Supplementary Material for ComFace: Facial Representation Learning with Synthetic Data for Comparing Faces

Yusuke Akamatsu[1,†], Terumi Umematsu[1], Hitoshi Imaoka[1], Shizuko Gomi[2], Hideo Tsurushima[2]
[1]NEC Corporation, Japan [2]University of Tsukuba, Japan
[†]yusuke-akamatsu@nec.com

A. Datasets

A.1. Synthetic Face Images

We utilize StyleGAN [17] and StyleGAN3 [16] to generate synthetic face images. We use StyleGAN and StyleGAN3 trained with Flickr-Faces-HQ dataset at 1024×1024. In StyleGAN, we use the attributes that vary weight provided by Ref. [25] and the attributes that vary age and smile provided by InterFaceGAN [28]. In StyleGAN3, we employ the 40 attributes included in the CelebA dataset [20] provided by Ref. [2]. For each attribute, StyleGAN generates 50,000 identities and StyleGAN3 generates 5,000 identities for synthetic face images. Among the attributes, identities are the same for StyleGAN, and identities are different for StyleGAN3. Therefore, StyleGAN generates 50,000 identities, and StyleGAN3 generates 5,000 identities × 40 attributes = 200,000 identities, resulting in a total of 250,000 identities. For each identity, we generate 101 synthetic images for each attribute with $\alpha = \{-5.0, -4.9, \cdots, 0.0, \cdots, 4.9, 5.0\}$ [1]. Since pairs of $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are randomly sampled from 101 images, the number of images used for training is an even number, up to 100 [2]. Finally, StyleGAN generates 50,000 identities × 3 attributes × 100 images = 15,000,000 (15M) images, and StyleGAN3 generates 5,000 identities × 40 attributes × 100 images = 20,000,000 (20M) images, for a total of 35M images to be used for FRL.

[1]In StyleGAN, face images with $\alpha = 0.0$ were excluded to avoid duplicating the same image since the identities are the same among the attributes.

[2]In curriculum learning, pairs are created with a restriction on the range of change $|\alpha_{\boldsymbol{y}_i} - \alpha_{\boldsymbol{x}_i}|$, which may be less than 100 images.

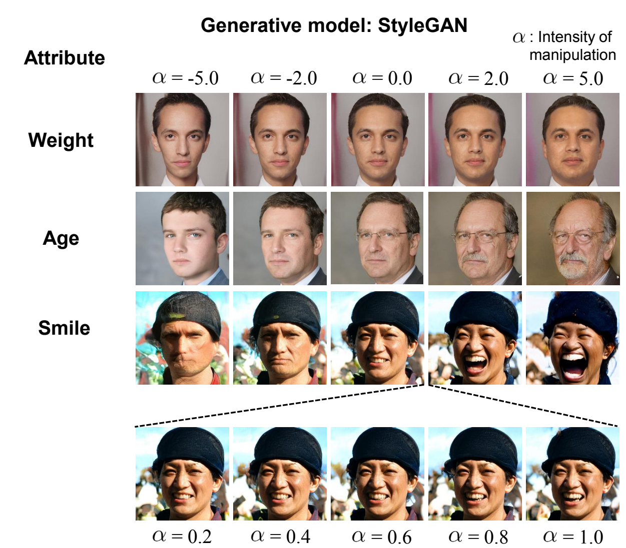

Figure A. Examples of face images generated by StyleGAN for all attributes.

