MARLIN: Masked Autoencoder for facial video Representation LearnINg

Zhixi Cai[1], Shreya Ghosh[1,2], Kalin Stefanov[1], Abhinav Dhall[1,3], Jianfei Cai[1],
Hamid Rezatofighi[1], Reza Haffari[1], Munawar Hayat[1]
[1]Monash University, [2] Curtin University, [3] Indian Institute of Technology Ropar
{zhixi.cai,kalin.stefanov,jianfei.cai,hamid.rezatofighi,gholamreza.haffari,
munawar.hayat}@monash.edu, shreya.ghosh@curtin.edu.au, abhinav@iitrpr.ac.in

Abstract

This paper proposes a self-supervised approach to learn universal facial representations from videos, that can transfer across a variety of facial analysis tasks such as Facial Attribute Recognition (FAR), Facial Expression Recognition (FER), DeepFake Detection (DFD), and Lip Synchronization (LS). Our proposed framework, named ***MARLIN****, is a facial video masked autoencoder, that learns highly robust and generic facial embeddings from abundantly available non-annotated web crawled facial videos. As a challenging auxiliary task, MARLIN reconstructs the spatio-temporal details of the face from the densely masked facial regions which mainly include eyes, nose, mouth, lips, and skin to capture local and global aspects that in turn help in encoding generic and transferable features. Through a variety of experiments on diverse downstream tasks, we demonstrate MARLIN to be an excellent facial video encoder as well as feature extractor, that performs consistently well across a variety of downstream tasks including FAR (1.13% gain over supervised benchmark), FER (2.64% gain over unsupervised benchmark), DFD (1.86% gain over unsupervised benchmark), LS (29.36% gain for Frechet Inception Distance), and even in low data regime. Our code and models are available at https://github.com/ControlNet/MARLIN.*

1. Introduction

Facial analysis tasks [34, 43, 70, 85] provide essential cues for human non-verbal behavior analysis, and help unfold meaningful insights regarding social interaction [36], communication [40], cognition [68] with potential applications in Human-Computer Interaction (HCI) and Affective Computing domains. Recently, we have witnessed significant progress in deep neural network models to solve facial analysis tasks such as Facial Attribute Recognition (FAR) [34, 85], Facial Expression Recognition (FER) [48], DeepFake Detection (DFD) [70], and Lip Synchronization (LS) [43]. While these deep models can achieve remarkable performance, they often require large-scale annotated datasets, which is not only a resource-expensive and time-consuming process but also infeasible for some applications requiring domain expertise for annotation (e.g. FER).

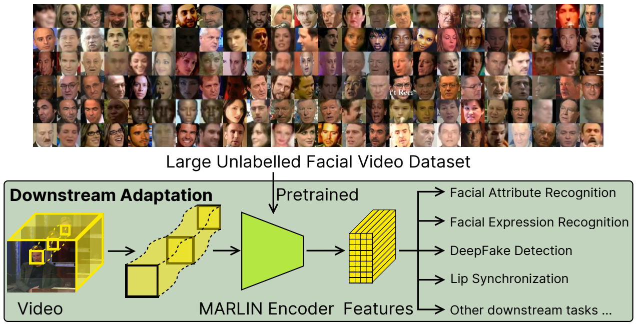
Large Unlabelled Facial Video Dataset
Downstream Adaptation
Pretrained
Video
MARLIN Encoder
Features
Facial Attribute Recognition
Facial Expression Recognition
DeepFake Detection
Lip Synchronization
Other downstream tasks ...

Figure 1. Overview of the proposed Masked Autoencoder for facial Representation LearnINg aka MARLIN. MARLIN aims to learn a universal facial representation from abundantly available non-annotated facial video data.

To this end, self-supervised pre-training [26, 37, 71] has lately emerged as an effective strategy to address the limitations of fully supervised methods, as it enables generic representation learning from non-annotated data, that can then be transferred across tasks having limited labels. For images of natural scenes and objects, self-supervised learning approaches using self-distillation [14], contrastive-learning [18, 19], solving pre-text tasks such as jigsaw puzzle [53], and more recently autoencoding [37, 71] have even outperformed the supervised learning approaches.

Despite the promises offered by these self-supervised methods in learning scalable and generic representations for natural scene images and videos, these have not yet been investigated for learning representations from facial video data. Facial representation learning requires tracking of fine-grained face specific details which might not be perfectly captured by linear tube masking [71]. Until now, most of the existing approaches associated with facial analysis tasks are highly specialized and develop task-specific models trained in a fully supervised manner [46, 54, 63], with very few recent efforts towards learning generic *image-based* facial encoding [10,84]. These closely related works [10, 84] either focus on exploring training dataset properties in terms of size and quality [10] or performing pre-training in visual-linguistic way [84]. These works [10,84] are hard to scale since they use static image-level facial information and the image-caption pairs are highly associated with context information rather than face.

In this paper, our goal is to learn *universal* and *task-agnostic* representations in a self-supervised manner for face-related downstream tasks (see Fig. 1). For this purpose, we employ a masked autoencoder [37, 71] with a facial-guided masking strategy that learns to reconstruct spatio-temporal details of a face from densely masked facial regions using non-annotated videos. Unlike existing approaches for natural scene videos [71], where the tube-masking is initialized with a static part of the video without any semantic information, our approach dynamically tracks face and then develops a facial part-guided tube masking strategy using an off-the-shelf face parser i.e. FaceX-Zoo [75]. Thus, we pose a more challenging task that encourages the model to learn spatio-temporal representations to cover local as well as global information. Inspired by prior works [27, 60] showing high-quality reconstruction results along with rich and generic latent features, we incorporate adversarial loss on top of masked encoding to enhance reconstruction quality. Our experimental results show that our proposed framework, MARLIN, 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, MARLIN, a *universal* and *task-agnostic* facial encoder that learns robust and transferable facial representation from abundantly available non-annotated web-crawled facial videos in a self-supervised fashion.
- As a challenging auxiliary task, we propose to reconstruct the spatio-temporal details of the face from the densely masked facial regions. The proposed facial region-guided tube masking (aka *Fasking*) strategy aims to learn local and global aspects from facial videos which in turn help encode generic and transferable features.
- Through extensive quantitative and qualitative analysis, we show that MARLIN learns rich, generic, transferable, and robust facial representation, that performs consistently well across a variety of downstream tasks including FAR (1.13% gain over supervised benchmark), FER (2.64% gain over unsupervised benchmark), DFD (1.86% gain over unsupervised benchmark), LS (29.36% gain for Frechet Inception Distance) and even in few shot settings.

Table 1. **Facial Analysis Tasks.** Overview of different face related tasks and relevant datasets down the lane.

