PrefAce: Face-Centric Pretraining with Self-Structure Aware Distillation

Siyuan Hu[1*], Zheng Wang[2], Peng Hu[3], Xi Peng[3], Jie Wu[2], Hongyuan Zhu[4†], Yew Soon Ong[1, 4]

[1]Nanyang Technological University,
[2]Wuhan University,
[3]Sichuan University,
[4]Institute for Infocomm Research (I^{2}R) & Centre for Frontier AI Research (CFAR), A*STAR, Singapore
husi0002@e.ntu.edu.sg, wangzwhu@whu.edu.cn, penghu.ml@gmail.com, pengx.gm@gmail.com, whu_wujie@163.com, hongyuanzhu.cn@gmail.com, asysong@ntu.edu.sg

Abstract

Video-based facial analysis is important for autonomous agents to understand human expressions and sentiments. However, limited labeled data is available to learn effective facial representations. This paper proposes a novel self-supervised face-centric pretraining framework, called PrefAce, which learns transferable video facial representation without labels. The self-supervised learning is performed with an effective landmark-guided global-local tube distillation. Meanwhile, a novel instance-wise update FaceFeat Cache is built to enforce more discriminative and diverse representations for downstream tasks. Extensive experiments demonstrate that the proposed framework learns universal instance-aware facial representations with fine-grained landmark details from videos. The point is that it can transfer across various facial analysis tasks, *e.g.*, Facial Attribute Recognition (FAR), Facial Expression Recognition (FER), DeepFake Detection (DFD), and Lip Synchronization (LS). Our framework also outperforms the state-of-the-art on various downstream tasks, even in low data regimes. Code is available at https://github.com/siyuan-h/PrefAce.

Introduction

Facial analysis tasks offer valuable insights into human non-verbal behavior, shedding light on social interaction (Haugh 2009), communication (Jack and Schyns 2015), and cognition (Storrs, Anderson, and Fleming 2021), with implications for Human-Computer Interaction (HCI) and Affective Computing. Recent strides in deep neural network models, encompassing Facial Attribute Recognition (FAR) (Zhu et al. 2022), Facial Expression Recognition (FER) (Li and Deng 2022), DeepFake Detection (DFD) (Tolosana et al. 2020), and Lip Synchronization (LS) (Kadam et al. 2021), have shown remarkable potential. However, the need for extensive annotated datasets poses challenges due to resource and time demands, especially for specialized applications requiring domain expertise for annotation like Facial Expression Recognition (FER).

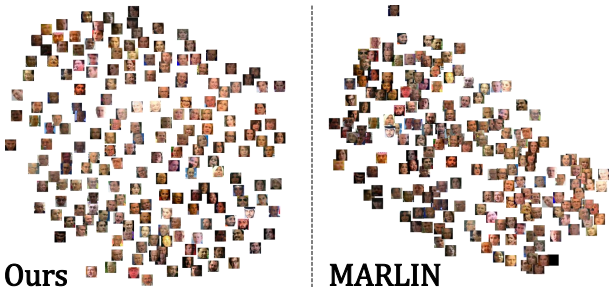


Figure 1: Distribution overview of t-SNE visualization on embeddings generated by PrefAce and state-of-the-art.

Self-supervised pretraining is a powerful solution to bypass the limitations of full supervision. Utilizing non-annotated data for generic representation learning, it transfers knowledge across tasks, enabling further training with limited annotation. In natural scene imagery, self-supervised approaches like self-distillation (Caron et al. 2021a), contrastive learning (Chen et al. 2020b), jigsaw puzzle solving (Noroozi and Favaro 2016), and masked autoencoders (He et al. 2022) have even outperformed supervised methods.

In facial analysis, prevailing methods often remain specialized and fully supervised. Recently, (Bulat et al. 2022) studied the efficacy of common self-supervised techniques in scalable, generic representation learning from uncurated images and videos. (Zheng et al. 2022) studied the self-supervised facial video representation learning in the setting of visual-linguistic pretraining. One step ahead, MARLIN (Cai et al. 2023) introducing facial region masking and adversarial training for face reconstruction to learn representations from videos modalities only which alleviate the requirement of image-caption pairs in (Zheng et al. 2022).

On the other hand, recent studies (Oquab et al. 2023) demonstrate that performing vanilla self-supervised learning on uncurated data will lead to inferior representations with less transferable to other tasks. Our empirical study of recent SOTA face pretraining method (e.g. MARLIN) also reveals that this phenomenon that it tends to learn common face features while ignores diverse and discriminative features (see

*This paper was completely accomplished when Hu Siyuan interned under Dr. Zhu Hongyuan's supervision at A*STAR.
†Corresponding author.

Fig. 2) which are useful for downstream tasks. Therefore, it is interesting to explore to balance learning both common and diverse face representations in the setting of uncurated data.

This paper propose a new method PrefAce aiming at enhancing learning *universal* and *task-agnostic* representations in a self-supervised manner for downstream facial analysis tasks. In PrefAce, we adopt the dual encoder architecture which is popular in self-supervised learning and compatible with most architectures (e.g. CNN and ViT). Moreover, we introduce a new multiscale landmark guided self-distillation to help the network focus on spatio-temporal fine-grained discriminative features. We also introduce a novel one-time FaceFeat Cache so that the encoder can try to produce diverse features by constrasting current video face instance with instances in all other videos. Our experimental results show that our proposed framework, PrefAce, learns highly generic facial encoding that scale and transfers well across diverse facial analysis tasks such as FER, DFD, FAR, and LS and achieve favorable performance gain w.r.t. state-of-the-art benchmarks. In summary, our main contributions are:

- We propose, PrefAce, a *universal* and *task-agnostic* self-supervised facial representation learning framework that learns common and diverse facial representation from uncurated web-crawled facial videos.
- We propose multiscale landmark guided self-distillation. The proposed strategy aims to learn fine-grained discriminative facial representations.
- We propose to use novel FaceFeat cache with instance-wise updating strategy to learn diverse and transferrable facial representations.
- Extensive quantitative and qualitative analysis demonstrate that PrefAce learns rich, generic, transferable, and robust facial representation, that outperforms state of the art across a variety of downstream tasks including FAR, FER, DFD, LS, and under few shot settings.