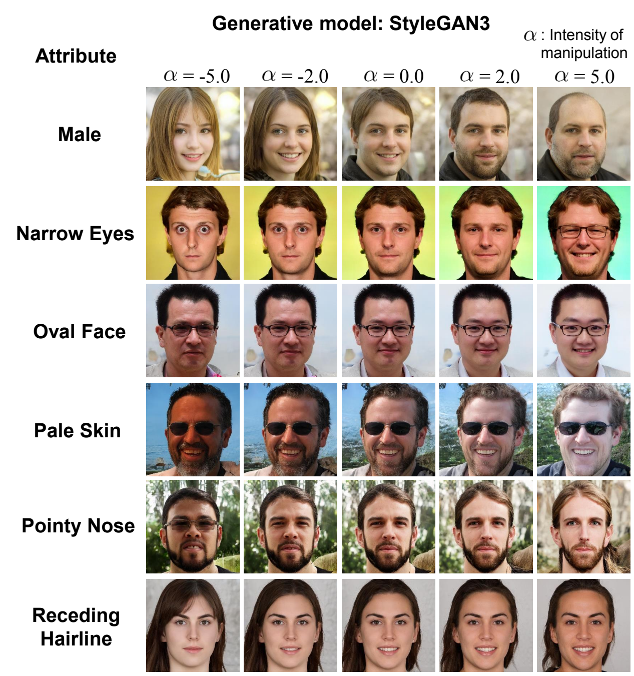

Figure B. Examples of face images generated by StyleGAN3 for part of 40 attributes.

Table A. References and boundaries used in facial manipulation for each attribute.

Attribute (Generative model)	Reference	Boundary
Weight (StyleGAN)	Ref. [25]	weight_orth_mouth.npy [3]
Age (StyleGAN)	InterFaceGAN [28]	stylegan_ffhq_age_w_boundary.npy [4]
Smile (StyleGAN)	InterFaceGAN [28]	stylegan_ffhq_smile_w_boundary.npy [4]
40 attributes in CelebA (StyleGAN3)	Ref. [2]	Aligned FFHQ/*attribute*_boundary.npy [5,6]

[3] https://github.com/LARC-CMU-SMU/facial-weight-change
[4] https://github.com/genforce/interfacegan
[5] *attribute* corresponds to the name of each attribute.
[6] https://github.com/yuval-alaluf/stylegan3-editing

Table B. ComFace settings for FRL.

Config	Value
Batch size	1024
Optimizer	Adam [18]
Learning rate	4.0e-4
Epochs	12
Learning rate schedule	halved in 10 epochs
Computing resource	32 NVIDIA A100 GPUs
Image size	224×224
Data augmentation	random horizontal flip ($p = 0.5$) + random color jitter ($p = 0.8$, brightness $= 0.4$, contrast $= 0.4$, saturation $= 0.4$, hue $= 0.1$) + random grayscale conversion ($p = 0.2$) + random resized crop (scale = (0.2,1.0)), p being a probability.

Figure A illustrates examples of face images generated by StyleGAN for all attributes. Figure B illustrates examples of face images generated by StyleGAN3 for part of the 40 attributes. Table A shows the references and boundaries used in face manipulation for each attribute.

A.2. Datasets for Downstream Tasks

Facial Expression Change Dataset: We use the public dataset DISFA [21, 22]. We apply a facial detection method [15] to face videos taken by right camera and extract images containing the entire face (see Fig. 4 of the main paper). The action unit (AU) intensity was annotated for each video frame with six levels from 0 (not present) to 5 (maximum intensity) for several AUs. Annotation was performed by a coder certified in use of the Facial Action Coding System (FACS) [12]. All participants gave informed consent. Twenty-five of the 27 gave permission for use of their images in publications. In our main paper, we include facial images of subjects who gave permission for publication. See Ref. [22] for other details.

Weight Change Dataset: We use the dataset collected in Ref. [1] (Edema-A) and our newly collected dataset (Edema-B). These datasets contain face images and weight data acquired on several dialysis days per patient. As in Ref. [1], we apply a facial detection method [15] to face videos and extract images containing the center of face. For both datasets, all procedures in the studies were approved by the Ethical Review Board and all patients gave informed consent. In the training phase, we perform transfer learning on weight change using all pairs within individuals, including pre- and post-dialysis (*i.e.*, comparing faces between pre-/pre-dialysis, post-/post-dialysis, and pre-/post-dialysis). In the test phase, we estimate weight change using pairs of pre- and post-dialysis (*i.e.*, comparing faces between pre-/post-dialysis only) according to Ref. [1].