Datasets	# Samples	Env.	Fmt.	Task	Year
LFW [39]	13,233	Wild	Img.	Identification	2008
VGG-FACE [54]	2.6M	Wild	Img.	Identification	2015
CelebA [50]	202,599	Wild	Img.	Attributes	2015
YouTubeFace [78]	3,425	Wild	Vid	Identification	2011
LRS2 [22]	144,482	Wild	Vid	Lip Sync.	2017
CelebV [79]	5	Wild	Vid	Reenact	2018
CMU-MOSEI [83]	23,453	Wild	Vid	Emo, Senti	2018
FaceForensics++ [62]	1,004	Wild	Vid	DeepFake	2019
VoxCeleb2 [23]	150,480	Wild	Vid	Speaker	2018
CelebV-HQ [85]	55,666	Wild	Vid	Attribute	2022

2. Related Work

Masked AutoEncoder. Masked autoencoder learns robust and transferable representations based on the hypothesis of reconstruction of the masked region. Masked autoencoding is motivated by context encoders [56] and denoising encoders [73]. After success of BERT [26] based masking, the vision community has also explored different design choices of masked auto encoding such as pixel level masking [17, 37, 80], token level masking [29] and deep feature based masking [6, 77], using vision Transformers [44, 52]. Similarly, for modeling spatio-temporal patterns of the input data, masked motion modelling [69] and tube masking [71] strategies have been incorporated recently. Along this line, MARLIN masks and reconstructs domain-specific facial parts to learn universal facial representation.

Facial Representation Learning. Till date, most of the existing facial analysis approaches are conducted in a task-specific way with fully supervised manner [46, 54, 63] 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. Tab. 1 shows an overview of the task-specific large-scale facial image or video datasets that have been curated over the past decade [1] to facilitate research in Face Verification (LFW [39], MS-celeb-1M [34], VGG-FACE [54], VGGFace2 [13]), Facial Attribute Recognition(CelebA [50], CelebV-HQ [85]), Facial Emotion Recognition (CMU-MOSEI [83]), DeepFake Detection (FF++ [62]) and Lip Synchronization (LRS2 [22]). 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 [10], and most importantly time consuming and resource expensive annotation process. To eliminate these drawbacks, some of the existing approaches [20, 81, 82] adopt data augmentation strategy via image or video synthesis as the surge in face generation technology fueled by Generative Adversarial Network (GAN) [20, 67, 81, 82] and other generation techniques [16, 35] aids realistic face generation even with the 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 domain 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 [3, 9, 10, 65, 84, 84, 86, 86]. The most closely related existing works [10, 84] either focus on exploring training dataset properties in terms of size and quality [10] or performing pre-training in visual-linguistic way [84]. These works [10, 84] are hard to scale since they use static image level facial information and the image-caption pairs are highly correlated with context information rather than face. In this work, we aim to develop a generic, universal, and task-agnostic facial encoder that learns from web-crawled non-annotated data. Our experimental analysis shows that MARLIN can align the latent space manifold to any desired downstream task specific label space. Thus, MARLIN has the capability to act as a strong facial encoder or feature extractor in many low-resource real world applications.

3. MARLIN

Our objective is to learn robust and transferable universal facial representation from abundantly available non-annotated facial video data [78]. If we think holistically, face specific tasks involve two different aspects: a) facial appearance related attributes such as parts of the face (nose, eyes, lips, hair, etc.), facial shape and texture which mainly need spatial investigation; and b) facial action such as emotion, Facial Action Coding System (FACS), lip synchronization which requires temporal information. Thus, spatio-temporal modeling is highly desirable in order to learn strong, robust, and transferable representation. To this end, our proposed framework, MARLIN, adopts a facial region guided masking strategy which poses a challenging auxiliary reconstruction task for self supervised representation learning (See Fig. 2). To facilitate learning from masked auto-encoder, we mainly choose the YouTube Faces [78] dataset that uses web-crawled facial videos from YouTube having variation in terms of different real life conditions.

3.1. Facial Representation Learning

Preliminaries. MARLIN consists of an encoder ($\mathcal{F}_{\phi_\mathcal{E}}$), decoder ($\mathcal{F}_{\phi_\mathcal{D}}$) and discriminator ($\mathcal{F}_{\phi_\Gamma}$) with embedding parameters $\phi_\mathcal{E}$, $\phi_\mathcal{D}$ and ϕ_Γ, respectively. Given a training dataset $\mathbb{D} = \{V_i\}_{i=1}^{N}$ where N is the number of videos in the dataset and $V \in \mathbb{R}^{C\times T_0\times H_0\times W_0}$ (C, T_0, H_0, W_0 are channel, temporal depth, height and width of the raw video, respectively). From the raw input video V, we track and crop the facial regions [75] followed by random temporal sampling represented as $v \in \mathbb{R}^{(C\times T\times H\times W)}$ (T, H, W are the modified temporal depth, height, and width of the derived video clip, respectively). The derived video clip v is further mapped to $(\mathbf{k}-\mathbf{n})$ visible and $\mathbf{n}$ masked tokens denoted as $\{\tilde{X}_v \in \mathbb{R}^{(\mathbf{k-n})\times\mathbf{e}}, \tilde{X}_m \in \mathbb{R}^{\mathbf{n}\times\mathbf{e}}\}$ by facial-region guided masking strategy ($\mathcal{F}_{\phi_f}$) with a pre-defined masking ratio $\mathbf{r} = \frac{\mathbf{n}}{\mathbf{k}}$. Here, $\mathbf{e}$ is the embedding dimension and $\mathbf{k}$ is the total number of tokens derived from v, i.e. $\mathbf{k} = \frac{T}{\mathbf{t}} \times \frac{H}{\mathbf{h}} \times \frac{W}{\mathbf{w}}$, given 3D cube tokens have dimension of $\mathbf{t} \times \mathbf{h} \times \mathbf{w}$ each. Thus, MARLIN injects facial region specific domain knowledge in the aforementioned token space to guide the representation learning via masking.

Algorithm 1 Facial-region Guided Masking Procedure

```
Require: $v \in \mathbb{R}^{(C\times T\times H\times W)}$, $\mathbf{r}$
seg_map ← FaceXZoo($v$)                                ▷ Face-Parsing,
seg_map∈{background, skin, left-eye,
right-eye, nose, mouth, hair}
$\mathbb{P}$ ={left-eye, right-eye, nose, mouth,
hair}                                                  ▷ Prioritize Regions
$\mathbf{k} = \frac{T}{\mathbf{t}} \times \frac{H}{\mathbf{h}} \times \frac{W}{\mathbf{w}}$          ▷ # of tokens for each $v$ (3D cube
tokens have dimension of $\mathbf{t} \times \mathbf{h} \times \mathbf{w}$ each)
$\mathbf{n} \leftarrow \mathbf{r} \times \mathbf{k}$                                  ▷ Number of masked tokens
$\tilde{X}_v \leftarrow \{\}$                                  ▷ Initialize visible tokens
patches ={background, skin, *shuffle($\mathbb{P}$)}
▷ Ordered list
for patch in patches do
    $\tilde{X}_v \leftarrow$ {patch}
    if len($\tilde{X}_v$) $==$ $(\mathbf{k} - \mathbf{r})$ then
        break
    end if
end for
$\tilde{X}_m \leftarrow \tilde{X} - \tilde{X}_v$                          ▷ $\tilde{X}$ is all tokens from $v$
```

The visible tokens $\tilde{X}_v$ are mapped to the latent space $\mathbf{z}$ by the following mapping function $\mathcal{F}_{\phi_\mathcal{E}} : \tilde{X}_v \to \mathbf{z}$. The latent space feature $\mathbf{z}$ is further fed to the decoder $\mathcal{F}_{\phi_\mathcal{D}}$ which reconstruct $\mathbf{z}$ to the $\mathbf{n}$ masked tokens $X_m^{'}$ by the following mapping $\mathcal{F}_{\phi_d} : \mathbf{z} \to X_m^{'}$. In the decoder, the corresponding visible and masked 3D cubes contain the flatten raw pixels denoted as $\mathbf{e} = \mathbf{C}\mathbf{thw}$. In brief given the visible tokens $\tilde{X}_v$, we reconstruct the masked tokens by the following function:

$$X_m^{'} = \mathcal{F}_{\phi_\mathcal{D}} \circ \mathcal{F}_{\phi_\mathcal{E}}(\tilde{X}_v) \qquad (1)$$

Reconstructing spatio-temporal facial patterns from raw pixels is quite challenging, we deploy a discriminator $\mathcal{F}_{\phi_\Gamma}$ with the adversarial training for better synthesis.