Related Work

Self-supervised Learning

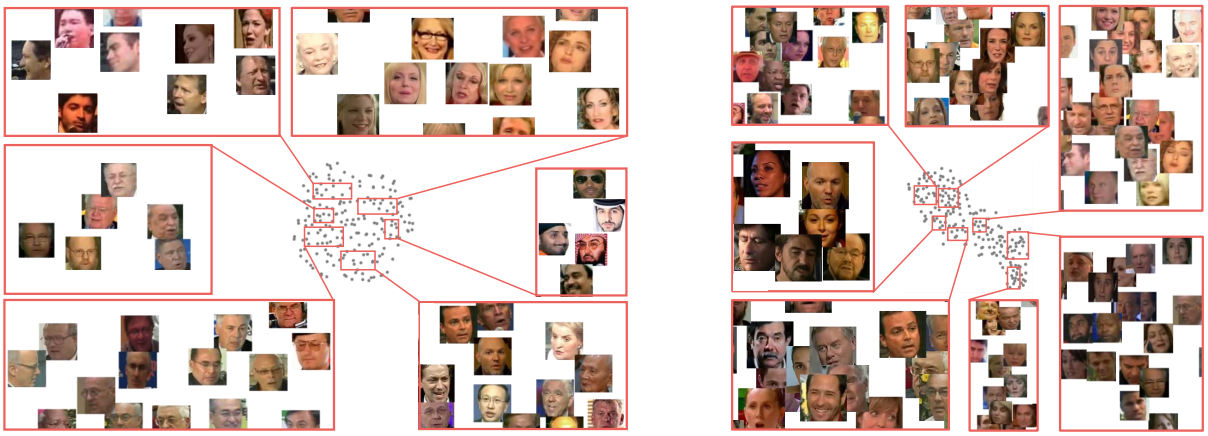
A large body of work on self-supervised learning focuses on discriminating between augmentations of instance (Chen et al. 2020a), thus learning the underlying invariance of representations. Recent works have shown that we can learn unsupervised features without discriminating between instances. Grill et al. (Grill et al. 2020) propose a metric-learning formulation called BYOL, where features are trained by matching them to representations obtained with a momentum encoder. Several other works echo this direction, showing that one can match more elaborate representations (Gidaris et al. 2020), train features matching them to a uniform distribution (Bojanowski and Joulin 2017) or by using whitening (Zbontar et al. 2021). Of particular interest, DINO (Caron et al. 2021b) proposes self-distillation, where probability outputs of dual encoders are matched on different regions. Moreover, the emergence of patch-based architectures, like ViTs, has led to a revisit of inpainting for pretraining (Bao, Dong, and Wei 2021), potentially in feature space (Assran et al. 2023). (He et al. 2021) shows that a masked autoencoder (MAE) learns features that provide substantial improvements when finetuned on downstream tasks.

However, all aforementioned works are image-level frameworks. When it comes to video-level self-supervised learning, most existing works are based on pixel information reconstruction, either in frame prediction (Pan et al. 2021), reconstruction (Tian et al. 2020), or masked autoencoder way (Tong et al. 2022). None is based on self-distillation. Different from all previous works, our approach proposes self-distillation for video learning, demonstrating that global-local distillation can well capture the spatio-temporal pattern of fine-grained regions in facial videos, thus further enhancing downstream task performance.

In contrastive learning framework, memory banks can be used in both supervised and self-supervised learning with different motivations. (Li et al. 2019) uses a memory bank to capture context information, and (He et al. 2020) maintains a memory bank by enqueueing negative samples in mini batches, thus equivalently increasing batch size. Different from previous works, we propose a memory dictionary called FaceFeat Cache, and utilize the memory to record facial characteristics of all videos, thus helps the network focus on the discriminative features of video instances in a contrast manner. Through extensive quantitative and qualitative analysis, it is proven that our proposed FaceFeat Cache can facilitate the network to generate more diverse, informative embeddings for downstream tasks.

Facial Representation Learning

Till date, most of the existing facial analysis approaches are conducted in a task-specific way with fully supervised manner on manually annotated data to enhance performance. Any state-of-the-art model's performance on benchmarked datasets is impacted by the quality and quantity of annotated data used during training. Various task-specific large-scale facial datasets have been curated over the past decade to facilitate research in Face Verification (MS-celeb-1M (Guo et al. 2016)), Facial Attribute Recognition(CelebA (Liu et al. 2015), CelebV-HQ (Zhu et al. 2022)), Facial Emotion Recognition (CMU-MOSEI (Bagher Zadeh et al. 2018)), DeepFake Detection (FF++ (Rossler et al. 2019)) and Lip Synchronization (LRS2 (Chung and Zisserman 2017)). However, data curation encounters several challenges such as requirements of specialized hardware (e.g. for FER and action unit data), the discrepancy in data distribution that prevent merging of multiple datasets (Bulat et al. 2022), and most importantly time consuming and resource expensive annotation process. To eliminate these drawbacks, some of the existing approaches adopt data augmentation via image or video synthesis as the surge in face generation fueled by Generative Adversarial Network (GAN) (Skorokhodov, Tulyakov, and Elhoseiny 2022) and other generation techniques (Hao et al. 2021) aids realistic face generation even with control over facial attributes. These generation techniques add variation in training set quantitatively, but in some cases it still lags in qualitative aspects due to do-



a) t-SNE visualization of PrefAce b) t-SNE visualization of MARLIN

Figure 2: t-SNE visualization of embedding groups generated by PrefAce and MARLIN. One can observe that embeddings generated by PrefAce are evenly distributed, with almost no overlap in between, and naturally grouped by attributes, e.g., gender, expression, head pose, hair color, etc. However, embeddings from MARLIN are overlapping and blurred in attributes. The aforementioned findings indicate that PrefAce can better capture characteristics of different faces.

main specific inconsistency and more importantly high network complexity.