Age Change Dataset: We use the public dataset FG-NET [24], which contains 1002 face images from 82 subjects with large variations of lighting, pose, and facial expression. We directly use face images from the original dataset. FG-NET has been used in many age estimation studies [6, 11, 29, 33]. One of the major aims of FG-NET project was to encourage research technology development in the area of face and gesture recognition by specifying and supplying suitable image sets. See Ref. [24] for other details.

B. Implementation Details

B.1. Details for FRL

Our ComFace settings for FRL are summarized in Table B. The data augmentation setting is similar to MoCo [14]. The model at the epoch with the lowest validation loss in FRL is used for the downstream tasks for comparing faces.

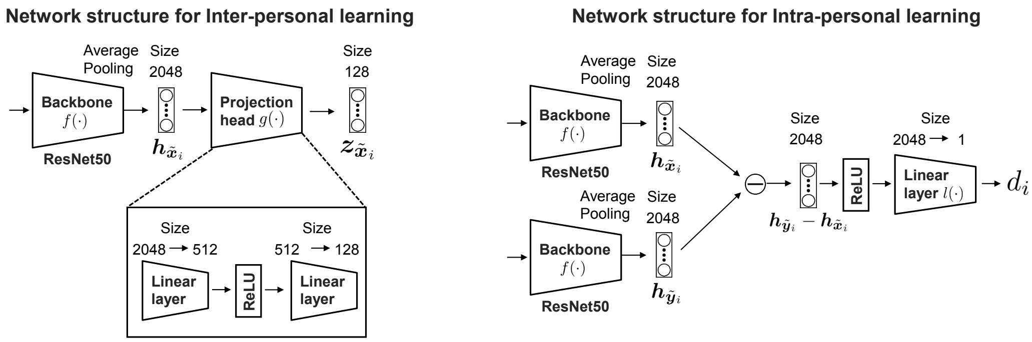

Figure C. Details of network structures for inter-personal and intra-personal learning.

Table C. Facial expression change estimation settings.

Config	Value
Batch size	16
Optimizer	Adam [18]
Learning rate	1.0e-5
Epochs	20 (linear evaluation) / 10 (fine-tuning)
Learning rate schedule	None
Computing resource	single NVIDIA GeForce RTX 3060 GPU
Image size	224×224
Data augmentation	random horizontal flip ($p = 0.5$) + random color jitter ($p = 0.8$, brightness $= 0.4$, contrast $= 0.4$, saturation $= 0.4$, hue $= 0.1$) + random grayscale conversion ($p = 0.2$) + random resized crop (scale = (0.5,1.0)), p being a probability.

Details of network structures for inter-personal and intra-personal learning are illustrated in Figure C. The network structure in intra-personal learning is similar to the Siamese network in Ref. [27].

B.2. Details for Downstream Tasks

We describe the implementation details for each downstream task. In transfer learning for all downstream tasks, we randomly sample $\boldsymbol{x}_i^{task}$ and $\boldsymbol{y}_i^{task}$ pairs for each epoch and construct mini-batches. The model at the epoch with the lowest validation loss in transfer learning is used for testing.

Facial Expression Change: The settings are summarized in Table C. The data augmentation setting is similar to ComFace setting for FRL. For fair comparisons, all comparative methods also follow the settings in Table C.

Weight Change: The settings are summarized in Table D. The data augmentation setting is the same as in the previous weight estimation method [1]. For fair comparisons, all comparative methods also follow the settings in Table D.

In the cross-dataset evaluation (Edema A→B in Table 2 of the main paper), the four models trained on Edema-A (from four-fold cross-validation) are tested on Edema-B. Table 2 of the main paper reports the average performance of the four models.

We also provide details of the experimental setup in comparison of ComFace with the previous method [1] (results are shown in Table 3 of the main paper). To ensure a fair comparison between ComFace and the previous method [1], we use the same test data for evaluation with the following setup.