3.2. Self-Supervised Representation Learning.

The self supervised pre-training strategy of MARLIN consists of three main components described below:

a) Facial-region Guided Tube Masking (Fasking). In order to capture spatio-temporal correspondence, we have deployed facial region specific tube masking strategy following [71]. We dynamically track and mask facial compo-

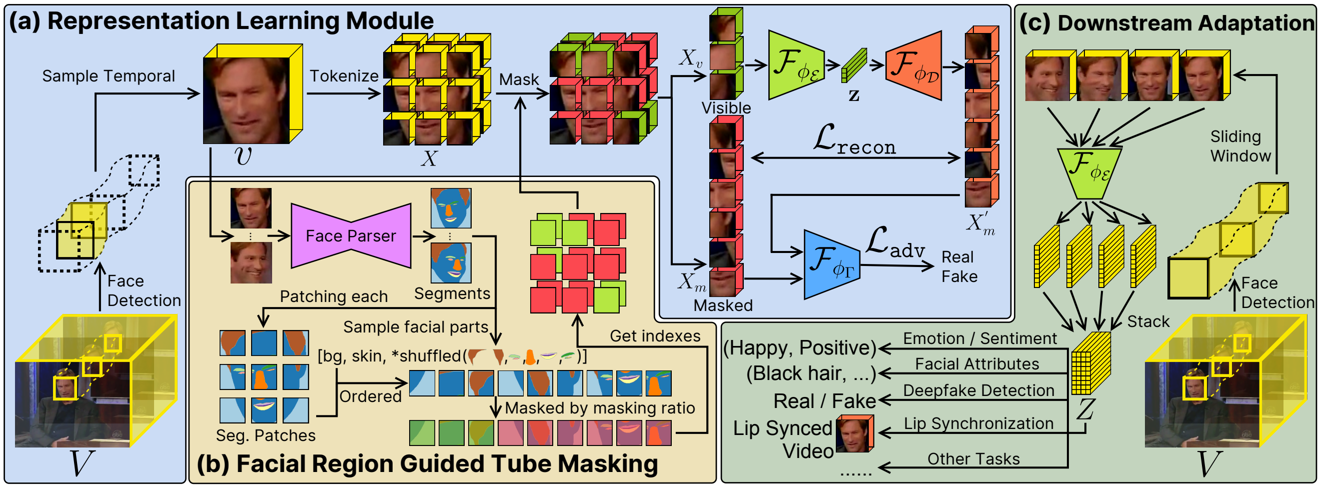

Figure 2. **Architectural overview of MARLIN (Best viewed in color).** MARLIN mainly consists of (a) Representation Learning Module, (b) Facial Region guided Tube Masking, and (c) Downstream Adaptation. (a) *Representation Learning Module:* MARLIN learns the facial representation from the unlabeled, web crawled video data in a self-supervised fashion (highlighted in Blue). (b) *Facial Region guided Tube Masking:* With the aid of facial region guided tube masking (highlighted in Yellow), MARLIN gets joint spatio-temporal attention which in turn facilitates downstream performance. The Face guided tube masking strategy injects domain knowledge into the pipeline. (c) *Downstream Adaptation:* For facial task specific downstream adaptation, MARLIN utilizes Linear Probing (LP) and Fine-Tuning (FT) to show the robustness, generalizability, and transferability of the learned feature (highlighted in Green).

nents across temporal axis for each spatio-temporal cube. Our facial regions based tube-masking strategy ensures the same facial region is masked throughout the temporal cube, thus posing a challenging reconstruction task and promoting learning local and global facial details (See Alg. 1). As the masked spatio-temporal cubes look like deformable bending tubes, we termed it as *Facial region-guided tube masking* aka *Fasking*.

We begin with face parsing using FaceXZoo [75] library which divides facial regions into the following parts {`left-eye`, `right-eye`, `nose`, `mouth`, `hair`, `skin`, `background`} (Fig. 2 (b)). Among the facial regions, we prioritize the following set $\mathbb{P}$ = {`left-eye`, `right-eye`, `nose`, `mouth`, `hair`} over skin and background to preserve face specific local and sparse features. In order to maintain pre-defined masking ratio $\mathbf{r}$, facial regions from the priority set $\mathbb{P}$ are masked across frames first followed by {`background`, `skin`} masking. Thus, Fasking generates $\mathbf{n}$ masked and $(\mathbf{k}-\mathbf{n})$ visible tokens. Across all the frames of the input v, we track specific facial regions from the pre-defined set to encode and reconstruct spatio-temporal changes to the model facial motion. The fasking strategy thus poses more challenges to the reconstruction while encoding subject specific appearance and fine-grained details.

b) Masked Autoencoder. After Fasking, $(\mathbf{k}-\mathbf{n})$ visible tokens are given input to the Encoder $\mathcal{F}_{\phi_\mathcal{E}}$ which maps the tokens to the latent space $\mathbf{z}$. The visible tokens serve as a reference to generate the masked counterpart of the face. Thus, the decoder $\mathcal{F}_{\phi_\mathcal{D}}$ maps the latent space $\mathbf{z}$ to the reconstructed masked tokens $X^{'}_{m}$. Please note that similar to VideoMAE [71], we adopt ViT [28] architecture as a backbone for MARLIN. A reconstruction loss ($\mathcal{L}_{\texttt{recon}}$) is imposed between masked cubes X_m and their reconstructed counterparts $X^{'}_{m}$ to guide the learning objective.

c) Adversarial Adaptation Strategy. To enhance the generation quality for rich representation learning, we incorporate adversarial adaptation on top of the masked autoencoder backbone. According to the prior literature [27,60], adversarial training enhances generation quality which in turn results in rich latent feature z. The discriminator $\mathcal{F}_{\phi_\Gamma}$ as shown in Fig. 2 is an MLP based network which imposes adversarial loss $\mathcal{L}_{\text{adv}}$ between X_m and their reconstructed counterparts $X^{'}_{m}$.

3.3. Overall MARLIN Loss

Alg. 2 summarizes the training process for the MARLIN framework. MARLIN mainly imposes (a) Reconstruction Loss and (b) Adversarial Loss to facilitate the training.

