Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection

Alexandros Haliassos[1,†] Rodrigo Mira[1] Stavros Petridis[1,2] Maja Pantic[1,2]

[1]Imperial College London [2]Meta AI

{alexandros.haliassos14,rs2517,stavros.petridis04,m.pantic}@imperial.ac.uk

Abstract

One of the most pressing challenges for the detection of face-manipulated videos is generalising to forgery methods not seen during training while remaining effective under common corruptions such as compression. In this paper, we examine whether we can tackle this issue by harnessing videos of real talking faces, which contain rich information on natural facial appearance and behaviour and are readily available in large quantities online. Our method, termed RealForensics, consists of two stages. First, we exploit the natural correspondence between the visual and auditory modalities in real videos to learn, in a self-supervised cross-modal manner, temporally dense video representations that capture factors such as facial movements, expression, and identity. Second, we use these learned representations as targets to be predicted by our forgery detector along with the usual binary forgery classification task; this encourages it to base its real/fake decision on said factors. We show that our method achieves state-of-the-art performance on cross-manipulation generalisation and robustness experiments, and examine the factors that contribute to its performance. Our results suggest that leveraging natural and unlabelled videos is a promising direction for the development of more robust face forgery detectors.

1. Introduction

Automatic face manipulation methods can realistically change someone's appearance or expression without requiring substantial human expertise or effort [37, 62, 67, 72, 94]. This technology's potential social harm has spurred considerable research efforts to detect forgery content [3, 24, 35, 44, 49, 50, 53, 63, 68, 81, 92, 112, 116, 118].

It is known that although deep learning-based detectors can achieve high accuracy on in-distribution data, performance often plummets on videos generated using novel manipulation methods (*i.e.*, not seen during training) [19, 34, 53, 68, 72, 105, 118].

†Corresponding author.

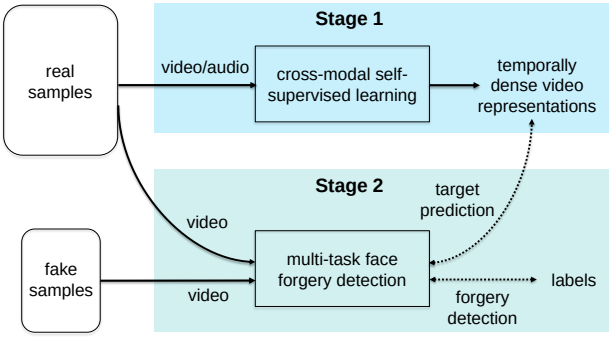

Figure 1. **Overview of our two-stage method**. First, we learn temporally dense video representations in a self-supervised way, by exploiting the correspondence between the visual and auditory modalities of real videos. Second, the network is presented with real and fake data and is tasked with performing face forgery detection while simultaneously predicting, for the real videos, the representations learned in stage 1. We use many more real than fake samples, as the former are more easily acquired.

Various frame-based methods (*i.e.*, that take a single frame as input) have been proposed to tackle cross-manipulation generalisation, including using data augmentation [105], truncating classifiers [19], using 3D decomposition [118], amplifying multi-band frequencies [79], and targeting the blending boundary between the background and the altered face [68]. Nevertheless, many still significantly underperform on novel forgery types or focus on low-level cues which can easily be corrupted by common perturbations like compression [53].

It is reasonable to believe that incorporating the temporal dimension can improve performance, especially since many synthesis methods do not take into account temporal consistency during the generation process [94]. However, as with frame-based methods, naively training deep networks on videos can lead to overfitting to the seen forgeries [53, 97, 114]. To counteract this, LipForensics [53] pre-trains on a large-scale lipreading dataset and then freezes part of the network to prevent it from focusing on low-level cues. It achieves strong performance in cross-

Figure 2. Top: consecutive frames of a fake video [37]. Bottom: same frames but heavily compressed. High-level semantics remain largely undisturbed under compression.

manipulation generalisation and robustness to common corruptions. On the other hand, (1) it requires pre-training on a *labelled* dataset, limiting its scalability; (2) it focuses exclusively on the mouth region; and (3) it freezes almost one third of the network when training on forgery data, which could sacrifice performance. A very recent method, FTCN [114], demonstrates high cross-manipulation generalisation by constraining all spatial convolutional kernel sizes to one. But, as we show, the impressive generalisation may come at a cost of reduced robustness to compression changes.

