we present the details of the CDB block. In Table 2, we present the details of the CFB block. The differences between CDB and CFB lie in the fundamental operations of arithmetic in their building blocks, such as the elementwise multiplication, division, addition and subtraction: \otimes , \oslash , \oplus , \ominus . Notice that both of them are built based on the Linear Block and the Dense Block, which are introduced in the next few subsections.

1.2 The Linear Block

The linear block is used to handle the conditions by taking a flow feature h and a condition feature c as input and output a conditional feature. As shown in Table 3, the linear block employs three fully connected (FC) layers to adjust and fuse h and c , where the Tanh is used as the activation function.

1.3 The Dense Block

The dense block is used to handle the feature disentanglement and fusion. Our inspiration comes from HiNet [6] which focuses on image steganography, where the dense block is built to process 2D image using 2D convolution (Conv2d). Differently, the dense block used in this paper is extended to process 1D feature by using 1D convolution (Conv1d) to support feature disentanglement and fusion. As shown in Table 4, the dense block employs five Conv1d layers for feature transformation. where the Tanh is used as the activation function and the output of all previous layers are concatenated by channel. During the forward process, dense block is used to disentangle the key features from the input. During the backward process, it is used for feature fusion.

2 METHODOLOGY COMPARISONS

In this part, we compare our method with the most related recent anonymization methods. Even though they involve CLIP [12] or identity disentanglement for anonymization, there are still distinguishing discrepancies between them and our method.

Table 1: The details of the CDB block. Top: the forward condition block. Bottom: the disentanglement block.

Input	Layers	Output	
h_{j-1}^2, k_i^f	Linear Block	(4928)	o_1
h_{j-1}^1, o_1	\oplus	(4928)	\hat{h}_j^1
o_1, k_i^f	Linear Block	(4928)	o_2
o_2	exp	(4928)	o_3
h_{j-1}^2, o_3	\otimes	(4928)	o_4
\hat{h}_j^1, k_i^f	Linear Block	(4928)	o_5
o_4, o_5	\oplus	(4928)	\hat{h}_j^2
\hat{h}_j^2	Dense Block	(4928)	o_6
\hat{h}_j^1, o_6	\oplus	(4928)	h_j^1
\hat{h}_j^1	Dense Block	(4928)	o_7
o_7	exp	(4928)	o_8
\hat{h}_j^2, o_7	\otimes	(4928)	o_9
h_j^1	Dense Block	(4928)	o_{10}
o_9, o_{10}	\oplus	(4928)	h_j^2

Table 2: The details of the CFB block. Top: the fusion block. Bottom: the backward condition block.

Input	Layers	Output	
g_{j-1}^1	Dense Block	(4928)	o_1
g_{j-1}^2, o_1	\ominus	(4928)	o_2
g_{j-1}^1	Dense Block	(4928)	o_3
o_3	exp	(4928)	o_4
o_2, o_4	\oslash	(4928)	\hat{g}_j^2
\hat{g}_j^2	Dense Block	(4928)	o_5
g_{j-1}^1, o_5	\ominus	(4928)	\hat{g}_j^1
\hat{g}_j^1, k_i^f	Linear Block	(4928)	o_6
\hat{g}_j^2, o_6	\ominus	(4928)	o_7
\hat{g}_j^1, k_i^f	Linear Block	(4928)	o_8
o_8	exp	(4928)	o_9
o_7, o_9	\oslash	(4928)	g_j^2
g_j^2, k_i^f	Linear Block	(4928)	o_{10}
\hat{g}_j^1, o_{10}	\ominus	(4928)	g_j^1

Table 3: The details of the linear block.

Input	Layers	Output	
z	FC1&Tanh	(1024)	o_z
c	FC2&Tanh	(1024)	o_c
o_z, o_c	FC3&Tanh	(4928)	o_{cz}

Table 4: The details of the dense block.