(a) Reconstruction Loss. Given an input masked tokens $\tilde{X}_m$, the masked auto-encoder module reconstruct it back to $X^{'}_{m}$. To this end, we minimize mean squared error loss in the 3D token space to update the weights of the $(\mathcal{F}_{\phi_\Gamma} \circ \mathcal{F}_{\phi_\mathcal{E}} \circ \mathcal{F}_{\phi_f})$ branch. The loss is defined as

$$\mathcal{L}_{\texttt{recon}} = \frac{1}{N}\sum_{i=1}^{N} ||X_m^{(i)} - X_m^{'(i)}||_2 \tag{2}$$

where N is the total number of data in $\mathbb{D}$, $X_m^{(i)}$ and $X_m^{'(i)}$ are the masked token and reconstruction of i-th data in $\mathbb{D}$.

Algorithm 2 Training procedure for MARLIN

```
Require: $\mathcal{F}_{\phi_f}, \mathcal{F}_{\phi_\mathcal{E}}, \mathcal{F}_{\phi_\mathcal{D}}, \mathcal{F}_{\phi_\Gamma}, \mathcal{F}_\theta, \mathbb{D}$, r, $\mathbb{D}_{down}$
while not converged do                                   ▷ MARLIN pre-training
    $v \leftarrow \texttt{sample batch}(\mathbb{D})$
    $\{\tilde{X}_m, \tilde{X}_v\} \leftarrow \mathcal{F}_{\phi_f}(v, r)$          ▷ Fasking (See Algo 1)
    $X'_m \leftarrow \mathcal{F}_{\phi_\mathcal{D}} \circ \mathcal{F}_{\phi_\mathcal{E}}(\tilde{X}_v)$          ▷ Train $\mathcal{F}_{\phi_\Gamma}$
    $\{\phi_\Gamma\} \leftarrow \nabla_{\{\phi_\Gamma\}} \mathcal{L}^{(d)}(X_m, X'_m)$
    $X'_m \leftarrow \mathcal{F}_{\phi_\mathcal{D}} \circ \mathcal{F}_{\phi_\mathcal{E}}(X_v)$          ▷ Train $\mathcal{F}_{\phi_\mathcal{E}}, \mathcal{F}_{\phi_\mathcal{D}}$
    $\{\phi_\mathcal{E}, \phi_\mathcal{D}\} \leftarrow \nabla_{\{\phi_\mathcal{E}, \phi_\mathcal{D}\}} \mathcal{L}^{(g)}(X_m, X'_m)$
end while
while not converged do                                   ▷ Downstream Adaptation
    $\{v, \mathbf{y}\} \leftarrow \texttt{sample batch}(\mathbb{D}_{\texttt{down}})$
    $\tilde{X} \leftarrow$ tokenize $v$
    $\mathbf{y}' \leftarrow \mathcal{F}_\theta \circ \mathcal{F}_{\phi_\mathcal{E}}(\tilde{X})$          ▷ Adapt downstream label
    if Linear Probing then                               ▷ Linear Probing
        $\{\theta\} \leftarrow \nabla_{\{\theta\}} \mathcal{L}_{\texttt{down}}(y, y')$
    else                                                 ▷ Fine-Tuning
        $\{\phi_\mathcal{E}, \theta\} \leftarrow \nabla_{\{\phi_\mathcal{E}, \theta\}} \mathcal{L}_{\texttt{down}}(y, y')$
    end if
end while
```

(b) Adversarial Loss. The adversarial adaptation considers the Wassenstain GAN loss [5] for better reconstruction of spatio-temporal facial patterns which in turn helps in learning rich representation. The loss is defined as follows:

$$\mathcal{L}^{(d)}_{\texttt{adv}} = \frac{1}{N\mathbf{n}} \sum_{i=1}^{N} \Big(\sum_{x'_m \in X'^{(i)}_m} \mathcal{F}_{\phi_\Gamma}(x'_m) - \sum_{x_m \in X^{(i)}_m} \mathcal{F}_{\phi_\Gamma}(x_m) \Big) \tag{3}$$

$$\mathcal{L}^{(g)}_{\texttt{adv}} = -\frac{1}{N\mathbf{n}} \sum_{i=1}^{N} \sum_{x'_m \in X'^{(i)}_m} \mathcal{F}_{\phi_\Gamma}(x'_m) \tag{4}$$

Thus, the overall learning objective $\mathcal{L}$ is formulated as follows, where λ_W is the weighting parameter:

$$\mathcal{L}^{(g)} = \mathcal{L}_{\texttt{recon}} + \lambda_W \mathcal{L}^{(g)}_{\texttt{adv}} \tag{5}$$

$$\mathcal{L}^{(d)} = \mathcal{L}^{(d)}_{\texttt{adv}} \tag{6}$$

During MARLIN's pre-training phase, $\mathcal{L}^{(d)}$ updates the parameters ϕ_{dis} and $\mathcal{L}^{(g)}$ updates the parameters ϕ_e, ϕ_d.

3.4. Downstream Adaptation

Our proposed MARLIN framework learns robust and transferable facial representation from the facial video in a self-supervised way. Following the standard evaluation protocols, we adopt Linear Probing (LP) and Fine-Tuning (FT) for downstream adaptation for different face relevant tasks (See Fig. 2 inference module). Given any task specific downstream dataset $\mathbb{D}_{\texttt{down}} = \{v_j, \mathbf{y}_j\}_{j=1}^{\mathcal{N}}$, we deploy linear fully-connected (FC) layers with embedding parameters θ to align the latent space to the downstream task specific label space on top of encoder module $\mathcal{F}_{\phi_\mathcal{E}}$. For linear probing, we freeze the backbone network $\mathcal{F}_{\phi_\mathcal{E}}$ and only update the $\mathcal{F}_\theta$. On the other hand for FT, we fine-tune the whole module i.e. $(\mathcal{F}_{\phi_\mathcal{E}} \circ \mathcal{F}_\theta)$. When MARLIN is used as a feature extractor for LP, it uses a sliding temporal window to extract features Z of the input face cropped video V as shown in Fig. 2 (c). The details of different downstream facial tasks are described below:

Facial Attribute Recognition (FAR) predicts the presence of appearance and action attributes such as gender, race, hair color, and emotion of a given face video. The problem of predicting facial attributes can be posed as a multi-label learning problem highly dependent on rich spatial encoding. For the downstream adaptation purpose, we use 28,532 train, 3,567 val, and 3,567 test videos from the CelebV-HQ [85] dataset. Following the prior works [33, 50, 84], 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 emotion (6-class) and sentiment (7-class and 2-class) of the concerned subject given a facial video. We evaluate the performance of MARLIN on CMU-MOSEI dataset [7] which is a conversational corpus having 16,726 train, 1,871 val, and 4,662 test data. Following the prior works [7, 25], we use overall accuracy(↑) as metrics.

Deepfake Detection (DFD) task predicts spatio-temporal facial forgery given a facial video from FF++(LQ) dataset [62]. For downstream adaptation, we use 3,600 train, 700 val, and 700 test sample videos from FF++(LQ) dataset [62]. Following prior literature [12, 58, 76], we use accuracy(↑) and AUC(↑) as the evaluation metrics.