In this work, we are motivated by the observation that fake videos often exhibit anomalous facial movements (including mouth, eyes, and brows) and expressions, as well as subtle changes in facial form over time. Such cues are high-level in nature and thus more resilient to corruptions which destroy low-level content, *e.g.*, compression or blurring (see Figure 2). We ask ourselves whether it is possible to guide a detector to focus on such cues by utilising *unlabelled* real videos, which are relatively easy to obtain with tools like face and voice activity detectors.

To this end, we propose a two-stage approach, termed *RealForensics* (see Figure 1). We first use self-supervision to exploit the known correspondence between the visual and auditory modalities in *natural videos*. Inspired by the state-of-the-art method in image representation learning BYOL [48], we use a cross-modal student-teacher framework, where a student processing the video stream must predict representations formed by a slowly-improving teacher from the audio stream, and vice versa. We learn *temporally dense* representations (one embedding per frame), since cues related to facial movements are often fast-varying. Our goal is to capture *all* shared information between the two modalities, including factors associated with lexical content [29], emotion [96], and identity [83]. Hence, we *directly predict* the teachers' outputs. In the second stage, the forgery detector is tasked with performing classification while simultaneously predicting video targets generated by the video student from the first stage. This prediction task incentivises the detector to focus on the aforementioned cues when classifying the samples and, as a result, alleviates overfitting.

Our contributions are as follows: (1) We present a novel two-stage detection approach that uses large amounts of natural talking faces for strong generalisation and robustness performance; this opens up the avenue for future forgery detection works to exploit the ubiquitous real videos online. (2) We propose, for the first stage, a non-contrastive self-supervised framework that learns temporally dense representations, and we validate its design for our task through ablations. (3) We achieve state-of-the-art performance in experiments that test cross-manipulation generalisation and robustness to common corruptions, and highlight the factors responsible for our method's performance.

2. Related Works

2.1. Face forgery detection

General approaches. Earlier works using convolutional neural networks (CNNs) include recasting steganalysis features as CNNs [32], constraining convolutional filters [13], and using shallow networks [3] to suppress high-level content. However, an unconstrained Xception [25] network outperforms these approaches on more recent forgery types [94]. Other works aim at detecting inconsistent head poses [108] or irregular eye blinking [69], although more recent fakes may not exhibit such anomalies. More recently, works have focused on attention mechanisms [35, 102, 112], exploiting the frequency spectrum [41, 43, 66, 73, 75, 79, 92], detecting anomalies in features from a face recognition network [104], or using extra identity information [6, 33, 38]. [44] and [111] use self-supervision for frame-based detection, but do not study the effect of using many real samples.

Cross-manipulation generalisation. Detectors often generalise poorly to unseen forgeries [19, 34, 53, 68, 105]. Approaches to improve generalisation include applying augmentations [105], reconstructing the input as an auxiliary task [34, 40, 85], mining frequency cues [75, 79], truncating classifiers [19], focusing on self-consistency [60, 68, 70, 113], or using spatio-temporal convolutional networks [45].

However, it has been shown that it is especially challenging to achieve cross-manipulation generalisation and at the same time perform well on corrupted data [53]. A closely related work to ours is LipForensics [53], which addresses this by finetuning a network that was pre-trained to perform lipreading. Unlike our method, it requires a large-scale *labelled* dataset and focuses exclusively on the mouth region. Very recently, [114] report high generalisation by reducing the spatial kernel sizes of convolutional layers to 1, thus learning temporal inconsistencies while ignoring spatial ones. By contrast, we target *spatio-temporal* irregularities that may be more consistent with human perception

of forgery cues. Finally, some recent works have focused on mismatches between the visual and auditory modalities in fake videos [5, 26, 65, 81, 117]. Our work, on the other hand, is *visual-only* at test-time: It uses the audio modality only for cross-modal supervision in an intermediate step, in which only real videos are used.