- Method [1]: This method performs pre-training on multiple patient data and then builds patient-specific models via transfer learning on per-patient data. As in the original paper [1], the patient-specific model uses 24 patients for pre-training and 15 patients for transfer learning and testing. For each of the 15 patients in transfer learning, we perform a leave-one-day-out cross-validation, where the data from one day are used for testing and the data from the other days are used for training, as in Ref. [1]. In the training data, per-patient data on randomly selected dialysis days (from 1 to 3 days, as shown in Table 3 of the main paper) are used for transfer learning. With the leave-one-day-out cross-validation, data from all days are used for testing.

Table D. Weight change estimation settings.

Config	Value
Batch size	16
Optimizer	Adam [18]
Learning rate	1.0e-4
Epochs	10 (fine-tuning)
Learning rate schedule	None
Computing resource	single NVIDIA GeForce RTX 3060 GPU
Image size	224×224
Data augmentation	random horizontal flip ($p = 0.5$) + random color jitter ($p = 0.8$, brightness $= 0.4$, contrast $= 0.4$, saturation $= 0.4$, hue $= 0.1$) + random grayscale conversion ($p = 0.2$), p being a probability.

Table E. Age change estimation settings.

Config	Value
Batch size	16
Optimizer	Adam [18]
Learning rate	4.6e-4
Epochs	10 (fine-tuning)
Learning rate schedule	None
Computing resource	single NVIDIA GeForce RTX 3060 GPU
Image size	224×224
Data augmentation	random horizontal flip ($p = 0.5$) + random color jitter ($p = 0.8$, brightness $= 0.4$, contrast $= 0.4$, saturation $= 0.4$, hue $= 0.1$) + random grayscale conversion ($p = 0.2$) + random resized crop (scale = (0.5,1.0)), p being a probability.

- ComFace (Ours): We perform a four-fold cross-validation as in the other validations, and the same 15 patient data as in the method [1] are used for testing. As in method [1], data from all days in 15 patients are used for testing. Note that in our method, test patients are not included in the training data.

Age Change: The settings are summarized in Table E. The data augmentation setting is similar to ComFace setting for FRL. For fair comparisons, all comparative methods also follow the settings in Table E.

B.3. Summary of Comparative and Proposed Methods

Table F summarizes the training datasets, training scales, training sources, and backbones for all comparative and proposed methods.

Table G describes pre-trained checkpoints for comparative methods. In the same manner as our method, the comparative methods perform transfer learning (linear evaluation or fine-tuning) from the pre-trained weights.

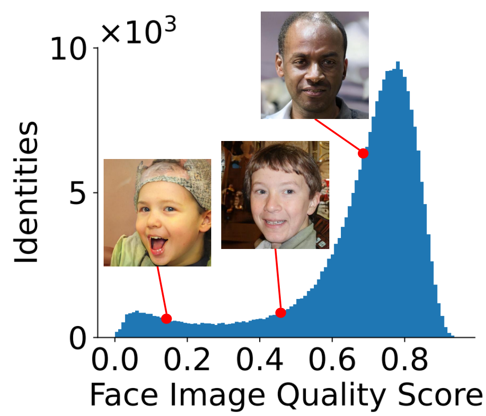

Figure D. FIQ score

C. Results

C.1. Quality Assessment of Synthetic Images

We confirmed the quality of synthetic images with face image quality (FIQ) assessment [30] as in previous synthetic data study for face recognition [13]. Figure D represents a histogram of FIQ scores for identities in synthetic images. We see that the majority of identities are good quality (we empirically confirmed the quality is good when the score ≥ 0.6). Table H shows transfer performance when

Table F. Training datasets, training scales, training sources, and backbones for all comparative and proposed methods.