Lip Synchronization (LS) is another line of research that require facial region specific spatio-temporal synchronization. This downstream adaptation further elaborates the adaptation capability of MARLIN for face generation tasks. For adaptation, we replace the facial encoder module in Wav2Lip [57] with MARLIN, and adjust the temporal window accordingly i.e. from 5 frames to **T** frames. For evaluation, we use the LRS2 [22] dataset having 45,838 train, 1,082 val, and 1,243 test videos. Following the prior literature [57, 74], we use Lip-Sync Error-Distance (LSE-D ↓), Lip-Sync Error-Confidence (LSE-C ↑) and Frechet Inception Distance (FID ↓) [38] as evaluation matrices.

4. Experiments and Results

We have comprehensively compared our method on different downstream adaptation tasks from quantitative (See Sec. 4.2) and qualitative (See Sec. 4.3 perspectives. Ad-

Table 2. **Facial Attribute Recognition.** Our proposed framework, MARLIN, trained on YTF [78] dataset and Linear Probed/Fine-Tuned on CelebV-HQ [85] benchmark dataset in terms of accuracy↑ and area under the curve↑. * shows supervised methods trained on the CelebV-HQ [85] dataset.

Method	Appearance		Action		Overall
	Acc.↑	AUC↑	Acc.↑	AUC↑	Acc.↑
R3D [72]*	92.34	0.9424	94.57	0.9173	93.45
MViTv1 [30]*	92.90	0.9452	95.13	0.9233	94.01
MViTv2 [49]*	92.77	0.954	95.15	0.9239	93.96
VideoMAE (FT) [71]	92.91	0.9529	95.37	0.9284	94.14
MARLIN (LP)	91.90	0.9373	95.25	0.9278	93.57
MARLIN (FT)	93.90	0.9561	95.48	0.9406	94.69

ditionally, we have performed extensive ablation studies to provide justification for our design choices.

4.1. Experimental Protocols

Datasets. We evaluate the MARLIN framework on different facial analysis tasks described in Sec. 3.4. In brief, we use CelebV-HQ [85] for facial attribute and action prediction, CMU-MOSEI dataset [7] for conversational emotion and sentiment prediction, FF++(LQ) dataset [62] for deepfake detection and LRS2 [22] for lip synchronization.

Settings. For fair comparisons, we follow the dataset specific experimental protocols mentioned in the task specific prior literature [7, 22, 33, 50, 62, 84]. Other than traditional evaluation, we perform few shot adaptation strategy as well to show the robustness and transferability of MARLIN.

Implementation Details. We implemented the method on PyTorch [55] with Nvidia RTX A6000 GPU. First of all, given any temporal chunk of a facial video, consecutive frames are highly redundant. Therefore, to consider semantically meaningful frames having significant motion across frames, we adopt the minimum temporal stride value to be 2. Given an input video (having dimension $3 \times 16 \times 224 \times 224$), the cube embedding layer generates $8 \times 14 \times 14$ 3D tokens of dimension $2 \times 16 \times 16$ to preserve spatio-temporal patterns. Using the Fasking strategy (See Algo. 1), MARLIN densely masks these tokens with a pre-defined masking ratio. Our empirical analysis suggests that MARLIN works favorably with a high masking ratio (90%). MARLIN's objective is to generate the masked part from the sparse visible tokens. After Fasking, each token is mapped to the latent space embedding dimension of 768. From this latent embedding, the masked part is reconstructed in the 3D token space that can further be mapped to the original video. For fair comparison, we use ViT-B as the backbone encoder, although the impact of other ViT-variants are depicted in ablation study. The pre-training hyperparameters are as follows: the base learning rate is linearly scaled with respect to the overall batch size, $\texttt{lr} = \texttt{base learning rate} \times \texttt{batch size}/256$. For self-supervised pre-training, we use AdamW optimizer with base learning rate $1.5e{-}4$, momentum $\beta_1 = 0.9, \beta_2 = 0.95$ with a learning rate scheduler (cosine decay) [51]. For linear probing, we use Adam optimizer with $\beta_1 = 0.5, \beta_2 = 0.9$ and base learning rate $1e{-}4$, weight decay 0. For fine-tuning, we use Adam optimizer with $\beta_1 = 0.5, \beta_2 = 0.9$ and base learning rate $1e{-}4$ without any weight decay.

Table 3. **Facial Expression and Sentiment Recognition.** Downstream adaptation results on MOSEI dataset [7] for Emotion, sentiment (7-class), and sentiment (2-class). Our proposed method, MARLIN, outperforms visual modality based emotion prediction methods. *Please note that SOTA for UMON [25] and GMF [4] utilize three modalities and thus, not directly comparable.* Here, YTF: YouTubeFace [78] and LAV represents linguistic, audio, and visual modality, respectively. * denotes supervised methods.

Tasks	Pre-train	Method	Mod.	Acc.↑
Emotion	–	MViTv1 [49]*	V	80.45
	–	UMONS [25]*	LAV	80.68
	–	GMF [4]*	LAV	81.14
	YTF [78]	VideoMAE [71]	V	80.39
	YTF [78]	MARLIN	V	80.60
Sentiment (7-Class)	–	MViTv1 [49]*	V	33.35
	YTF [78]	VideoMAE [71]	V	33.78
	YTF [78]	MARLIN	V	34.63
Sentiment (2-Class)	MOSEI [7] and IEMOCAP [11]	CAE-LR [45]	V	71.06
	YTF [78]	VideoMAE [71]	V	72.96
	YTF [78]	MARLIN	V	73.70

4.2. Quantitative Analysis

4.2.1. Comparison with SOTA Facial Analysis Tasks.

We compare the performance of MARLIN with different downstream facial analysis tasks following standard task specific evaluation protocols [7, 22, 33, 50, 62, 84].

Facial Attributes. In Tab. 2, we compare the LP and FT adaptation performance of MARLIN with the popular trans-

Table 4. **Deepfake Detection.** We compare the Fine-Tuning (FT) results on MARLIN for FaceForensic++ [62] dataset. * denotes supervised methods.

Pre-train	Method	Acc.(%)↑	AUC↑
–	Steg.Features [32]*	55.98	–
–	LD-CNN [24]*	58.69	–
–	Constraied Conv. [8]*	66.84	–
–	CustomPooling CNN [61]*	61.18	–
–	MesoNet [2]*	70.47	–
–	Face X-ray [47]*	–	0.6160
–	Xception [21]*	86.86	0.8930
–	F^3-Net [58]*	93.02	0.9580
–	P3D [59]*	–	0.6705
–	R3D [72]*	–	0.8772
–	I3D [15]*	–	0.9318
–	M2TR [76]*	–	0.9395
–	ST-M2TR [76]*	–	0.9531
YTF [78]	VideoMAE [71]	87.57	0.9082
YTF [78]	MARLIN	89.43	0.9305

Table 5. **Lip Synchronization.** We compare Linear Probing (LP) and Fine-Tuning (FT) results on the LRS2 [22] dataset.