To this end, there are very few recent works that aim to learn *image-based* task specific facial encoding with limited supervision. The most closely related existing works either focus on exploring training dataset properties in terms of size and quality (Bulat et al. 2022) or performing pretraining in visual-linguistic way (Zheng et al. 2022). These works are hard to transfer to video-related tasks since they use static image level facial information or require the image-caption pairs.

To tackle these problems, MARLIN perform facial representation learning using a masked autoencoder way using video modality only. However, we find that facial region-guided tube masking introduced by MARLIN will be degraded to entire face masking due to its extreme mask ratio, leading to an unsatisfactory loss of granularity. Different from MARLIN, we propose that self-distillation is more suitable for facial representation learning, as coarse-to-fine and spatial-temporal features can be properly learnt with global-local distillation as demonstrated in experiment.

Proposed Method

Our objective is to learn universal, informative and transferable facial representations from abundantly available non-annotated facial video data (Wolf, Hassner, and Maoz 2011). Different from pretraining of natural scenes, face specific tasks present two characteristics: a) Facial data has various layout (nose, eyes, lips, hair, etc.) due to different facial expressions. Beyond learning the appearance of individual landmarks, it is important to learn about the relationships between the local landmarks and the global face. b) Facial data not only have common attributes among demographics but also with specific attributes (e.g. eye color, shape and texture) to differentiate between instances, thus requiring more informative representations for downstream tasks.

To this end, we propose PrefAce, an unsupervised face-centric pretraining framework by focusing the backbone network to learn representations that not only can preserver the local-to-global and shared-to-specific facial attributes with an introduction of its architecture and loss functions.

Architecture

PrefAce consists of a teacher network (g_{θ_t} with parameters θ_t) and a student network (g_{θ_s} with parameters θ_s) respectively, which follows recent conventional setting of self-supervised learning works (Caron et al. 2021a) to learn complementary information from two augmented views. Different from previous work with image-level learning only, our proposed framework focus on video-level learning.

Given a training dataset $\mathbb{D} = \{V_i\}_{i=1}^{N}$ where N is the number of videos in the dataset and $V \in \mathbb{R}^{C\times T\times H_0\times W_0}$ are global video clips (C, T, H_0, W_0 are channel, temporal depth, height and width of the raw video, respectively). The derived embeddings of global frame denoted as $\{X_{global} \in \mathbb{R}^{\frac{T}{\mathbf{t}}\times\frac{H_0}{\mathbf{h}}\times\frac{W_0}{\mathbf{w}}}\}$.

Besides global frame, random local regions are also cropped as a complement. Across all the frames of the input V, we track specific facial regions v from the pre-defined set to encode the spatio-temporal changes and model landmark motion, and randomly crop local tubes to further facilitate the network to learn changes of skin correlated to landmark motion. The local frames are also embedded as $\{X_{local} \in \mathbb{R}^{\frac{T}{\mathbf{t}}\times\frac{H}{\mathbf{h}}\times\frac{W}{\mathbf{w}}}\}$, respectively.

Moreover, we perform face parsing which segments facial regions into the following parts: left eye, right eye, nose, mouth, hair, skin and background, using (Wang et al. 2021) given its efficiencies. Among the facial regions, we focus on the following set $\mathbb{P} = \{\texttt{left-eye}, \texttt{right-eye}, \texttt{nose}, \texttt{mouth}, \texttt{hair}\}$ over skin and background to preserve face specific local details. The cropping strategy thus facilitate the network

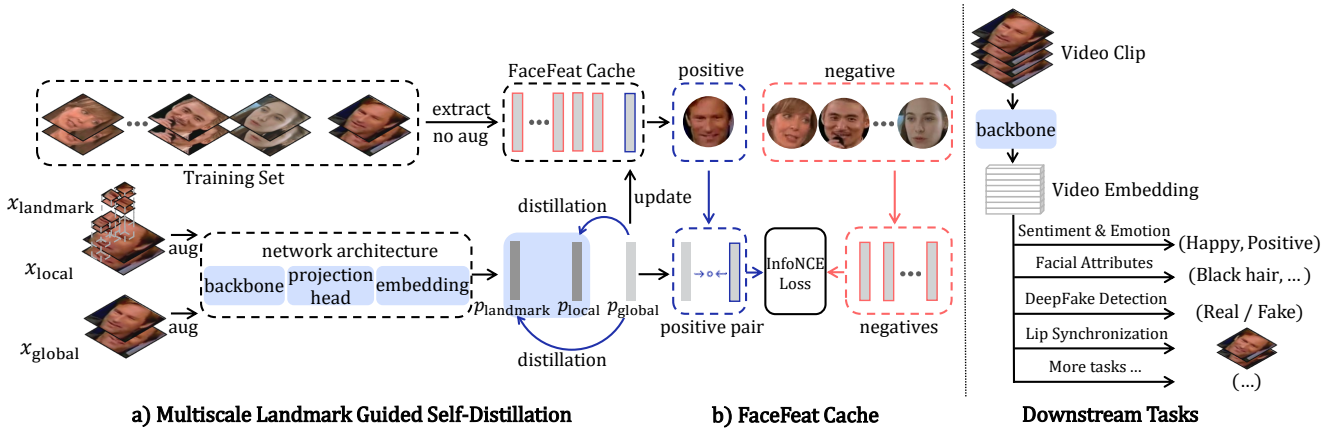


Figure 3: Overview of PrefAce. PrefAce mainly consists of two parts: (a) Multiscale Landmark Guided Self-Distillation: aids PrefAce to perform knowledge distillation between student and teacher network under the guidance of spatio-temporal landmark guided tube cropping, encouraging the network to learn spatio-temporal and coarse-to-fine consistent representations. (b) FaceFeat Cache: facilitates PrefAce to update video instance features so that the network can learn diverse representations.

consistently focus on the landmarks while tracking and encoding correlated skin details. The final embeddings will be $\{X_{landmark} \in \mathbb{R}^{\frac{T}{\mathbf{t}} \times \frac{H}{\mathbf{h}} \times \frac{W}{\mathbf{w}}}\}$, respectively.