Method	Dataset	Training Scale	Training Source	Backbone
Scratch	-	-	-	ResNet50
General Pre-training:				
ImageNet [10]	ImageNet	1.28M	Images+Human labels	ResNet50
VGGFace2 [4]	VGGFace2	3.31M	Images+Human labels	ResNet50
Visual Representation Learning:				
SimCLR [7]	ImageNet	1.28M	Images	ResNet50
MoCo v2 [8, 14]	ImageNet	1.28M	Images	ResNet50
SwAV [5]	ImageNet	1.28M	Images	ResNet50
Barlow Twins [32]	ImageNet	1.28M	Images	ResNet50
Facial Representation Learning:				
Bulat *et al.* [3]	VGGFace	3.4M	Face images	ResNet50
FaRL [34]	LAION-FACE [34]	20M	Face images+Text	ViT-B/16
PCL [19]	VoxCeleb1 [23]+VoxCeleb2 [9]	unknown	Face images	16-layer CNN
ComFace (Ours)	Synthetic data	35M	Synthetic face images+Intensity α	ResNet50

Table G. Pre-trained checkpoints for comparative methods.

Method	Backbone	Pre-trained checkpoint
General Pre-training:		
ImageNet [10]	ResNet50	resnet50 [7]
VGGFace2 [4]	ResNet50	resnet50_scratch_weight.pkl [8]
Visual Representation Learning:		
SimCLR [7]	ResNet50	resnet50-1x.pth [9,10]
MoCo v2 [8, 14]	ResNet50	moco_v2_200ep_pretrain.pth.tar [11]
SwAV [5]	ResNet50	swav_200ep_pretrain.pth.tar [12]
Barlow Twins [32]	ResNet50	resnet50.pth [13]
Facial Representation Learning:		
Bulat *et al.* [3]	ResNet50	flr_r50_vgg_face.pth [14]
FaRL [34]	ViT-B/16	FaRL-Base-Patch16-LAIONFace20M-ep16.pth [15]
PCL [19]	16-layer CNN	best.pth [16]

[7] https://github.com/huggingface/pytorch-image-models, timm==0.4.12
[8] https://github.com/cydonia999/VGGFace2-pytorch
[9] https://github.com/google-research/simclr
[10] Original TensorFlow checkpoints were converted to PyTorch format by https://github.com/tonylins/simclr-converter
[11] https://github.com/facebookresearch/moco
[12] https://github.com/facebookresearch/swav
[13] https://github.com/facebookresearch/barlowtwins
[14] https://github.com/1adrianb/unsupervised-face-representation
[15] https://github.com/FacePerceiver/FaRL
[16] https://github.com/DreamMr/PCL

using synthetic images of identities selected based on FIQ scores. We find that it is better to use all synthetic images for pre-training without restricting by FIQ scores (our final setting). Therefore, only high quality images are not necessarily required for representation learning.

C.2. Evaluation for Several Backbones

Table I compares performance in several backbones for ComFace. ResNet50 is more suitable for ComFace than ViT and other scales of ResNets. This could be because the dataset size for downstream tasks is not large and a medium-scale network performs better.

C.3. Evaluation for Other Major AUs

In addition to AU6 and 12 in the main paper, we evaluate the performance of the other major AUs. To directly evaluate the learned representation, we use linear evaluation. Table J shows correlation coefficients for several AUs. We find that ComFace is superior to the other methods for most AUs, demonstrating its generalization ability for a variety of facial expressions.

C.4. Settings of Linear Evaluation

In linear evaluation of the main paper, the backbone is frozen and the linear layer is trained from scratch. This

Table H. Transfer performance with respect to FIQ score v. Line indicated in gray is our final setting. Results are evaluated in fine-tuning.

Score v	AU6 Corr.↑	AU12 Corr.↑	Edema-A Acc.↑	Edema-B Acc.↑	Age Corr.↑
$v \geq 0$	0.663	0.831	88.6	96.3	0.870
$v \geq 0.1$	0.645	0.824	85.6	93.8	0.863
$v \geq 0.2$	0.647	0.826	88.7	95.9	0.845
$v \geq 0.4$	0.653	0.828	85.1	94.3	0.852
$v \geq 0.6$	0.641	0.831	86.6	95.7	0.852

Table I. Performance for several backbones. Results are evaluated in fine-tuning.