Method	**LSE-D↓**	**LSE-C ↑**	**FID↓**
Speech2Vid [41]	14.230	1.587	12.320
LipGAN [42]	10.330	3.199	4.861
Wav2Lip [57]	7.521	6.406	4.887
AttnWav2Lip [74]	7.339	6.530	–
Wav2Lip + ViT [28]	8.996	2.807	13.352
Wav2Lip + ViT + VideoMAE [71]	7.316	5.096	4.097
Wav2Lip + ViT + MARLIN	7.127	5.528	3.452

former (i.e. MViT-v1 [30] and MViT-v2 [49]) and CNNs (i.e. R3D [72]) on CelebV-HQ [85] dataset. From the table, it is observed that MARLIN's FT version outperforms supervised MViT-v2 [49] transformer architecture by 1.13% (92.77% → 93.90%) for appearance attributes and 0.33% (95.15% → 95.48%) for action attributes. Similar patterns are also been observed with the R3D CNN module as well. We attribute MARLIN's performance gain to the pre-training strategy that encodes generic, robust, and transferable features from any input facial video.

Emotion and Sentiment. In Tab. 3, we similarly compare the LP and FT adaptation performance of conversational emotion and sentiment in terms of accuracy(↑) and AUC(↑) on CMU-MOSEI [83] dataset. *Please note that the MARLIN is a visual modality only encoder.* The results suggest that MARLIN performs competitively with SOTA methods [25, 45, 49], especially it outperforms unsupervised SOTA CAE-LR [45] by 2.64% (71.06% → 73.70%) on 2-class sentiment task. For emotion and 7-class sentiment as well, it outperforms supervised benchmarks [49] marginally. These results also indicate that MARLIN learns highly generic, robust, and transferable feature representation from pre-training.

DeepFake Detection. In Tab. 4, we compare the performance of video manipulation on FaceForensics++ [62] dataset and report results in terms of video-level accuracy(↑) and AUC(↑). The results indicate that MARLIN performs favorably against the supervised SOTA methods [2, 8, 15, 21, 24, 32, 47, 59, 61, 72]. This is the first SSL work that uses only spatio-temporal visual information anomaly to detect video manipulation. Unless F^3-Net, which uses frequency aware pattern over the temporal dimension to detect forgeries in a supervised fashion. Whereas MARLIN irrespective of frequency pattern learns facial representation and can detect anomalies from the spatio-temporal signal.

Lip Synchronization. For a fair comparison, we adopt the following experimental setups: *1) Wav2Lip+ViT:* To compare the contribution of ViT architecture [28] wrt SOTA CNNs and MARLIN where the weights of ViT is trained from scratch on LRS2 [22] dataset. *2) Wav2Lip+ViT+VideoMAE:* To compare the contribution of vanilla VideoMAE with ViT backbone pre-trained on YTF [78] dataset. *2) Wav2Lip+ViT+MARLIN:* To compare the contribution of MARLIN pre-trained on YTF [78] with SOTA [57, 66, 74] and different design aspects. The experimental results are depicted in Tab. 5 with LSE-D↓, LSE-C ↑ and FID ↓ as evaluation metrics following standard protocol [38, 57, 66, 74]. The improvement of lip sync score (LSE-D↓: 7.521 → 7.127; FID ↓: 4.887 → 3.452) indicates that MARLIN learns rich spatio-temporal patterns which are transferable and robust. It is also interesting to observe that MARLIN is adaptive to very fine grained features specific to the face as well.

Table 6. **Few shot adaptation on different facial tasks.** Comparison of different methods for few shot adaptation.

Data→	**MOSEI [7]**			**FF++ [62]**	**CelebV-HQ [85]**	
Task→	**Emo.**	**7-Sen.**	**2-Sen.**	**DeepFake**	**Appr.**	**Act.**
Anno.%	**Acc.↑**	**Acc.↑**	**Acc.↑**	**AUC↑**	**AUC↑**	**AUC↑**
100%	80.60	34.63	73.70	0.9305	0.9373	0.9278
50%	80.59	33.73	73.33	0.8681	0.9273	0.9270
10%	79.89	33.56	72.26	0.7459	0.8996	0.9201
1%	78.61	30.09	71.89	0.6252	0.8423	0.9063

4.2.2. Few-Shot Adaptation.

Few shot adaptation has recently gained attention due to its adaptation capability with very low data regime [9, 65, 84, 86]. Following the standard evaluation protocol [9, 65, 84, 86], we also investigate the adaptation capability of MARLIN. Given any downstream dataset, we use limited train set labels to align the output manifold while keeping the test set fixed via LP (MOSEI, CelebV-HQ) and FT (FF+) strategy. From Tab. 6, a slight drop in performance is observed across different tasks which further demonstrates that MARLIN learns generic, transferable, and adaptive information.

4.2.3. Ablation Studies.

We have performed extensive ablation studies to show the

Table 7. Contribution of different modules, encoder architectures, and masking strategies towards overall MARLIN framework. Fasking: Facial Guided Masking, AT: Adversarial Training

Datasets →	**MOSEI [7]**			**FF++ [62]**	
	Emo.	**7-Sent.**	**2-Sent.**		
	Acc.	Acc.	Acc.	**Acc.**	**AUC.**
	(%↑)	(%↑)	(%↑)	(%↑)	(↑)
Modules ↓					
VideoMAE	80.39	33.78	72.96	87.57	0.9082
+ Fasking	80.55	34.58	73.54	87.29	0.9154
+ AT	80.58	34.05	73.17	88.00	0.9096
+ Both (MARLIN)	**80.60**	**34.63**	**73.70**	**89.43**	**0.9305**
Encoder Arch. ↓					
ViT-S	80.38	33.40	72.69	87.43	0.8863
ViT-B	80.60	34.63	73.70	89.43	0.9305
ViT-L	**80.63**	**35.28**	**74.83**	**90.71**	**0.9377**
Masking Strategy ↓					
Random	80.40	34.10	72.96	87.29	0.8797
Frame	79.33	33.99	72.90	86.57	0.8835
Tube	80.58	34.05	73.17	88.00	0.9096
Fasking	**80.60**	**34.63**	**73.70**	**89.43**	**0.9305**

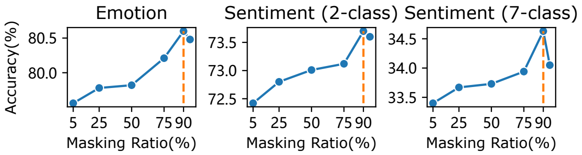

Figure 3. **Impact of Masking Ratio** Comparison of different masking ratios for emotion and sentiment prediction in CMU-MOSEI dataset [7]. Empirically, it suggests 90% masking works best for MARLIN.

effectiveness of each component.

1) Masking ratio. We use different masking ratios in the range [0.05 - 0.95] and repeat the pre-training followed by LP on CMU-MOSEI [83] dataset. From Fig. 3, we see that $\sim$ 90% masking ratio is optimal for the MARLIN. With a less masking ratio (i.e. ≤ 0.5), more information is available for the reconstruction task which degrades the feature quality. Similarly, beyond $\sim$ 90%, the reconstruction task becomes more challenging, leading to a performance drop. With the empirical evidence, we set the masking ratio to be $\sim$ 90% throughout all of our experiments. **2) Masking strategies.** We further compare the proposed *Fasking* strategy with existing masking strategies [31,71] *i.e. Frame*, *Random* and *Tube-Masking*. The empirical results in Tab. 7 demonstrate that *Fasking* is better. **3) Different modules.** We progressively integrate each module and observe its impact on downstream performance on CMU-MOSEI [83] and FF++ [62] while keeping other components fixed. From Tab. 7, we see that the addition of Fasking and Adversarial Training (AT) improves the performance, reflecting the importance of each component. **4) Encoder architectures.** To investigate the impact of the backbone encoder architectures, and compare ViT-S, ViT-B, and ViT-L (See Tab. 7). We observe that the larger model size enhances the performance. For fair comparison, we use a ViT-B encoder.