Lastly, to enhance the quality of facial representation generated for downstream tasks, the network should focus on the meaningful parts of faces, e.g., fine-grained differences of shape, texture and motion in landmarks between instances. Therefore, we employ a memory dictionary as FaceFeat Cache to store representation of all video face instances. The memory is initialized using representations of videos without augmentation from the teacher network at the beginning of the epoch, and momentum updated during training. The network is expected to perfectly match the representations of videos' different versions as they correspond to the same instance, while minimize the similarity with other videos with different instances, encouraging the network to focus on the characteristics of faces.

Loss Function

PrefAce mainly optimized the following loss function:

$$\mathcal{L} = \mathcal{L}_{distill} + \mathcal{L}_{ID} \tag{1}$$

with (a) Multi-scale Landmark Guided Self-distillation Loss $\mathcal{L}_{distill}$ and (b) Instance Learning Loss $\mathcal{L}_{ID}$.

Multi-scale Landmark Guided Self-distillation Loss The learning objective is to train a student network g_{θ_s} to match the output of a given teacher network g_{θ_t}, parameterized by θ_s and θ_t respectively. As the inputs of the student and teacher networks are different augmentations of the same video, self-distillation aims to learn the spatio-temporal invariance despite cropping and random distortion. Given an input video clip $v \in \mathbb{R}^{(C \times T \times H \times W)}$, the video is mapped to embeddings denoted as $\{X \in \mathbb{R}^{\mathbf{k} \times \mathbf{e}}\}$, where $\mathbf{e}$ is the embedding dimension and $\mathbf{k}$ is the total number of tokens derived from v, i.e. $\mathbf{k} = \frac{T}{\mathbf{t}} \times \frac{H}{\mathbf{h}} \times \frac{W}{\mathbf{w}}$. With 3D video embeddings, both networks take the 2D mean as input and generate probability distributions over K dimensions denoted by P_s and P_t The probability P is obtained by normalizing the video mean representation of the network g with a softmax function. More precisely,

$$P_s(x)^{(i)} = \frac{\exp(\overline{g}_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^{K} \exp(\overline{g}_{\theta_s}(x)^{(k)}/\tau_s)}, \tag{2}$$

with $\tau_s > 0$ a temperature parameter that controls the sharpness of the output distribution, and a similar formula holds for P_t with temperature τ_t. Given a fixed teacher network g_{θ_t}, we learn to match these distributions by minimizing the cross-entropy loss w.r.t. the parameters of the student network θ_s:

$$\min_{\theta_s} H(P_t(x), P_s(x)), \tag{3}$$

where $H(a, b) = -a \log b$.

After aforementioned landmark guided tube crop, the augmented set contains two *global* views, x_1^g and x_2^g and several *local* views of smaller resolution. All crops are passed through the student while only the *global* views are passed through the teacher. As the cropping is performed in a tube manner, our video-level self-distillation encourages spatio-temporal "local-to-global" correspondences. We minimize the loss:

$$\mathcal{L}_{distill} = \min_{\theta_s} \sum_{x \in \{x_1^g, x_2^g\}} \sum_{\substack{x' \in V \\ x' \neq x}} H(P_t(x), P_s(x')). \tag{4}$$

Instance Learning Loss The Instance Learning loss is in the form of InfoNCE loss (He et al. 2020). The objective of Instance Learning in PrefAce is to maximize the similarity between representations of variations of the same video, while minimize the similarity between representations of different videos, aiming to capture the instance-consistent characteristics of facial videos. At the beginning of each epoch, PrefAce will calculate instance representation ϕ as memory initialization for all videos in the training dataset, and no augmentation is performed in order to preserve representation

Method	Appearance		Action		Overall
	Acc.↑	AUC↑	Acc.↑	AUC↑	Acc.↑
R3D*	92.34	0.9424	94.57	0.9173	93.45
MViTv1*	92.90	0.9452	95.13	0.9233	94.01
MViTv2*	92.77	0.954	95.15	0.9239	93.96
VideoMAE (FT)	92.91	0.9529	95.37	0.9284	94.14
MARLIN (LP)	91.90	0.9373	95.25	0.9278	93.57
PrefAce (LP)	**92.31**	**0.9427**	**95.55**	**0.9309**	**93.93**
MARLIN (FT)	93.90	0.9561	95.48	0.9406	94.69
PrefAce (FT)	**94.84**	**0.9590**	**96.35**	**0.9462**	**95.60**

Table 1: Facial Attribute Recognition. Our proposed framework, PrefAce, trained on YTF dataset and Linear Probed/Finetuned on CelebV-HQ benchmark dataset in terms of accuracy↑ and area under the curve↑. * shows supervised methods trained on the CelebV-HQ dataset.

correctness. Note that there are two global views for each video, thus we calculate the Instance Loss for both views and take the mean:

$$\mathcal{L}_{g1} = -\log \frac{exp(x_1^g \cdot \phi_+/\tau)}{\sum_{k=1}^{K} exp(x_1^g \cdot \phi_k/\tau)}, \tag{5}$$

$$\mathcal{L}_{g2} = -\log \frac{exp(x_2^g \cdot \phi_+/\tau)}{\sum_{k=1}^{K} exp(x_2^g \cdot \phi_k/\tau)}, \tag{6}$$

$$\mathcal{L}_{ID} = \frac{1}{2}(\mathcal{L}_{g1} + \mathcal{L}_{g2}). \tag{7}$$

At each step, FaceFeat Cache will be updated consistently by corresponding video mean representations, with a momentum m controlling the update speed:

$$\forall x \in X, \phi_k \leftarrow m\phi_k + (1-m) \cdot \frac{1}{2}(\overline{g}_{x_1} + \overline{g}_{x_2}). \tag{8}$$