Input	Layers	Output	
z	Conv1d&Tanh	(32, 4928)	o_z^1
Concat(z, o_z^1)	Conv1d&Tanh	(32, 4928)	o_z^2
Concat(z, o_z^1, o_z^2)	Conv1d&Tanh	(32, 4928)	o_z^3
Concat(z, o_z^1, o_z^2, o_z^3)	Conv1d&Tanh	(32, 4928)	o_z^4
Concat($z, o_z^1, o_z^2, o_z^3, o_z^4$)	Conv1d	(1, 4928)	o_z^5

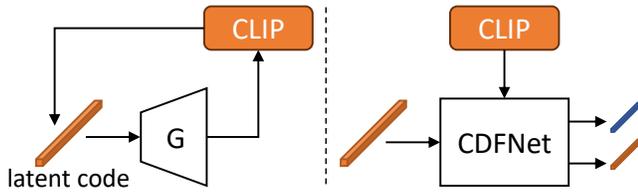


Figure 1: Differences between existing CLIP-based anonymization method, e.g., CLIP2Protect (left), and ours CDFNet (right). G denotes the generator of CLIP2Protect.

CLIP for Anonymization. The pretrained CLIP model is used in both CLIP2Protect [13] and our LReOrg framework, but their focuses are different. As illustrated in Figure 1, the CLIP model is used to manipulate the latent code in the StyleGAN [8] space by using a textual loss to hide the adversarial perturbations into the makeup effect. However, in our framework, CLIP is mainly used to generate the cross-modal conditions of our CDFNet to supervise the disentanglement or fusion of the desired identity or attribute information in the latent space. Besides, CLIP2Protect is time consuming for inference by finetuning the StyleGAN generator for each input image, which also has limited protection ability because the generated anonymous faces can be easily re-identified by face recognition models (i.e. low PSR rates). Our experimental results in the submitted manuscript have already verified this.

Disentanglement for Anonymization. Previously, the disentanglement (e.g. IdentityDP [16]) is employed to obtain two kinds of information by using identity encoder (e.g. pretrained ArcFace [1]) and attribute encoder. The identity encoder is used to extract identity feature and anonymize it with a modified one (e.g. by adding Laplace noise [2, 14, 16]). The attribute encoder is used to extract feature of the other information or facial attributes to be preserved. Instead of only obtaining two kinds of information, as shown in Figure 2, our proposed CDFNet is designed to disentangle fine-grained information (e.g. identity, gender, age, expression and makeup) based on the cross-modal keyword conditions so that it can support more effective and flexible anonymization by letting users to determine which information (identity or attributes) to be anonymized or preserved according to the practical requirements. Compared to one closely related work IdentityDP, our results in the submitted manuscript show better privacy protection ability and privacy-utility tradeoff.

3 MORE EXPERIMENTAL RESULTS

In this part, we provide more qualitative and quantitative results in terms of feature representation, controllability and comparison with state-of-the-art (SOTA) methods on privacy-utility tradeoff.

3.1 Feature Representation

To verify the disentanglement ability of our CDFNet in the latent space, we plot the T-SNE [15] embedding results of the disentangled identity features in Figure 3. According to the plot, it is obvious that the feature points exhibit clustering properties and the faces in each cluster share more similarities than the faces from the other



Figure 2: Differences between existing disentanglement method for anonymization (left) and our solution (right). E denotes the disentanglement network.

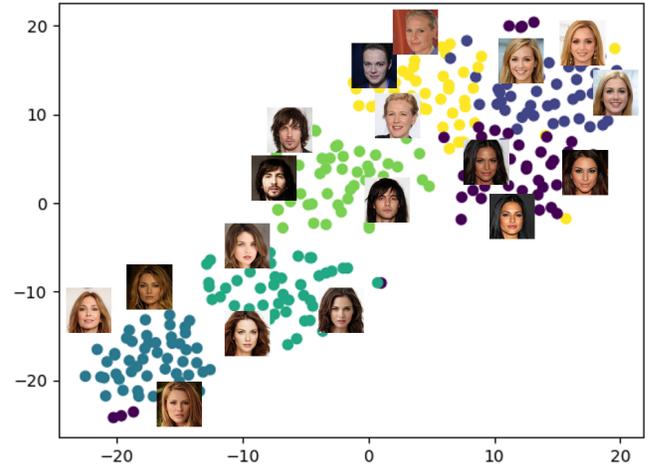


Figure 3: The T-SNE embedding results of the disentangled identity features extracted by using our CDFNet.

clusters. This suggests that the disentangled features correspond to identity representation of the faces.

3.2 Controllability

In addition to directly control the anonymization and recovery of the facial attributes, we show more controllability results of the proposed approach.