Backbone	AU6 Corr.↑	AU12 Corr.↑	Edema-A Acc.↑	Edema-B Acc.↑	Age Corr.↑
ResNet18	0.658	0.826	87.2	93.2	**0.870**
ResNet50	**0.663**	**0.831**	**88.6**	**96.3**	**0.870**
ResNet101	0.660	0.820	86.4	93.6	0.853
ViT-B/16	0.609	0.825	81.8	93.9	0.843

Table J. Correlation coefficients for several AUs

Method	AU1	AU2	AU4	AU5	AU9	AU17	AU20	AU25
General Pre-training:								
ImageNet [10]	0.209	0.075	0.177	0.004	0.065	-0.023	0.060	0.519
VGGFace2 [4]	0.196	0.213	0.403	0.022	0.017	0.128	-0.004	0.656
Visual Representation Learning:								
SimCLR [7]	0.292	0.190	0.330	-0.071	0.065	0.001	0.073	0.629
MoCo v2 [8, 14]	0.113	0.073	0.363	-0.030	0.089	0.016	0.014	0.637
SwAV [5]	0.256	0.192	0.370	0.049	0.008	-0.034	0.152	0.678
Barlow Twins [32]	0.251	0.177	0.495	0.070	0.106	0.013	0.104	0.690
Facial Representation Learning:								
Bulat *et al.* [3]	0.095	0.207	0.295	0.010	0.255	0.031	0.068	0.549
FaRL [34]	0.270	0.262	**0.627**	0.199	0.158	0.125	0.148	0.735
PCL [19]	0.303	0.122	0.229	0.024	0.007	0.030	-0.024	0.566
ComFace (Ours)	**0.443**	**0.468**	0.582	**0.451**	**0.463**	**0.155**	**0.160**	**0.746**

setting is the same for all methods in order to ensure fair comparisons. On the other hand, ComFace can also train the linear layer from the pre-trained weights. We therefore consider the following two settings here: In linear evaluation, (a) the backbone is frozen and the linear layer is trained from scratch (our final setting) (b) the backbone is frozen and the linear layer is trained from the pre-trained weights. Table K compares the performance of two settings in linear evaluation. It shows that setting(a) performs better than setting(b). This may be because the backbone is frozen and it is more reasonable to train the linear layer from scratch in order to match the linear layer with the backbone in transfer learning. Nevertheless, setting(b) still outperforms other comparative methods (see Table 1 of the main paper).

Table K. Performance in two linear evaluation settings: (a) backbone is frozen and linear layer is trained from scratch (b) backbone is frozen and linear layer is trained from pre-trained weights. Line indicated in gray is our final setting.

	AU6 Linear		AU12 Linear	
Linear Evaluation Setting	MAE↓	Corr.↑	MAE↓	Corr.↑
(a) Backbone: Frozen, Linear: Scratch	0.639	0.648	0.663	0.786
(b) Backbone: Frozen, Linear: Pre-trained	0.704	0.565	0.711	0.752

C.5. Settings of Fine-tuning

In fine-tuning of the main paper, the backbone and linear layer are trained from the pre-trained weights. To evaluate the effectiveness of training the linear layer from the pre-trained weights, we compare the following two settings here: In fine-tuning, (c) the backbone is trained from the pre-trained weights and the linear layer is trained from scratch (d) the backbone and linear layer are trained from the pre-trained weights (our final setting). Table L compares the performance of two settings in fine-tuning. It shows that setting(d) performs better than setting(c). This result suggests that the linear layer trained by intra-personal learning in FRL is useful for downstream tasks for comparing faces. To achieve the best performance, we use setting(d) in fine-tuning.