Experiment

Finetuning Details

Our proposed PrefAce framework learns universal and transferable facial representation from videos in a self-supervised way. Following standard evaluation protocols, we perform Linear Probing (LP) and Fine-Tuning (FT) on teacher network for downstream adaptation in different tasks. Given any task specific downstream dataset $\mathbb{D}_{down} = \{v_j, \mathbf{y}_j\}_{j=1}^{\mathcal{N}}$, we deploy linear fully-connected (FC) layers with embedding parameters θ_{FC} to align the latent space to downstream task specific label space on top of the teacher network g_{θ_t}. For Linear Probing, we freeze the teacher network g_{θ_t} and only update θ_{FC}. On the other hand for Finetuning, we finetune the entire teacher network i.e. $(g_{\theta_t} \circ \theta_{FC})$.

Implementation Details The network is trained with PyTorch (Paszke et al. 2019) on Nvidia A100 GPUs. First of all, given a temporal chunk of facial videos, consecutive frames are highly redundant. Therefore, we set the minimum temporal stride to 2 to capture semantically meaningful frames with significant motion across frames. Given an input video with dimension $3 \times 16 \times 224 \times 224$, with the multiscale landmark guided tube crop strategy, the augmented input set contains 2 global video views of size $3 \times 16 \times 224 \times 224$ and 10 local video views of size $3 \times 16 \times 96 \times 96$. The cube embedding layer generates $8 \times 14 \times 14$ 3D tokens, each of dimension $2 \times 16 \times 16$ to preserve spatio-temporal patterns. Afterwards, each token is mapped to the latent embedding of dimension 768. PrefAce's objective is to match the output of student network (g_{θ_s}) to the output of teacher network (g_{θ_t}). Moreover, the similarity between facial embeddings of different videos shall be minimized. For fair comparison, we use ViT-Base as the backbone encoder, and the impact of ViT-variants is depicted in ablation study.

The pretraining hyperparameters are as follows: the base learning rate is linearly scaled with respect to the overall batch size, $lr = base\ learning\ rate \times batch\ size/256$. For self-supervised pretraining, we use AdamW optimizer with base learning rate $7.5e{-}4$, momentum $\beta_1 = 0.9, \beta_2 = 0.999$ with a learning rate scheduler (cosine decay). For LP and FT, we use AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.98$ and base learning rate $1e{-}4$, without weight decay.

Tasks	Pretrain	Method	Mod.	Acc.↑
Sentiment (2-Class)	MOSEI and IEMOCAP	CAE-LR	V	71.06
	YTF	VideoMAE	V	72.96
	YTF	MARLIN	V	73.70
	YTF	PrefAce	V	**74.05**
Sentiment (7-Class)	–	MViTv1*	V	33.35
	YTF	VideoMAE	V	33.78
	YTF	MARLIN	V	34.63
	YTF	PrefAce	V	**35.03**
Emotion	–	MViTv1*	V	80.45
	–	UMONS*	LAV	80.68
	–	GMF*	LAV	81.14
	YTF	VideoMAE	V	80.39
	YTF	MARLIN	V	80.60
	YTF	PrefAce	V	**81.29**

Table 2: Facial Expression and Sentiment Recognition. Downstream adaptation results on MOSEI dataset for sentiment (2-class), sentiment (7-class), and emotion analysis. Our proposed method, PrefAce, outperforms visual modality based sentiment and emotion prediction methods. *Please note that SOTA for UMON and GMF utilize three modalities and thus, not directly comparable.* Here, YTF: YouTube Faces dataset and LAV represents linguistic, audio, and visual modality, respectively. * denotes supervised methods.

Tasks and Evaluations

The details of different downstream facial tasks. For fair comparison, we follow the dataset specific experimental protocols mentioned in task specific prior literatures. Besides standard evaluations, we also perform few shot adaptation to demonstrate the robustness and transferability of PrefAce.

Facial Attribute Recognition (FAR) predicts appearance and action attributes such as gender, race, hair color, and emotion of a given face video. The problem of predict-

Pretrain	Method	Acc.(%)↑	AUC↑
–	Steg.Features*	55.98	–
–	LD-CNN*	58.69	–
–	Constraied Conv.*	66.84	–
–	CustomPooling CNN*	61.18	–
–	MesoNet*	70.47	–
–	Face X-ray*	–	0.6160
–	Xception*	86.86	0.8930
–	P3D*	–	0.6705
–	R3D*	–	0.8772
–	I3D*	–	0.9318
YTF	VideoMAE	87.57	0.9082
YTF	MARLIN	89.43	0.9305
YTF	PrefAce	**92.30**	**0.9371**

Table 3: Deepfake Detection. We compare the Fine-Tuning (FT) results on PrefAce for FaceForensic++ dataset. * denotes supervised methods.

ing facial attributes can be posed as a multi-label classification highly dependent on rich spatial encoding. For downstream adaptation, we use 28,532 train, 3,567 val, and 3,567 test videos from CelebV-HQ dataset. Following prior work (Zheng et al. 2022), we report average accuracy(↑), Area Under the Curve (AUC↑) over all attributes.

Facial Expression Recognition (FER) task encodes spatio-temporal facial muscle movement patterns to predict sentiment (2-class and 7-class) and emotion (6-class) of the concerned subject given a facial video. We evaluate the performance of PrefAce on CMU-MOSEI dataset which is a conversational corpus having 16,726 train, 1,871 val, and 4,662 test data. Following prior work (Delbrouck et al. 2020), we use overall accuracy(↑) as the metric.