In our formulation, the identity is regarded as a special case of 'attribute'. With the help of our anonymizer, we can obtain an anonymized feature \hat{f}_a for the CDFNet feature f_a of each attribute. To visually show the anonymization process, we propose to manipulate the anonymized feature \hat{f}_a in a linear manner by using

$$f_m = \xi \cdot (\hat{f}_a - f_a) + f_a \quad (1)$$

as the intermediate value, where $\xi \in [0, 1]$ is a coefficient.

In Figure 4, we demonstrate some representative linear manipulation results by gradually varying f_m from f_a to \hat{f}_a . It is obvious that we can achieve some direct control of the results by using different ξ with salient changes on the corresponding attributes. Thus, our LReOrg can enable an interactive anonymization by regarding ξ as a notable parameter to adjust the privacy protection level as well as the ability of attribute preservation.

In a separate folder ('Manipulation') of this supplementary material, we also show some gif figures to illustrate the linear manipulation process of our approach: Original->Anonymization->Recovery.

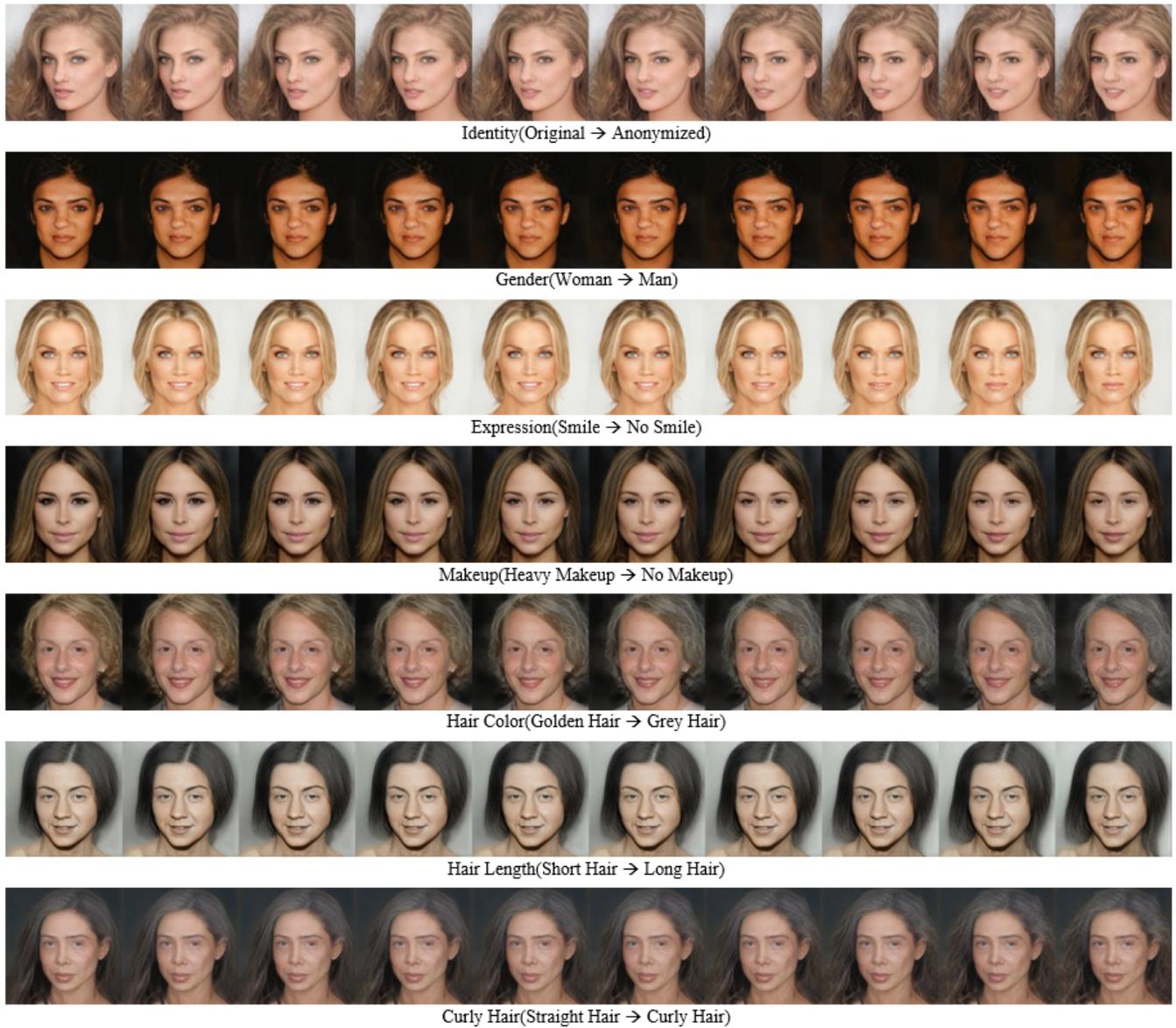


Figure 4: The sequential face images generated by performing linear manipulation towards different kinds of attributes.