2.2. Self-supervised learning

Image SSL. Recently, contrastive learning using the InfoNCE loss [88] has become a popular approach in image representation learning [20, 21, 54, 57, 88, 100, 107]. In this paradigm, the similarity between two views of an image is maximised, while different images ("negatives") are repelled. Contrastive learning has also been used to learn dense visual representations [91, 106]. However, recent works that remove negatives generally outperform contrastive approaches [11, 16–18, 22, 48, 109]. Our work is partly inspired by BYOL [48], which uses a slowly-evolving teacher network that produces targets for a student to predict. The first stage of our approach can be viewed as an extension of BYOL to the audiovisual setting, in which we have a student-teacher pair for each modality and the output representations are temporally dense. Recent works [86] and [42,93] also use BYOL-style training but are for audio-only learning and action recognition, respectively.

Audiovisual SSL. Many works exploit audiovisual correspondence for video action recognition [7–10,29,64,76,82, 90]. However, these approaches learn a single representation for a video clip, which is less suitable for modelling the fine-grained movements of a speaking face. The very recent work [77] proposes to learn, in a contrastive manner, both global and local representations that are agnostic to the specific downstream task. In contrast, aside from methodological differences, our work focuses on cross-dataset generalisation and robustness for face forgery detection. Audiovisual methods have also been proposed for applications involving faces (*e.g.*, audiovisual synchronisation and biometric matching). In general, methods that model lexical content tend to contrast samples from the same video for identity invariance [29–31]. Conversely, works that learn identity embeddings tend to match misaligned video-audio pairs from the same person for invariance to lexical content [83, 84]. We posit that it is beneficial to capture both types of information for forgery detection and hence directly predict aligned embeddings.

Generalisation via self-supervision. It has been shown that using self-supervision as an auxiliary task, *e.g.*, predicting rotations [47] or solving jigsaw puzzles [87], can improve generalisation on the main task at hand [15,46,58]. We use a similar idea for improving generalisation for forgery detection, but we *learn* the targets in a separate stage before using them to define the auxiliary task.

3. Method

RealForensics comprises two stages (see Figure 3). The first stage involves learning temporally dense video representations using cross-modal self-supervision from many natural talking faces. These representations are subsequently used as prediction targets in the second stage to regularise the binary forgery classification task.

3.1. Stage 1: representation learning

Given real videos and the corresponding audio, we aim to learn video representations that capture information associated with facial appearance and behaviour. Cues like facial movements are temporally fine-grained by nature, and hence we wish to learn *temporally dense* representations, *i.e.*, an embedding per frame. We use a student-teacher framework without contrasting negatives for the following reasons. (1) This style of training has resulted in state-of-the-art results in image representation learning [48]; (2) it incentivises the network to retain all information shared by the two modalities [48]; and (3) it obviates the need for large batch sizes [20] or a queue [54] to store the negatives.

Formulation. We assume access to a large dataset $\mathcal{D}_r$ of real talking faces. A sample $x \in \mathcal{D}_r$ is a video $x^v \in \mathbb{R}^{T_v \times H \times W \times 3}$ (of T_v video frames, height H, and width W) with its corresponding audio, represented as a log-mel spectrogram, $x^a \in \mathbb{R}^{T_a \times L}$ (of T_a audio frames and L mel filters). We ensure that $T_a = 4T_v$.