C.6. Comparison with Visual Representation Learning Using Synthetic Images

We compare ComFace with a recent visual representation learning method using synthetic data, StableRep [17] [31]. Table M shows the results of weight change estimation in representation learning methods using synthetic data. Although our model has a smaller training scale than StableRep, ComFace still has superior transfer performance. We expect that this is because the representations learned using synthetic face images from StyleGANs are more suitable for the estimation of intra-personal facial changes than those learned using general images from Stable Diffusion [26].

C.7. Full Results of Age Change Estimation

Table N shows the full results of estimating age change in fine-tuning. We can see that ComFace outperforms all other methods.

[17] The pre-trained checkpoint is cc12m_1x.pth from https://github.com/google-research/syn-rep-learn/tree/main/StableRep.

Table L. Performance in two fine-tuning settings: (c) backbone is trained from pre-trained weights and linear layer is trained from scratch (d) backbone and linear layer are trained from pre-trained weights. Line indicated in gray is our final setting.

Fine-tuning Setting	AU6 Corr.↑	AU12 Corr.↑	Edema-A Acc.↑	Edema-B Acc.↑	Age Corr.↑
(c) Backbone: Pre-trained, Linear: Scratch	0.666	0.829	84.4	91.9	0.841
(d) Backbone: Pre-trained, Linear: Pre-trained	0.663	0.831	88.6	96.3	0.870

Table M. Results of weight change estimation in representation learning methods using synthetic data. Results are evaluated in fine-tuning. Training scales, generative models for synthesis, and backbones are described.

				Edema-A			Edema-B			Edema-A→B		
Method	Scale	Generative model	Backbone	MAE↓	Corr.↑	Acc.↑	MAE↓	Corr.↑	Acc.↑	MAE↓	Corr.↑	Acc.↑
StableRep [31]	100M	Stable Diffusion	ViT-B/16	1.434	0.725	86.5	**1.439**	0.799	93.9	1.715	0.809	93.3
ComFace (Ours)	35M	StyleGANs	ResNet50	**1.394**	**0.750**	**88.6**	1.523	**0.801**	**96.3**	**1.668**	**0.819**	**93.8**

Table N. Results of estimating age change. Results are evaluated in fine-tuning.

Method	MAE↓	Corr.↑
Scratch	8.980	0.514
General Pre-training:		
ImageNet [10]	7.863	0.614
VGGFace2 [4]	8.783	0.511
Visual Representation Learning:		
SimCLR [7]	6.947	0.729
MoCo v2 [8, 14]	9.254	0.431
SwAV [5]	6.368	0.780
Barlow Twins [32]	6.686	0.758
Facial Representation Learning:		
Bulat *et al.* [3]	6.327	0.783
FaRL [34]	5.249	0.851
PCL [19]	8.998	0.451
ComFace (Ours)	**4.914**	**0.870**

References

[1] Y. Akamatsu, Y. Onishi, H. Imaoka, J. Kameyama, et al. Edema estimation from facial images taken before and after dialysis via contrastive multi-patient pre-training. *IEEE Journal of Biomedical and Health Informatics*, 27(3):1419–1430, 2023. 2, 3, 4

[2] Y. Alaluf, O. Patashnik, Z. Wu, A. Zamir, et al. Third time's the charm? image and video editing with StyleGAN3. In *Proc. European Conf. Computer Vision Workshops*, pages 204–220. Springer, 2022. 1, 2

[3] A. Bulat, S. Cheng, J. Yang, A. Garbett, et al. Pre-training strategies and datasets for facial representation learning. In *Proc. European Conf. Computer Vision (ECCV)*, pages 107–125. Springer, 2022. 5, 6, 7

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, et al. Vggface2: A dataset for recognising faces across pose and age. In *Proc. Int. Conf. Automatic Face & Gesture Recognition (FG)*, pages 67–74, 2018. 5, 6, 7

[5] M. Caron, I. Misra, J. Mairal, P. Goyal, et al. Unsupervised learning of visual features by contrasting cluster assignments. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9912–9924, 2020. 5, 6, 7