4.3. Qualitative Aspects

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

1) Facial Attributes. We visualize the important regions that MARLIN focused on using Gradient-weighted Class Activation Mapping (Grad-CAM) [64]. In Fig. 4 top, the heat-map results are based on LP on top of MARLIN's feature on CelebV-HQ [85] dataset (appearance task) and it indicates that MARLIN focus on facial attributes such as hair, spectacle, hat, etc. **2) Lip Synchronization.** In Fig. 4 bottom, we presents the generation results for lower part of faces which is a challenging task. The top, middle and bottom rows show ground truth, vanilla Wav2Lip [57]'s output and MARLIN's output along with the closeup looks, respectively. Here, Wav2lip's CNN encoder failed to locate the lip region (as shown in the Wav2lip row of Fig. 4 highlighted in red) whereas MARLIN despite pre-trained on fasking strategy is adaptive enough to generate more accurate spatio-temporal pattern for LS.

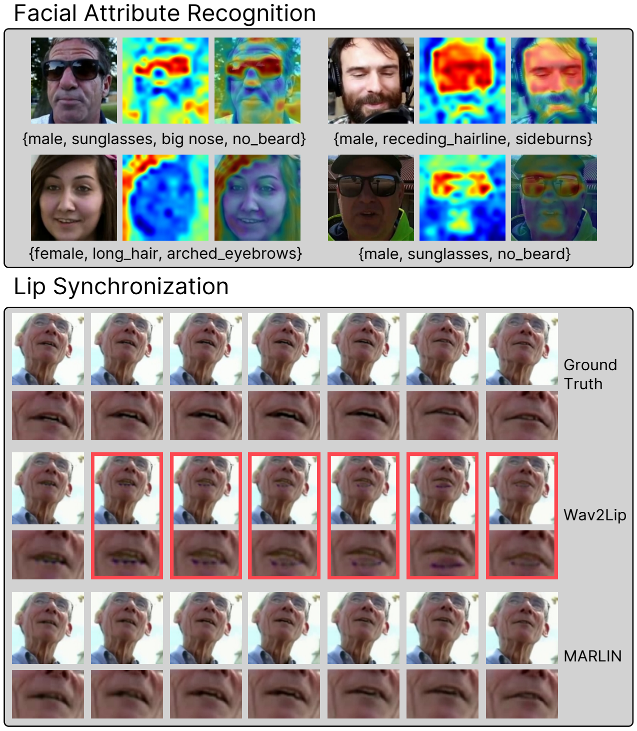
Facial Attribute Recognition
{male, sunglasses, big nose, no_beard}
{male, receding_hairline, sideburns}
{female, long_hair, arched_eyebrows}
{male, sunglasses, no_beard}
Lip Synchronization
Ground Truth
Wav2Lip
MARLIN

Figure 4. **Qualitative Analysis.** *Top:* Qualitative results for MARLIN for facial attribute recognition task. *Bottom:* Qualitative results for MARLIN for facial lip synchronization task.

5. Conclusion

In this paper, we aim to learn a universal and generic facial encoder, MARLIN, which is adaptive, robust and transferable for different facial analysis tasks. As a challenging auxiliary task, MARLIN reconstructs the spatio-temporal details of the face from the densely masked facial regions to capture local and global aspects which in turn helps in encoding generic and transferable features. **Broader Impact.** We believe that MARLIN can act as a good feature extractor for different downstream facial analysis tasks. Owing to the rich facial features, it would be easy to deploy MARLIN in low resource (e.g. mobile devices, Jetson Nano platforms) devices for real world applications. **Limitations.** As the model is trained on YouTube Face dataset [78], there could be potential bias in terms of race and cultural background of the identities. Potential bias can also be introduced in the model as we use the existing face detection library [75]. We will eliminate these limitations in our updated versions.

Acknowledgements.
This material is based on research sponsored by DARPA under agreement number HR001122C0029. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. M. Hayat is supported by the ARC DECRA Fellowship DE200101100.

References

[1] Insaf Adjabi, Abdeldjalil Ouahabi, Amir Benzaoui, and Abdelmalik Taleb-Ahmed. Past, Present, and Future of Face Recognition: A Review. *Electronics*, 9(8):1188, Aug. 2020. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute. 2

[2] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, Dec. 2018. ISSN: 2157-4774. 6, 7

[3] Shivangi Aneja and Matthias Nießner. Generalized Zero and Few-Shot Transfer for Facial Forgery Detection, June 2020. arXiv:2006.11863 [cs, eess]. 3

[4] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated multimodal networks. *Neural Computing and Applications*, 32(14):10209–10228, July 2020. 6

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223. PMLR, July 2017. ISSN: 2640-3498. 5

[6] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language, Oct. 2022. arXiv:2202.03555 [cs]. 2

[7] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. 5, 6, 7, 8

[8] Belhassen Bayar and Matthew C. Stamm. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, IH&amp;MMSec '16, pages 5–10, New York, NY, USA, June 2016. Association for Computing Machinery. 6, 7

[9] Bjorn Browatzki and Christian Wallraven. 3FabRec: Fast Few-Shot Face Alignment by Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6110–6120, 2020. 3, 7

[10] Adrian Bulat, Shiyang Cheng, Jing Yang, Andrew Garbett, Enrique Sanchez, and Georgios Tzimiropoulos. Pre-training strategies and datasets for facial representation learning, July 2022. arXiv:2103.16554 [cs]. 2, 3

[11] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, Dec. 2008. 6

[12] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do You Really Mean That? Content Driven Audio-Visual Deepfake Dataset and Multimodal Method for Temporal Forgery Localization. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–10, Sydney, Australia, Nov. 2022. 5

[13] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, May 2018. 2

[14] 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

[15] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6, 7

[16] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. Pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5799–5809, 2021. 2

[17] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative Pretraining From Pixels. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1691–1703. PMLR, Nov. 2020. ISSN: 2640-3498. 2

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, Nov. 2020. ISSN: 2640-3498. 1

[19] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc., 2020. 1

[20] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 2

[21] Francois Chollet. Xception: Deep Learning With Depthwise Separable Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017. 6, 7

[22] J. Chung and A. Zisserman. Lip reading in profile. *British Machine Vision Conference, 2017*, 2017. Publisher: British Machine Vision Association and Society for Pattern Recognition. 2, 5, 6, 7

[23] J. S. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*, 2018. 2

[24] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting Residual-based Local Descriptors as Convolutional Neural Networks: an Application to Image Forgery Detec-

tion. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, IH&amp;MMSec '17, pages 159–164, New York, NY, USA, June 2017. Association for Computing Machinery. 6, 7

[25] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 1–7, 2020. arXiv: 2006.15955. 5, 6, 7