Deepfake Detection (DFD) task predicts spatio-temporal facial forgery given a facial video from FF++(LQ) dataset. For downstream adaptation, we use 3,600 train, 700 val, and 700 test sample videos from FF++(LQ) dataset. Following prior literature (Cai et al. 2022), we use accuracy(↑) and AUC(↑) as the evaluation metrics.

Lip Synchronization (LS) is another research area that requires facial region specific spatio-temporal synchronization. This downstream adaptation further elaborates the adaptation capability of PrefAce for face generation tasks. For adaptation, we replace the facial encoder module in Wav2Lip (Prajwal et al. 2020) with PrefAce, and adjust the temporal window accordingly i.e. from 5 frames to $\mathbf{T}$ frames. For evaluation, we use the LRS2 dataset having 45,838 train, 1,082 val, and 1,243 test videos. Following the prior literature (Wang et al. 2022), we use Lip-Sync Error-Distance (LSE-D↓), Lip-Sync Error-Confidence (LSE-C↑) and Frechet Inception Distance (FID↓) (Heusel et al. 2017) as evaluation matrices.

Quantitative Analysis

Comparison with SOTA Facial Analysis Tasks

We compare the performance of PrefAce with different downstream facial analysis tasks following standard task specific evaluation protocols.

Facial Attributes In Tab. 1, we compare the Linear Probing and Finetuning adaptation performance of PrefAce with popular transformers (i.e. MViT-v1 (Fan et al. 2021) and MViT-v2 (Li et al. 2022)) and CNNs (i.e. R3D (Tran et al. 2018)) on CelebV-HQ dataset. From the table, it is observed that PrefAce's Finetuned version outperforms supervised MViT-v2 transformer architecture by 2.07% (92.77% $\rightarrow$ 94.84%) on appearance attributes and 1.20% (95.15% $\rightarrow$ 96.35%) on action attributes, and outperforms unsupervised SOTA MARLIN transformer architecture by 0.94% (93.90% $\rightarrow$ 94.84%) on appearance attributes and 0.87% (95.48% $\rightarrow$ 96.35%) on action attributes. We credit PrefAce's performance gain to the multiscale landmark guided self-distillation that helps the network focus on key parts of input facial videos, and the FaceFeat Cache helps the network to encode informative characteristics from facial videos.

Emotion and Sentiment In Tab. 2, we compare the Linear Probing and Finetuning adaptation performance of sentiment and emotion analysis in terms of accuracy(↑) and AUC(↑) on CMU-MOSEI dataset. *Please note that the PrefAce uses only visual modality.* The results suggest that PrefAce performs competitively with SOTA methods (Cai et al. 2023; Koromilas and Giannakopoulos 2022; Li et al. 2022; Delbrouck et al. 2020), outperforming unsupervised SOTA MARLIN on all 3 tasks. Specifically, PrefAce outperforms unsupervised SOTA MARLIN by 0.35% (73.70% $\rightarrow$ 74.05%) on 2-class sentiment task, 0.40% (34.63% $\rightarrow$ 35.03%) on 7-class sentiment task and 0.69% (80.60% $\rightarrow$ 81.29%) on 6-class emotion task. These results also indicate that PrefAce learns highly generic, robust, and transferable feature representations.

DeepFake Detection In Tab. 3, we compare the performance of video manipulation on FaceForensics++ dataset and report results in terms of video-level accuracy(↑) and AUC(↑). Besides performing favorably against both the supervised SOTA methods (Afchar et al. 2018; Fridrich and Kodovsky 2012; Cozzolino, Poggi, and Verdoliva 2017; Bayar and Stamm 2016; Rahmouni et al. 2017; Li et al. 2020; Chollet 2017; Qiu, Yao, and Mei 2017; Tran et al. 2018; Carreira and Zisserman 2017), PrefAce outperforms unsupervised SOTA MARLIN by 2.87% (89.43% $\rightarrow$ 92.30%) on

Method	LSE-D↓	LSE-C↑	FID↓
Speech2Vid	14.230	1.587	12.320
LipGAN	10.330	3.199	4.861
Wav2Lip	7.521	6.406	4.887
AttnWav2Lip	7.339	6.530	–
Wav2Lip + ViT	8.996	2.807	13.352
Wav2Lip + ViT + VideoMAE	7.316	5.096	4.097
Wav2Lip + ViT + MARLIN	7.127	5.528	3.452
Wav2Lip + ViT + PrefAce	**7.013**	**5.785**	**3.264**

Table 4: Lip Synchronization. We compare Linear Probing (LP) and Fine-Tuning (FT) results on the LRS2 dataset.

Data→	MOSEI			FF++	CelebV-HQ	
Task→	Emo.	7-Sen.	2-Sen.	DeepFake	Appr.	Act.
Anno.%	Acc.↑	Acc.↑	Acc.↑	AUC↑	AUC↑	AUC↑
100%	**81.29** (80.60)	**35.03** (34.63)	**74.05** (73.70)	**0.9371** (0.9305)	**0.9427** (0.9373)	**0.9309** (0.9278)
50%	**81.04** (80.59)	**34.22** (33.73)	**73.35** (73.33)	**0.9000** (0.8681)	**0.9332** (0.9273)	**0.9330** (0.9270)
10%	**80.56** (79.89)	**33.59** (33.56)	72.03 (72.26)	**0.7817** (0.7459)	**0.9069** (0.8996)	**0.9286** (0.9201)
1%	**79.16** (78.61)	**30.34** (30.09)	**71.97** (71.89)	**0.6380** (0.6252)	**0.8574** (0.8423)	**0.9112** (0.9063)

Table 5: Few shot adaptation on different facial tasks. Comparison between PrefAce and SOTA MARLIN on sentiment & emotion analysis, DeepFake detection and appearance & action classification. Results of MARLIN are shown in brackets.

accuracy and 0.66% (93.05% → 93.71%) on AUC. PrefAce uses only video modality to detect anomalies, and outperforms unsupervised SOTA method significantly even under few-shot setting, which will be introduced shortly.