Our architecture consists of a student and teacher pair for each modality. The teachers produce targets that the students from the other modality must predict. Specifically, teacher video and audio backbone networks, f_t^v and f_t^a, produce embeddings $e_t^v = f_t^v(x^v)$ and $e_t^a = f_t^a(x^a)$ from the inputs, which are then passed through projectors, g_t^v and g_t^a, to yield dense video and audio targets, $z_t^v = \text{norm}(g_t^v(e_t^v)) \in \mathbb{R}^{T_v \times C}$ and $z_t^a = \text{norm}(g_t^a(e_t^a)) \in \mathbb{R}^{T_v \times C}$, where C is the dimensionality of the embeddings and $\text{norm}(\cdot)$ denotes l_2 normalisation across the channel dimension. Note that the audio backbone subsamples the temporal dimension such that the video and audio embeddings have the same shape. The students have the same architecture as their corresponding teachers, except that each student additionally contains a predictor, whose job is to predict the targets from the other modality. Let the video and audio predictions be $p^v = \text{norm}(h^v(z_s^v))$ and $p^a = \text{norm}(h^a(z_s^a))$, respectively, where h^v and h^a denote the predictors and z_s^v and z_s^a are the unnormalised student representations after the student projectors; then the loss is

$$\mathcal{L} = \frac{1}{2}||\text{sg}\,(z_t^v) - p^a||_F^2 + \frac{1}{2}||\text{sg}\,(z_t^a) - p^v||_F^2, \quad (1)$$

where $||\cdot||_F$ denotes the Frobenius norm, and sg, which stands for "stop-gradient," emphasises that the targets are

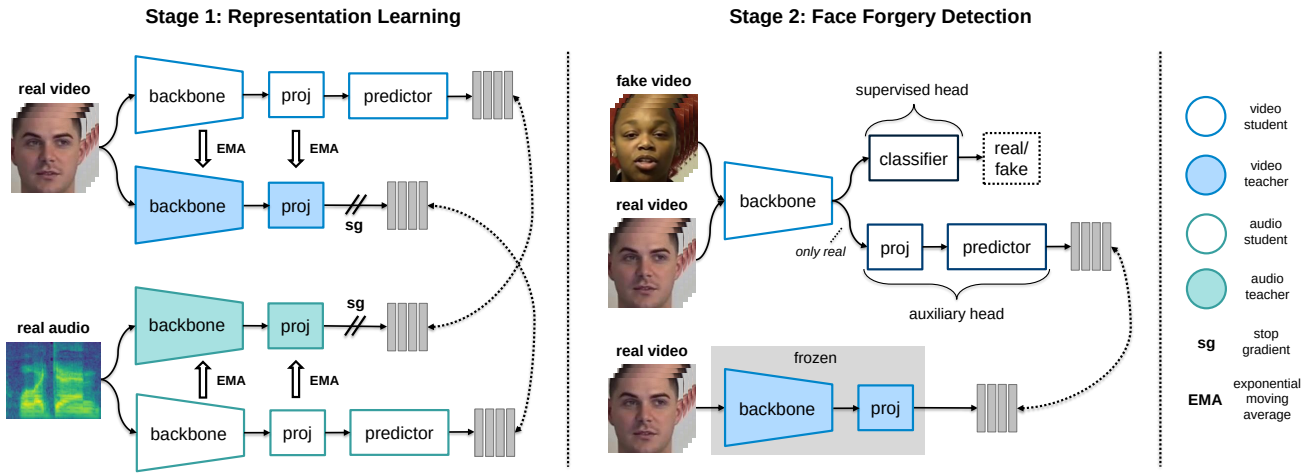

Figure 3. **The two stages of RealForensics**. In stage 1, the aim is to learn, in a self-supervised manner, frame-wise representations that capture information on natural facial behaviour and appearance. We utilise an audiovisual, cross-modal, student-teacher framework, whereby the student networks ingest real video and audio and try to predict the corresponding targets generated from the other modality. We also randomly mask the student inputs (omitted from the diagram for clarity). The teacher networks are momentum encoders that are updated via an exponential moving average (EMA), as in [48]. In stage 2, the detector performs face forgery classification, while predicting the video targets produced by the (now frozen) video teacher from stage 1; only real videos contribute to the prediction loss. The video student from stage 1 is used to initialise the backbone. This multi-task formulation likely incentivises the network to detect forgeries based on stable cues that generalise well to unseen forgeries and are robust to low-level perturbations. Best viewed in colour.