[6] P. Chen, X. Zhang, Y. Li, J. Tao, et al. DAA: A delta age adain operation for age estimation via binary code transformer. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 15836–15845, 2023. 2

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proc. Int. Conf. Machine Learning (ICML)*, pages 1597–1607, 2020. 5, 6, 7

[8] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 5, 6, 7

[9] J. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep speaker recognition. In *Proc. Interspeech*, 2018. 5

[10] J. Deng, W. Dong, R. Socher, L. Li, et al. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 5, 6, 7

[11] Z. Deng, H. Liu, Y. Wang, C. Wang, et al. PML: Progressive margin loss for long-tailed age classification. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 10503–10512, 2021. 2

[12] P. Ekman and W. V. Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 2

[13] M. Falkenberg, A. Bensen Ottsen, M. Ibsen, and C. Rathgeb. Child face recognition at scale: Synthetic data generation and performance benchmark. *arXiv preprint arXiv:2304.11685*, 2023. 4

[14] K. He, H. Fan, Y. Wu, S. Xie, et al. Momentum contrast for unsupervised visual representation learning. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 2, 5, 6, 7

[15] H. Imaoka, H. Hashimoto, K. Takahashi, A. F. Ebihara, et al. The future of biometrics technology: from face recognition to related applications. *APSIPA Transactions on Signal and Information Processing*, 10:e9, 2021. 2

[16] T. Karras, M. Aittala, S. Laine, E. Härkönen, et al. Alias-free generative adversarial networks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 852–863, 2021. 1

[17] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 1

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2, 3, 4

[19] Y. Liu, W. Wang, Y. Zhan, S. Feng, et al. Pose-disentangled contrastive learning for self-supervised facial representation. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 9717–9728, 2023. 5, 6, 7

[20] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2015. 1

[21] S. M. Mavadati, M. H. Mahoor, K. Bartlett, and P. Trinh. Automatic detection of non-posed facial action units. In *Proc. Int. Conf. Image Processing (ICIP)*, pages 1817–1820, 2012. 2

[22] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, et al. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 2

[23] A. Nagrani, J. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *Proc. Interspeech*, 2017. 5

[24] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes. Overview of research on facial ageing using the FG-NET ageing database. *IET Biometrics*, 5(2):37–46, 2016. 2

[25] V. Pinnimty, M. Zhao, P. Achananuparp, and E. Lim. Transforming facial weight of real images by editing latent space of StyleGAN. *arXiv preprint arXiv:2011.02606*, 2020. 1, 2

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, et al. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 6

[27] O. Schlesinger, N. Vigderhouse, D. Eytan, and Y. Moshe. Blood pressure estimation from ppg signals using convolutional neural networks and siamese network. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 1135–1139, 2020. 3

[28] Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the latent space of GANs for semantic face editing. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 9243–9252, 2020. 1, 2

[29] N. Shin, S. Lee, and C. Kim. Moving window regression: A novel approach to ordinal regression. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 18760–18769, 2022. 2

[30] P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuchner, et al. SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 5651–5660, 2020. 4

[31] Y. Tian, L. Fan, P. Isola, H. Chang, and D. Krishnan. StableRep: Synthetic images from text-to-image models make strong visual representation learners. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 6, 7

[32] J. Zbontar, L. Jing, I. Misra, Y. LeCun, et al. Barlow Twins: Self-supervised learning via redundancy reduction. In *Proc. Int. Conf. Machine Learning (ICML)*, pages 12310–12320, 2021. 5, 6, 7

[33] C. Zhang, S. Liu, X. Xu, and C. Zhu. C3AE: Exploring the limits of compact model for age estimation. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 12587–12596, 2019. 2

[34] Y. Zheng, H. Yang, T. Zhang, J. Bao, et al. General facial representation learning in a visual-linguistic manner. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 18697–18709, 2022. 5, 6, 7