Lip Synchronization For fair comparison, we compare the following settings: *1) Wav2Lip+ViT:* Contribution of ViT architecture where the weights of ViT is trained from scratch on LRS2 dataset. *2) Wav2Lip+ViT+MARLIN:* Contribution of unsupervised SOTA MARLIN pretrained on YTF. *3) Wav2Lip+ViT+PrefAce:* Contribution of PrefAce pretrained on YTF, with (Si et al. 2021; Prajwal et al. 2020; Wang et al. 2022) and different design aspects. The experimental results are depicted in Tab. 4 with LSE-D↓, LSE-C↑ and FID↓ as evaluation metrics following standard protocol (Prajwal et al. 2020). The improvement of lip sync score (LSE-D↓: 7.127 → 7.013; FID↓: 3.452 → 3.264) indicates that PrefAce learns rich and transferable spatio-temporal patterns for landmarks.

Few-Shot Adaptation Few shot adaptation has recently gained attention due to its importance in low data regimes. Following standard evaluation protocols, we also investigate the adaptation capability of PrefAce. We use limited train set labels while keeping the test set fixed via Linear Probing (MOSEI, CelebV-HQ) and Finetuning (FF+) strategy. From Tab. 5, it is observed that PrefAce outperforms SOTA MARLIN across different tasks which further demonstrates that PrefAce learns generic, transferable, and adaptive information.

Ablation Studies

We have performed extensive ablation studies to show the effectiveness of each component.

Different modules We progressively integrate each module into the framework and observe the effectiveness on downstream performance on CMU-MOSEI and FF++ while keeping other components fixed. From Tab. 6, we can see that both the multiscale landmark guided self-distillation and the FaceFeat Cache are helpful in improving the model performance, demonstrating that the overall design of PrefAce is effective.

Encoder architectures To investigate the impact of the backbone encoder architectures, we compare ViT-S, ViT-B (See Tab. 6). It is observed that enlarging model size yields performance gain. For fair comparison, ViT-B encoder is employed.

Datasets→	MOSEI			FF++	
	Emo. Acc. (%↑)	7-Sent. Acc. (%↑)	2-Sent. Acc. (%↑)	Acc. (%↑)	AUC. (↑)
Modules↓					
Vanilla	80.25	33.82	73.00	88.39	0.9076
+ Landmark Distill	80.61	34.70	73.48	89.82	0.9144
+ FaceFeat Cache	80.67	34.64	73.32	88.67	0.9100
+ Both (PrefAce)	**81.29**	**35.03**	**74.05**	**92.30**	**0.9371**
Network Arch.↓					
ViT-S	80.69	34.28	72.90	89.53	0.8912
ViT-B	81.29	35.03	74.05	92.30	0.9371

Table 6: Contribution of different modules and network architectures towards overall PrefAce framework.

Qualitative Aspects

In order to understand the effectiveness of the learned features, we further conducted following qualitative analysis.

Representation Distribution To demonstrate that PrefAce can learn more robust and informative representations, we visualize the generated embedding space of PrefAce and MARLIN, on the test set of the YTF dataset using t-SNE. In Fig. 2, each t-SNE point is represented by the first frame of its corresponding video. It can be observed that embeddings generated by PrefAce are more evenly distributed, with almost no overlap between different videos. It's important to highlight that the majority of the few overlaps occur among videos of the same individual, which is reasonable. Moreover, embeddings generated by PrefAce are naturally grouped by attributes, e.g., gender, expression, head pose, hair color, etc. The aforementioned findings indicates that the embeddings generated by PrefAce focus more on characteristics of different faces, thus are more informative.

Conclusion

We propose a new self-supervised facial pretraining framework, PrefAce, that can generate universal, robust, and informative representations for downstream facial analysis tasks through multi-scale landmark guided self-distillation. To tackle the problem that previous works ignoring the diverse and discriminative features, we propose FaceFeat Cache, encouraging the network to explore characteristics of different faces, thus further enhancing downstream task performance.

Acknowledgements

This work is supported by A*STAR AME Programmatic Funding A18A2b0046, RobotHTPO Seed Fund under Project C211518008 and EDB Space Technology Development Grant under Project S22-19016-STDP.

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *WIFS*, 1–7. ISSN: 2157-4774.

Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; and Ballas, N. 2023. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *arXiv preprint arXiv:2301.08243*.

Bagher Zadeh, A.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *ACL*, 2236–2246. Melbourne, Australia.

Bao, H.; Dong, L.; and Wei, F. 2021. BEiT: BERT Pre-Training of Image Transformers. *arXiv preprint arXiv:2106.08254*.

Bayar, B.; and Stamm, M. C. 2016. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer. In *MMSec*, 5–10.

Bojanowski, P.; and Joulin, A. 2017. Unsupervised learning by predicting Noise. In *ICML*.

Bulat, A.; Cheng, S.; Yang, J.; Garbett, A.; Sanchez, E.; and Tzimiropoulos, G. 2022. Pre-training strategies and datasets for facial representation learning. ArXiv:2103.16554 [cs].

Cai, Z.; Ghosh, S.; Stefanov, K.; Dhall, A.; Cai, J.; Rezatofighi, H.; Haffari, R.; and Hayat, M. 2023. Marlin: Masked autoencoder for facial video representation learning. In *CVPR*, 1493–1504.

Cai, Z.; Stefanov, K.; Dhall, A.; and Hayat, M. 2022. Do You Really Mean That? Content Driven Audio-Visual Deepfake Dataset and Multimodal Method for Temporal Forgery Localization. In *DICTA*, 1–10. Sydney, Australia.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021a. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV*, 9650–9660.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021b. Emerging Properties in Self-Supervised Vision Transformers. arXiv:2104.14294.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 6299–6308.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. *preprint arXiv:2002.05709*.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. E. 2020b. Big Self-Supervised Models are Strong Semi-Supervised Learners. In *NeurIPS*, volume 33, 22243–22255. Curran Associates, Inc.