treated as constants. The total loss is averaged over all samples. The students are optimised via gradient descent, and the teachers are exponential moving averages of the students. That is, if we denote the video teacher weights as ψ^v and the corresponding student weights as θ^v, then at each iteration

$$\psi^v \leftarrow \mu\psi^v + (1-\mu)\theta^v, \tag{2}$$

where μ is a momentum parameter close to 1. The audio teacher weights are updated similarly.

Transformer as predictor. BYOL shows that the predictor is a necessary component to avoid *representation collapse*, a situation where the representations for all samples are the same [48]. We observe the same for our framework (see Section 5). Whereas BYOL outputs global representations and thus uses an MLP as a predictor, we find that a shallow transformer is suitable for our dense representation learning task (see the appendix for an ablation).

Random masking. We also find that random masking results in better representations (see Section 5). For videos, we zero random rectangular regions in frames [115], consistent across a whole video clip, as well as erase a random number of consecutive frames. For spectrograms, we erase a random number of consecutive audio frames and frequency bins. This is similar to the SpecAugment method [89], but without the time warping step. We apply this masking only to the inputs of the students. Intuitively, this forces the students to make use of context to infer the missing information and prevents them from overly relying on specific features of the input, *e.g.*, the mouth region.

Implementation details. Unless specified otherwise, we use the following settings for this stage (see the appendix for more details).

- *Inputs*. We extract the faces using face detection and alignment. A clip consists of 25 frames. The log-mel spectrograms contain 80 mel filters and 100 audio frames. During training, we randomly crop the video clips to size 140×140 and resize them to 112×112. We randomly apply horizontal flipping and grayscale transformation, each with probability 0.5. As mentioned, we also randomly mask the students' inputs.
- *Backbones.* The video backbone is a Channel-Separated Convolutional Network (CSN) [101]; we set the temporal strides to 1 to prevent temporal subsampling. The audio backbone is a ResNet18 [55], with the strides in the convolutional layers modified such that it subsamples the temporal dimension by 4, thus matching the temporal span of the video backbone's output.
- *Projectors.* The projection network used for both the video and audio modalities is a single 1×1 convolutional layer with output dimension of 256, followed by batch normalisation (BN) [61]. We find that this BN layer helps with training, similarly to [22].

- *Predictors.* The predictor for both modalities is a 1-block transformer encoder. It follows the design of a ViT block [39]. We use 8 attention heads, each with dimension 64, MLP dimension of 2048, and replace layer normalisation [12] with batch normalisation [61] before the MLP.
- *Optimisation.* We use the AdamP optimiser [59] with learning rate 7×10^{-4} and weight decay 10^{-2}. We train for 150 epochs, with an initial 20-epoch linear warmup followed by a cosine decay schedule for the learning rate [74]. The predictors' lr are kept fixed [22]. The EMA momentum parameters for the teachers are set to 0.999.

3.2. Stage 2: multi-task forgery detection

The aim of this work is to learn a *visual-only* forgery detector. Indeed, many forgery datasets do not officially release audio along with the videos [62, 72, 94]. As a result, at this stage we discard the audio student-teacher pair, after having served its purpose in stage 1.

We propose to use the video teacher from stage 1 to produce targets for our network to predict. At the same time, the network performs forgery detection, in a multi-task fashion. Note that the teacher is frozen in this stage. Using this auxiliary loss likely encourages the network to classify real and fake videos by focusing on high-level spatio-temporal characteristics of facial appearance and behaviour.