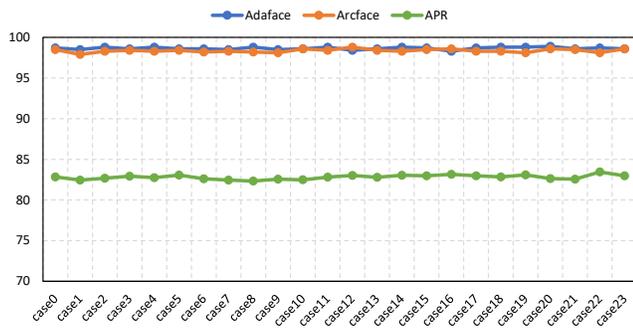


Figure 5: The PSR and APR curves of the different recovery cases of the non-sensitive attributes.

3.3 Impact of Attribute Recovery Orders

In this part, we explore the influences of the recovery orders of non-sensitive attributes with respect to the performance of privacy protection in the Recovery Reorganization module of LReOrg. In Table 5, we present 24 recovery cases of the four example attributes (i.e. Gender, Age, Expression, Makeup) by exchanging their permutations. In Figure 5, we report the PSR and APR curves of the generated face images under different cases. It is easy to observe that each group of the PSR and APR data almost follow a straight line with limited variances, which indicates that the recovery order have non-salient influences on privacy protection and utility preservation. Therefore, users do not need to care too much about the order of non-sensitive attribute recovery for using LReOrg.

Table 5: Ablation study settings of the attribute recovery orders of our recovery reorganization module.

Case	Recovery Order
case0	('Gender', 'Age', 'Expression', 'Makeup')
case1	('Gender', 'Age', 'Makeup', 'Expression')
case2	('Gender', 'Expression', 'Age', 'Makeup')
case3	('Gender', 'Expression', 'Makeup', 'Age')
case4	('Gender', 'Makeup', 'Age', 'Expression')
case5	('Gender', 'Makeup', 'Expression', 'Age')
case6	('Age', 'Gender', 'Expression', 'Makeup')
case7	('Age', 'Gender', 'Makeup', 'Expression')
case8	('Age', 'Expression', 'Gender', 'Makeup')
case9	('Age', 'Expression', 'Makeup', 'Gender')
case10	('Age', 'Makeup', 'Gender', 'Expression')
case11	('Age', 'Makeup', 'Expression', 'Gender')
case12	('Expression', 'Gender', 'Age', 'Makeup')
case13	('Expression', 'Gender', 'Makeup', 'Age')
case14	('Expression', 'Age', 'Gender', 'Makeup')
case15	('Expression', 'Age', 'Makeup', 'Gender')
case16	('Expression', 'Makeup', 'Gender', 'Age')
case17	('Expression', 'Makeup', 'Age', 'Gender')
case18	('Makeup', 'Gender', 'Age', 'Expression')
case19	('Makeup', 'Gender', 'Expression', 'Age')
case20	('Makeup', 'Age', 'Gender', 'Expression')
case21	('Makeup', 'Age', 'Expression', 'Gender')
case22	('Makeup', 'Expression', 'Gender', 'Age')
case23	('Makeup', 'Expression', 'Age', 'Gender')