Chollet, F. 2017. Xception: Deep Learning With Depthwise Separable Convolutions. In *CVPR*, 1251–1258.

Chung, J.; and Zisserman, A. 2017. Lip reading in profile. *BMVC*.

Cozzolino, D.; Poggi, G.; and Verdoliva, L. 2017. Recasting Residual-based Local Descriptors as Convolutional Neural Networks: an Application to Image Forgery Detection. In *MMSec*, 159–164. ISBN 978-1-4503-5061-7.

Delbrouck, J.-B.; Tits, N.; Brousmiche, M.; and Dupont, S. 2020. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, 1–7. ArXiv: 2006.15955.

Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; and Feichtenhofer, C. 2021. Multiscale Vision Transformers. In *ICCV*, 6824–6835.

Fridrich, J.; and Kodovsky, J. 2012. Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security*, 7(3): 868–882. Conference Name: IEEE Transactions on Information Forensics and Security.

Gidaris, S.; Bursuc, A.; Komodakis, N.; Pérez, P.; and Cord, M. 2020. Learning representations by predicting bags of visual words. In *CVPR*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; Piot, B.; Kavukcuoglu, K.; Munos, R.; and Valko, M. 2020. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *ECCV*, Lecture Notes in Computer Science, 87–102. Cham: Springer International Publishing. ISBN 978-3-319-46487-9.

Hao, Z.; Mallya, A.; Belongie, S.; and Liu, M.-Y. 2021. GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. *arXiv:2104.07659 [cs]*. ArXiv: 2104.07659.

Haugh, M. 2009. Face and Interaction. *Face, Communication and Social Interaction*, 1–30.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2021. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*, 16000–16009.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*, volume 30. Curran Associates, Inc.

Jack, R. E.; and Schyns, P. G. 2015. The Human Face as a Dynamic Tool for Social Communication. *Current Biology*, 25(14): R621–R634.

Kadam, A.; Rane, S.; Mishra, A. K.; Sahu, S. K.; Singh, S.; and Pathak, S. K. 2021. A Survey of Audio Synthesis and

Lip-syncing for Synthetic Video Generation. *EAI Endorsed Transactions on Creative Technologies*, 8(28): e2–e2.

Koromilas, P.; and Giannakopoulos, T. 2022. Unsupervised Multimodal Language Representations using Convolutional Autoencoders. ArXiv:2110.03007 [cs].

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020. Face X-Ray for More General Face Forgery Detection. In *CVPR*, 5001–5010.

Li, S.; Chen, D.; Liu, B.; Yu, N.; and Zhao, R. 2019. Memory-based neighbourhood embedding for visual recognition. In *ICCV*, 6102–6111.

Li, S.; and Deng, W. 2022. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, 13(3): 1195–1215. Conference Name: IEEE Transactions on Affective Computing.

Li, Y.; Wu, C.-Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. In *CVPR*, 4804–4814.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *ICCV*, 3730–3738.

Noroozi, M.; and Favaro, P. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *ECCV*, Lecture Notes in Computer Science, 69–84. Cham: Springer International Publishing. ISBN 978-3-319-46466-4.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193.

Pan, T.; Song, Y.; Yang, T.; Jiang, W.; and Liu, W. 2021. Videomoco: Contrastive video representation learning with temporally adversarial examples. 11205–11214.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, volume 32. Curran Associates, Inc.

Prajwal, K. R.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 484–492. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-7988-5.

Qiu, Z.; Yao, T.; and Mei, T. 2017. Learning Spatio-Temporal Representation With Pseudo-3D Residual Networks. In *ICCV*, 5533–5541.

Rahmouni, N.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2017. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, 1–6. ISSN: 2157-4774.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Niessner, M. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *ICCV*, 1–11.

Si, S.; Wang, J.; Qu, X.; Cheng, N.; Wei, W.; Zhu, X.; and Xiao, J. 2021. Speech2Video: Cross-Modal Distillation for Speech to Video Generation. In *INTERSPEECH*, 1629–1633.

Skorokhodov, I.; Tulyakov, S.; and Elhoseiny, M. 2022. StyleGAN-V: A Continuous Video Generator With the Price, Image Quality and Perks of StyleGAN2. In *CVPR*, 3626–3636.

Storrs, K. R.; Anderson, B. L.; and Fleming, R. W. 2021. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, 5(10): 1402–1417. Number: 10 Publisher: Nature Publishing Group.

Tian, Y.; Che, Z.; Bao, W.; Zhai, G.; and Gao, Z. 2020. *Self-supervised Motion Representation via Scattering Local Motion Cues*, volume 12359 LNCS. Springer International Publishing.

Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; and Ortega-Garcia, J. 2020. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, 64: 131–148.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. ArXiv:2203.12602 [cs] type: article.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *CVPR*, 6450–6459.

Wang, G.; Zhang, P.; Xie, L.; Huang, W.; and Zha, Y. 2022. Attention-Based Lip Audio-Visual Synthesis for Talking Face Generation in the Wild. ArXiv:2203.03984 [cs, eess].

Wang, J.; Liu, Y.; Hu, Y.; Shi, H.; and Mei, T. 2021. FaceX-Zoo: A PyTorch Toolbox for Face Recognition. In *ACM MM*, MM '21, 3779–3782. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8651-7.

Wolf, L.; Hassner, T.; and Maoz, I. 2011. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 529–534. ISSN: 1063-6919.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. *arXiv preprint arXiv:2103.03230*.

Zheng, Y.; Yang, H.; Zhang, T.; Bao, J.; Chen, D.; Huang, Y.; Yuan, L.; Chen, D.; Zeng, M.; and Wen, F. 2022. General Facial Representation Learning in a Visual-Linguistic Manner. In *CVPR*, 18697–18709.

Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; and Loy, C. C. 2022. CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. ArXiv:2207.12393 [cs].