Formulation. We again use our dataset of real faces $\mathcal{D}_r$, but we now also assume access to a dataset of fake videos, $\mathcal{D}_f$.[1] Our full dataset is thus $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_f$. Our architecture consists of a shared backbone f with weights θ_b and two heads: a supervised head with weights θ_s for the forgery classification loss and an auxiliary one q with weights θ_a for the target prediction loss. The auxiliary loss is given by

$$\mathcal{L}_a(\mathcal{D}_r; \theta_b, \theta_a) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_r} ||q\left(f(x^v; \theta_b); \theta_a\right) - t(x^v)||_F^2, \quad (3)$$

where t is the teacher from stage 1, and the auxiliary head's and teacher's outputs are l_2-normalised as in stage 1.

The supervised loss $\mathcal{L}_s(\mathcal{D}; \theta_b, \theta_s)$ is a logit-adjusted version of binary cross entropy, as proposed in [80], to address any class imbalance (see the appendix for details). Moreover, to obtain the logits, we l_2-normalise the feature vectors and the weights of the last linear layer (and set its bias to 0), thus obtaining a cosine classifier [103]. This combines better with the auxiliary loss, which can also be cast in terms of cosine similarity. Finally, the objective is given by

$$\min_{\theta_b, \theta_s, \theta_a} \mathcal{L}_s(\mathcal{D}; \theta_b, \theta_s) + w\mathcal{L}_a(\mathcal{D}_r; \theta_b, \theta_a), \quad (4)$$

where w is a scaling factor, which we set to 1.

[1] In practice, the real samples for this stage include our auxiliary dataset as well as the real samples from the forgery dataset.

Implementation details. The video teacher is transferred from stage 1 and remains frozen henceforth. The backbone's architecture is the same as the video backbone in stage 1, and we initialise it with the learned weights. The auxiliary head is comprised of a randomly initialised projector and predictor as in stage 1. The supervised head is a cosine classifier, as previously described. A batch consists of 32 fake and 256 real samples, to effectively make use of the many more real samples available. We use the AdamP optimiser with learning rate 3×10^{-4} and the same preprocessing and augmentations described in stage 1. We train for 150 epochs and use the validation set for early stopping.

4. Experiments

Auxiliary dataset. We use the LRW dataset [28] without the labels for our extra real samples. It contains 500,000 videos of talking faces with hundreds of different identities. This dataset was also used by LipForensics [53], which allows for fairer comparisons. In addition, its size strikes a balance between meaningful results and non-prohibitive computational costs. We present results for another dataset, VoxCeleb2 [27], in Section 5.

Forgery datasets. We use the following forgery datasets: (1) **FaceForensics++** (FF++) [94] consists of 1,000 real videos and 4,000 fake videos, generated using two face swapping methods, Deepfakes [1] and FaceSwap [2], and two face reenactment methods, Face2Face [99] and NeuralTextures [98]. Unless stated otherwise, we use the mildly compressed version of the dataset (c23). As in [53, 94], we take the first 270 frames for each training video, and the first 110 frames for each validation/testing video. (2) **FaceShifter** [67] and (3) **DeeperForensics** [62] are state-of-the-art face swapping methods that have been applied to the real videos of FF++; we use the test videos, according to the FF++ split. (4) **CelebDF-v2** [72] is a challenging face swapping dataset with 518 test videos. (5) **DFDC** is a subset of the Deepfake Detection Challenge Dataset (DFDC) [37] used in [53]. It features 3,215 videos, many of which have been subjected to strong perturbations.

Evaluation metrics. Following *e.g.*, [2, 53, 94, 114, 117], we use accuracy and area under the receiver operating characteristic curve (AUC) for evaluation. We use video-level metrics: For a single video we first uniformly sample non-overlapping clips and then average all clip predictions across the video.

4.1. Cross-manipulation generalisation

A deployed detector is expected to recognise fake videos that were created using methods *not seen during training*, a non-trivial task in practice [53, 68, 85, 114]. In this section, we follow the protocol used in [53, 70, 85] to evaluate our detector's ability to generalise to unseen manipulations.