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1, 2

[27] Jeff Donahue and Karen Simonyan. Large Scale Adversarial Representation Learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2, 4

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. 4, 7

[29] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are Large-scale Datasets Necessary for Self-Supervised Pre-training?, Dec. 2021. arXiv:2112.10740 [cs]. 2

[30] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 6, 7

[31] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked Autoencoders As Spatiotemporal Learners, Oct. 2022. arXiv:2205.09113 [cs]. 8

[32] Jessica Fridrich and Jan Kodovsky. Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, June 2012. Conference Name: IEEE Transactions on Information Forensics and Security. 6, 7

[33] Shreya Ghosh, Abhinav Dhall, and Nicu Sebe. Automatic Group Affect Analysis in Images via Visual Attribute and Feature Networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1967–1971, Oct. 2018. ISSN: 2381-8549. 5, 6

[34] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 87–102, Cham, 2016. Springer International Publishing. 1, 2

[35] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. *arXiv:2104.07659 [cs]*, Apr. 2021. arXiv: 2104.07659. 2

[36] Michael Haugh. Face and Interaction. *Face, Communication and Social Interaction*, pages 1–30, 2009. 1

[37] 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, 2

[38] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 5, 7

[39] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Oct. 2008. 2

[40] Rachael E. Jack and Philippe G. Schyns. The Human Face as a Dynamic Tool for Social Communication. *Current Biology*, 25(14):R621–R634, July 2015. 1

[41] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You Said That?: Synthesising Talking Faces from Audio. *International Journal of Computer Vision*, 127(11):1767–1779, Dec. 2019. 7

[42] Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. Towards Automatic Face-to-Face Translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 1428–1436, New York, NY, USA, Oct. 2019. Association for Computing Machinery. 7

[43] Anup Kadam, Sagar Rane, Arpit Kumar Mishra, Shailesh Kumar Sahu, Shubham Singh, and Shivam Kumar Pathak. A Survey of Audio Synthesis and Lip-syncing for Synthetic Video Generation. *EAI Endorsed Transactions on Creative Technologies*, 8(28):e2–e2, Apr. 2021. 1

[44] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in Vision: A Survey, Feb. 2021. arXiv: 2101.01169. 2

[45] Panagiotis Koromilas and Theodoros Giannakopoulos. Unsupervised Multimodal Language Representations using Convolutional Autoencoders, Jan. 2022. arXiv:2110.03007 [cs]. 6, 7

[46] Pavel Korshunov and Sebastien Marcel. DeepFakes: a New Threat to Face Recognition? Assessment and Detection, Dec. 2018. arXiv: 1812.08685. 2

[47] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-Ray for More General Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020. 6, 7

[48] Shan Li and Weihong Deng. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Comput-*

ing, 13(3):1195–1215, July 2022. Conference Name: IEEE Transactions on Affective Computing. 1

[49] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 6, 7

[50] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 2, 5, 6

[51] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*, July 2022. 6

[52] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing Properties of Vision Transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 23296–23308. Curran Associates, Inc., 2021. 2

[53] Mehdi Noroozi and Paolo Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 69–84, Cham, 2016. Springer International Publishing. 1

[54] O. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *BMVC 2015 - Proceedings of the British Machine Vision Conference 2015*, 2015. Publisher: British Machine Vision Association. 2

[55] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 6

[56] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context Encoders: Feature Learning by Inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 2

[57] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 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, pages 484–492, New York, NY, USA, Oct. 2020. Association for Computing Machinery. 5, 7, 8

[58] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 86–103, Cham, 2020. Springer International Publishing. 5, 6

[59] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning Spatio-Temporal Representation With Pseudo-3D Residual Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 6, 7

[60] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *4th International Conference on Learning Representations, ICLR 2016*, 2016. 2, 4

[61] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6, Dec. 2017. ISSN: 2157-4774. 6, 7

[62] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019. 2, 5, 6, 7, 8

[63] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 2

[64] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 8

[65] Ying Shu, Yan Yan, Si Chen, Jing-Hao Xue, Chunhua Shen, and Hanzi Wang. Learning Spatial-Semantic Relationship for Facial Attribute Recognition With Limited Labeled Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11916–11925, 2021. 3, 7

[66] Shijing Si, Jianzong Wang, Xiaoyang Qu, Ning Cheng, Wenqi Wei, Xinghua Zhu, and Jing Xiao. Speech2Video: Cross-Modal Distillation for Speech to Video Generation. In *INTERSPEECH 2021*, pages 1629–1633, Aug. 2021. 7

[67] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. StyleGAN-V: A Continuous Video Generator With the Price, Image Quality and Perks of StyleGAN2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3626–3636, 2022. 2

[68] Katherine R. Storrs, Barton L. Anderson, and Roland W. Fleming. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, 5(10):1402–1417, Oct. 2021. Number: 10 Publisher: Nature Publishing Group. 1

[69] Xinyu Sun, Peihao Chen, Liangwei Chen, Thomas H. Li, Mingkui Tan, and Chuang Gan. M^{3}Video: Masked Motion Modeling for Self-Supervised Video Representation Learning, Oct. 2022. arXiv:2210.06096 [cs]. 2

[70] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, Dec. 2020. 1

[71] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training, Mar. 2022. arXiv:2203.12602 [cs] type: article. 1, 2, 3, 4, 6, 7, 8

[72] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6, 7

[73] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *The Journal of Machine Learning Research*, 11:3371–3408, Dec. 2010. 2

[74] Ganglai Wang, Peng Zhang, Lei Xie, Wei Huang, and Yufei Zha. Attention-Based Lip Audio-Visual Synthesis for Talking Face Generation in the Wild, Mar. 2022. arXiv:2203.03984 [cs, eess]. 5, 7

[75] Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei. FaceX-Zoo: A PyTorch Toolbox for Face Recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pages 3779–3782, New York, NY, USA, Oct. 2021. Association for Computing Machinery. 2, 3, 4, 8

[76] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, ICMR '22, pages 615–623, New York, NY, USA, June 2022. Association for Computing Machinery. 5, 6

[77] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked Feature Prediction for Self-Supervised Visual Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 2

[78] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534, June 2011. ISSN: 1063-6919. 2, 3, 6, 7, 8

[79] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. ReenactGAN: Learning to Reenact Faces via Boundary Transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 603–619, 2018. 2

[80] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A Simple Framework for Masked Image Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 2

[81] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video Generation using VQ-VAE and Transformers, Sept. 2021. Number: arXiv:2104.10157 arXiv:2104.10157 [cs]. 2

[82] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating Videos with Dynamics-aware Implicit Generative Adversarial Networks. In *International Conference on Learning Representations*, Mar. 2022. 2

[83] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention Recurrent Network for Human Communication Comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. Section: Main Track: NLP and Machine Learning. 2, 7, 8

[84] 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, 3, 5, 6, 7

[85] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A Large-Scale Video Facial Attributes Dataset, July 2022. arXiv:2207.12393 [cs]. 1, 2, 5, 6, 7, 8

[86] Nan Zhuang and Cheng Yang. Few-Shot Knowledge Transfer for Fine-Grained Cartoon Face Generation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, July 2021. ISSN: 1945-788X. 3, 7