3.4 Comparison of Embedding Distance

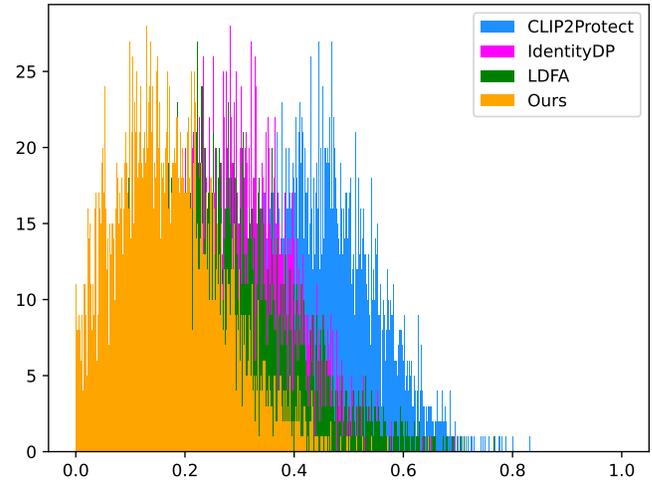
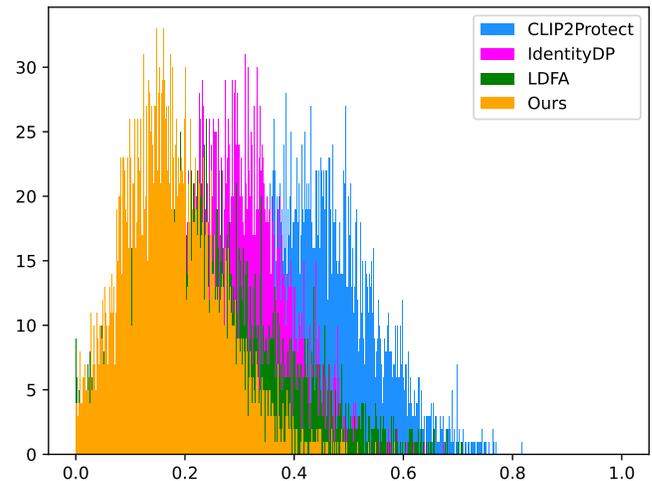
In order to better show the face anonymization ability of our method, we visualize the histograms of the embedding distances of face recognition models by following [10]. Figure 6 and Figure 7 present the metrics of Arcface [1] and Adaface [9] in histogram, respectively. In both evaluations, we use cosine similarity as the measurement of distance. This means that a distribution of smaller mean and variance indicates better anonymization performance. Obviously, our method shows the smallest mean and variance compared with the SOTA methods LDFA [10], IdentityDP [16] and CLIP2Protect [13], which indicates a more effective performance in face privacy protection.

3.5 More Visual Results

In Figure 8, we present more visual comparison results between our pre-trained LReOrg model and the representative and SOTA methods on another popular FFHQ dataset [7]: DP1 [3], DP2 [4], CIAGAN [11], IdentityDP [16], DartBlur [5], LDFA [10] and CLIP2Protect [13]. The face images generated by DP1 and CLIP2Protect may share some visual similarities with the original input faces. The results of CIAGAN, IdentityDP, DP2 and LDFA may have noticeable artifacts, while DartBlur can not generate visually pleasing images. In contrast, LReOrg(Ours) exhibits more acceptable results for anonymization.

REFERENCES

- [1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*. 4690–4699.

**Figure 6: The comparison results of the embedding distances by using the extracted Arcface features.****Figure 7: The comparison results of the embedding distances by using the extracted Adaface features.**

- [2] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *Proceedings of the International Conference on Theory and Applications of Models of Computation*. 1–19.
- [3] Häkon Hukkelås, Rudolf Mester, and Frank Lindseth. 2019. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*. Springer, 565–578.
- [4] Häkon Hukkelås and Frank Lindseth. 2023. DeepPrivacy2: Towards Realistic Full-Body Anonymization. In *WACV*. 1329–1338.
- [5] Baowei Jiang, Bing Bai, Haozhe Lin, Yu Wang, Yuchen Guo, and Lu Fang. 2023. DartBlur: Privacy Preservation With Detection Artifact Suppression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16479–16488.
- [6] Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. 2021. HiNet: deep image hiding by invertible network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4733–4742.
- [7] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.
- [8] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