Method	Train on remaining three			
	DF	FS	F2F	NT
Xception [94]	93.9	51.2	86.8	79.7
CNN-aug [105]	87.5	56.3	80.1	67.8
Patch-based [19]	94.0	60.5	87.3	84.8
Face X-ray [68]	99.5	93.2	94.5	92.5
CNN-GRU [95]	97.6	47.6	85.8	86.6
LipForensics [53]	99.7	90.1	99.7	99.1
AV DFD [117]	100.	90.5	99.8	98.3
FTCN [114]	99.9	99.9	99.7	99.2
CSN	98.8	87.9	98.7	88.6
RealForensics (ours)	100.	97.1	99.7	99.2

Table 1. **FF++ cross-manipulation generalisation.** AUC scores (%) for each FF++ manipulation type after training on the remaining types. We use the test sets of Deepfakes (DF), FaceSwap (FS), Face2Face (F2F), and NeuralTextures (NT), as well as the real test videos. Top-2 best methods are underlined.

Method	CDF	DFDC	FSh	DFo	Avg
Xception [94]	73.7	70.9	72.0	84.5	75.3
CNN-aug [105]	75.6	72.1	65.7	74.4	72.0
Patch-based [19]	69.6	65.6	57.8	81.8	68.7
Face X-ray [68]	79.5	65.5	92.8	86.8	81.2
CNN-GRU [95]	69.8	68.9	80.8	74.1	73.4
Multi-task [85]	75.7	68.1	66.0	77.7	71.9
DSP-FWA [71]	69.5	67.3	65.5	50.2	63.1
Two-branch [79]	76.7	—	—	—	—
LipForensics [53]	82.4	73.5	97.1	97.6	87.7
FTCN [114]	**86.9**	74.0	98.8	98.8	89.6
CSN	69.4	68.1	87.9	89.3	78.7
RealForensics (ours)	**86.9**	**75.9**	**99.7**	**99.3**	**90.5**

Table 2. **Cross-dataset generalisation.** AUC scores (%) on CelebDF-v2 (CDF), DeepFake Detection Challenge (DFDC), FaceShifter (FSh), and DeeperForensics (DFo), after training on FaceForensics++. Best results are in **bold**.

Method	Settings		Accuracy	
	Arch	# params	FSh	DFo
LipForensics [53]	RN+TCN [78]	36.0	87.5	90.4
FTCN [114]	FTCN [114]	26.6	93.9	91.1
RealForensics (ours)	CSN [101]	**21.4**	**97.1**	**97.1**

Table 3. **Parameters and generalisation accuracy.** Number of parameters (in millions), at test-time, for related state-of-the-art methods, and accuracy on FaceShifter (FSh) and DeeperForensics (DFo) after training on FaceForensics++. Best results are in **bold**.

Table 1 shows results obtained by RealForensics on each manipulation type in the FF++ dataset after training on the remaining types. Our detector works on par with the state-of-the-art without (1) using auxiliary labelled supervision [53], (2) heavily constraining the network by freezing large parts [53] or removing spatial convolutions [114], nor (3) using audio at test-time [117]. We also outperform the baseline of training a CSN [101] network on the forgery data (with the same augmentations as RealForensics), indicating the effectiveness of leveraging real data using our approach.

We also evaluate *cross-dataset* generalisation by training on FF++ and then testing *a single model* on unseen, challenging datasets: CelebDF-v2 [72], DFDC [37], FaceShifter [67], and DeeperForensics [62]. The AUC results are given in Table 2. Our method achieves state-of-the-art results on all datasets, suggesting that our detector performs well when exposed to more advanced forgeries than originally trained on. RealForensics also beats the CSN baseline by a large margin. Finally, as seen in Table 3, we achieve higher generalisation accuracy on FaceShifter and DeeperForensics than related methods, with fewer network parameters at test-time.

