Appendices In this supplementary material, we conduct a series of studies on the proposed models as follows.

1. More ablation study

Study on face-aware retrieval system We evaluate the model performances with/without the face-aware retrieval system as shown in Table 1. As we can see, the face-aware retrieval improves the model performances on all three tasks.

Table 1. Study on face-aware retrieval system.

Model		CelebA	RAF-DB	300W
ProS-1M-real	w/o	91.42	89.34	3.35
	w/ (ours)	91.58	89.83	3.32

The different number of prototypes, architecture and training time: We compare the performances of the proposed ProS-1M-syn model on the different numbers of prototypes, architecture, and training epochs. The results are shown in Table 2. As we can observe, the performances are improved with the increasing number of prototypes from 1 [1], 512,1024 and start degrading at 2048. Therefore, we set the default number of prototypes as 1024. In addition, we evaluate the model with a longer training time (100 *vs* 20 epochs) and a larger model ViT-B/16 (85M *vs* 21M). We can observe the longer training iterations and a larger model size do slightly improve the model performances.

Table 2. Ablation study of different number of prototypes, training epochs and model architecture on ProS-1M-syn, which is trained on 1024 prototypes, 20 epochs and ViT-S/16.

		CelebA	RAF-DB	300W
# of prototypes	1	90.45	86.48	3.71
	512	91.46	88.46	3.38
	1024 (ours)	91.57	89.06	3.36
	2,048	91.53	88.85	3.38
epochs	20 (ours)	91.57	89.06	3.36
	100	91.59	89.44	3.35
architectures	ViT-S/16 (ours)	91.57	89.06	3.36
	ViT-B/16	91.52	89.53	3.35

Data size: We study how the data size of face images could influence the final performance. In particular, we study the training data size of 0.2M, 0.5M, 1M, and 8M on real images. We report the results in Table 3. As we can observe, the more training images we use, the better performance.

2. Models comparison

The differences between the proposed method and existing ones [2, 3] are shown in Table 4. Compared with DINO, we add the prototypes and use the Sinkhorn regularization [4]. Compared with SwAV, we explore the momentum encoder and vision transformer architecture.

[1] we use the loss in Dino [3]

Table 3. Study on data size.

Size	CelebA	300W	RAF-DB
0.2M	91.45	3.57	81.75
0.5M	91.53	3.48	85.91
1M	91.58	3.32	89.83
8.6M (full)	91.88	3.27	91.04

Table 4. Comparison between proposed ProS, DINO [3] and SwAV [2]

Methods	Momentum	Prototype	Operation (teacher)	Architecture	Dataset
SwAV [2]		✓	Sinkhorn [4]	ResNet	ImageNet
DINO [3]	✓		Centering	Vision Transformer	ImageNet
ProS(ours)	✓	✓	Sinkhorn [4]	Vision Transformer	MS1M

2.1. Pre-training models on face dataset

We re-implement the pre-training methods such as DINO [3], MAE [5], and MSN [1] models on the synthetic 1M images and evaluate the downstream tasks as shown in Table 5. For a fair comparison, we use the ViT-S/16 architecture for these methods and linearly scale the learning rate based on the data size. As we can observe, ProS still outperforms the other baselines, especially on the expression estimation task at RAF-DB dataset. This indicates the superiority of the proposed method compared with the other baselines when trained with the same face dataset.

Table 5. Experimental comparison with DINO [3], MAE [5], and MSN [1] methods on facial datasets

Methods	CelebA	RAF-DB	300W
DINO [3]	91.45	87.48	3.41
MAE [5]	91.28	87.73	3.38
MSN [1]	91.43	88.19	3.38
ProS-1M-syn (ours)	91.57	89.06	3.36

3. Linear probe

We analyze the feature learned from ProS-1M-syn model by fine-tuning with frozen vision-transformer backbone and the study results are shown in Table 6. As we can observe, the linear probe results from synthetic data are better on face attribute estimation. While, the model from real images achieves better performance on expression classification and face alignment.

Experiments on face parsing As shown in Table 7, ProS fails to achieve excellent results on the face parsing on LaPa dataset. One reason could be that the learned features mostly cover the facial region but not the hair region, which can also be observed in the parsing result in the ”Hair” class.

Table 6. Study on linear probe with frozen ViT-S/16 backbone.

Dataset	CelebA					LFWA				
Portion	0.2%	0.5%	1%	2%	100%	5%	10%	20%	50%	100%
# of training data	325	843	1,627	3,255	162,770	313	626	1,252	3,131	6,263
ProS-1M-syn$_{lp}$	87.42	88.64	89.17	89.67	90.23	82.55	83.26	83.98	84.76	85.14
ProS-1M-real$_{lp}$	87.30	88.24	88.80	89.31	90.62	81.02	82.13	83.02	84.08	84.72

	AffectNet8			RAF-DB			
Methods	Full	10%	2%	Full	10%	2%	1%
ProS-1m-syn$_{lp}$	42.06	38.48	33.78	80.04	73.40	64.86	56.23
ProS-1m-real$_{lp}$	43.01	40.56	37.56	75.46	69.20	60.07	55.64

	WFLW					300W		
Methods	0.7%	5%	10%	20%	100%	1.5%	10%	100%
ProS-1M-syn$_{lp}$	10.73	8.00	7.39	6.94	6.12	5.56 11.12 6.64	4.32 8.33 5.12	3.66 6.72 4.26
ProS-1M-real$_{lp}$	9.47	7.25	6.76	6.35	5.68	5.31 10.44 6.32	4.17 7.90 4.90	3.58 6.39 4.13

Table 7. Comparison with SOTA methods on LaPa [6] dataset.

Subset	Skin	Hair	L-E	R-E	U-L	I-M	L-L	Nose	L-B	R-B	Mean
FaRL [8]	97.52	95.11	92.33	92.09	88.69	90.70	90.05	97.55	91.57	91.34	92.70
AGRNet [7]	97.7	96.5	91.6	91.1	88.5	90.7	90.1	97.3	89.9	90.0	92.3
ProS-1M-syn	96.95	93.20	91.09	90.86	87.58	89.47	89.26	97.45	90.47	89.60	91.60
ProS-1M-real	97.05	93.55	91.02	91.20	88.01	89.73	89.26	97.40	90.34	89.95	91.70
ProS-full-real	97.13	93.57	91.42	91.32	88.27	90.10	89.51	97.52	90.88	90.27	92.00

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. *arXiv preprint arXiv:2204.07141*, 2022. 1

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 1

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 1

[4] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 1

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1

[6] Yinglu Liu, Hailin Shi, Hao Shen, Yue Si, Xiaobo Wang, and Tao Mei. A new dataset and boundary-attention semantic segmentation for face parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11637–11644, 2020. 2

[7] Gusi Te, Wei Hu, Yinglu Liu, Hailin Shi, and Tao Mei. Agrnet: Adaptive graph representation learning and reasoning for face parsing. *IEEE Transactions on Image Processing*, 30:8236–8250, 2021. 2

[8] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022. 2