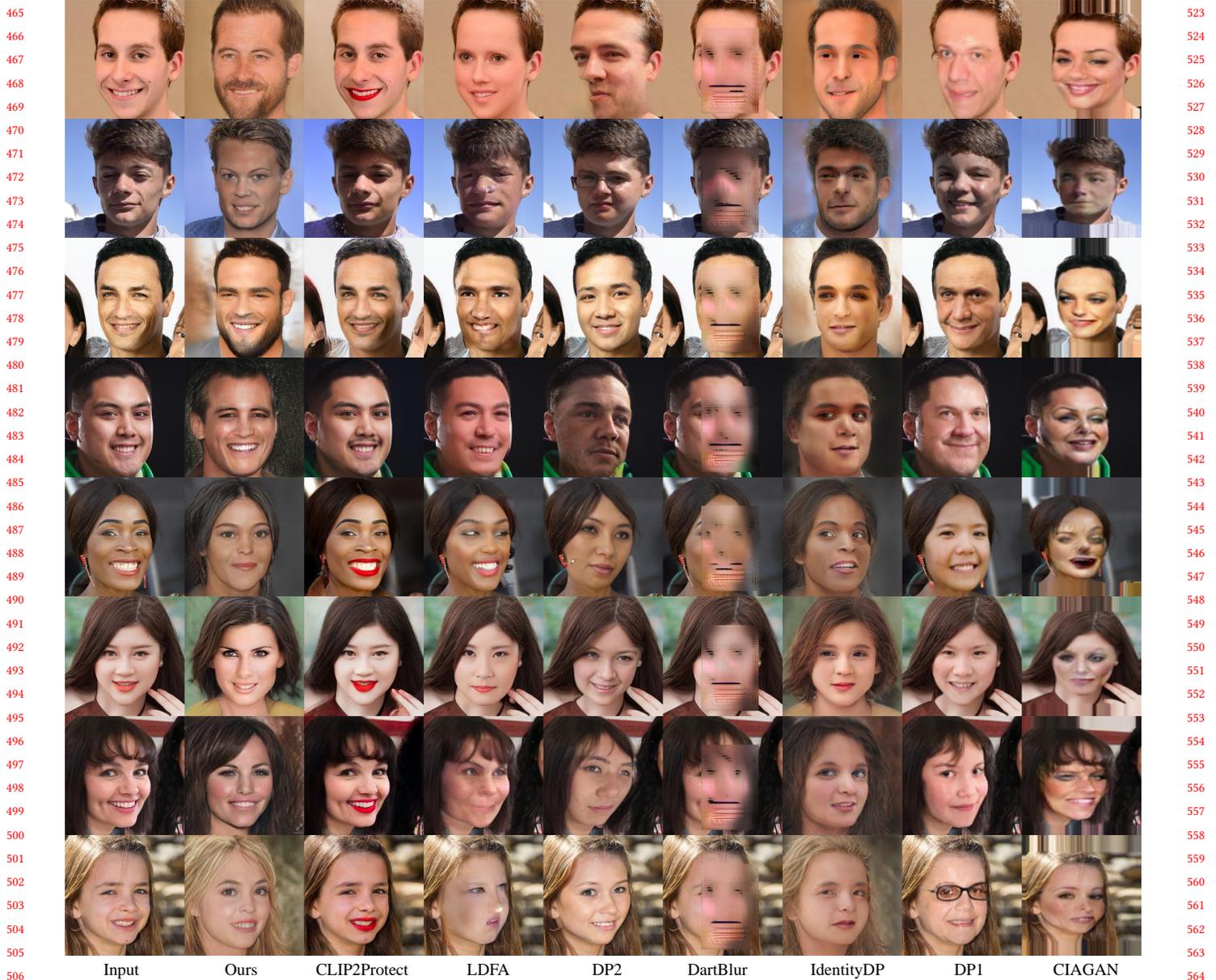


Figure 8: More visual comparison results of our method with that of SOTA. The first column presents the original input faces.

[9] Minchul Kim, Anil K Jain, and Xiaoming Liu. 2022. AdaFace: Quality Adaptive Margin for Face Recognition. In *CVPR*. 18750–18759.

[10] Marvin Klemp, Kevin Rösch, Royden Wagner, Jannik Quehl, and Martin Lauer. 2023. LDFA: Latent Diffusion Face Anonymization for Self-Driving Applications. In *CVPRW*. 3198–3204.

[11] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. 2020. CIAGAN: conditional identity anonymization generative adversarial networks. In *CVPR*. 5447–5456.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[13] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. 2023. Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20595–20605.

[14] Latanya Sweeney. 1997. Guaranteeing anonymity when sharing medical data, the datafly system. In *Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, 51.

[15] Laurens Van der Maaten and Geoffrey Hinton. 2012. Visualizing non-metric similarities in multiple maps. *Machine learning* 87, 1 (2012), 33–55.

[16] Yunqian Wen, Bo Liu, Ming Ding, Rong Xie, and Li Song. 2022. Identitydp: Differential private identification protection for face images. *Neurocomputing* 501 (2022), 197–211.