4.2. Robustness to common corruptions

In addition to good cross-manipulation generalisation, detectors should also be able to withstand common corruptions that videos may be subjected to on social media. We follow [53] to assess robustness to *unseen* perturbations. As in [53], we train on FF++ with grayscale clips and no augmentation other than horizontal flipping and random cropping, to avoid any intersection between train- and test-time perturbations. The set of perturbations, proposed in [62], are changes in saturation and contrast, block-wise occlusions, Gaussian noise and blur, pixelation and video compression. Each perturbation type is applied at five different intensity levels. Table 4 presents the average AUC across all intensity levels for each corruption type. RealForensics suffers significantly less from common corruptions than frame-based methods that target low-level cues, such as [19, 68], and also outperforms LipForensics and FTCN. (We use FTCN's publicly available model[2], which was trained on FF++ c23.) We notice that, relative to RealForensics and LipForensics, FTCN struggles on Gaussian noise and video compression (see also Figure 4), which disturb temporal coherence. This may be explained by FTCN's lack of spatial convolutions.

5. Ablations

In this section, we present ablations to understand the factors responsible for our method's performance. See the appendix for more ablations.

Framework ablation. In Table 5, we ablate different components of our method and inspect its generalisation per-

[2] https://github.com/yinglinzheng/FTCN

Method	Clean	Saturation	Contrast	Block	Noise	Blur	Pixel	Compress	Avg
Xception [94]	99.8	99.3	98.6	**99.7**	53.8	60.2	74.2	62.1	78.3
CNN-aug [105]	99.8	99.3	99.1	95.2	54.7	76.5	91.2	72.5	84.1
Patch-based [19]	99.9	84.3	74.2	99.2	50.0	54.4	56.7	53.4	67.5
Face X-ray [68]	99.8	97.6	88.5	99.1	49.8	63.8	88.6	55.2	77.5
CNN-GRU [95]	99.9	99.0	98.8	97.9	47.9	71.5	86.5	74.5	82.3
LipForensics [53]	99.9	**99.9**	**99.6**	87.4	73.8	**96.1**	95.6	95.6	92.5
FTCN [114]	99.4	99.4	96.7	97.1	53.1	95.8	98.2	86.4	89.5
RealForensics (ours)	99.8	99.8	**99.6**	98.9	**79.7**	95.3	**98.4**	**97.6**	**95.6**

Table 4. **Robustness to common corruptions.** Average AUC scores (%) across five intensity levels for each corruption type proposed in [62]. We also present, for each method, the average score across all corruptions. Best results are in **bold**. For a more detailed analysis, see the appendix.

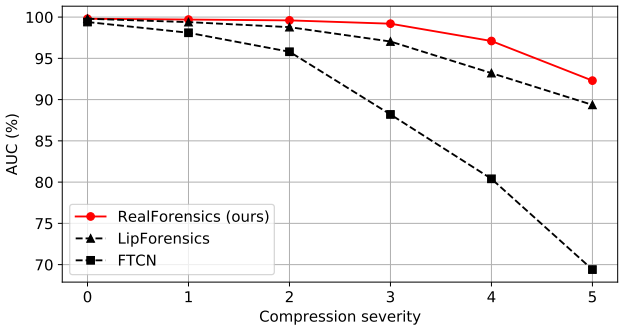

Figure 4. **Robustness to compression**. AUC scores (%) on FaceForensics++ (FF++) at various H.264 video compression rates (23, 30, 32, 35, 38, 40), after training on FF++ with light compression (rate of 23).

formance on FaceShifter and DeeperForensics after training on FaceForensics++. We make the following observations. First, simply training a CSN [101] model without our two-stage framework leads to a drop in accuracy of about 14%. Second, transferring the weights from stage 1 to the video backbone and finetuning the network on forgery data, without the auxiliary loss in stage 2, results in a drop of about 2%. This suggests that forcing the network to predict the video representations along with its main task has a positive regularisation effect. Finally, we observe modest improvements by employing logit adjustment [80] for imbalanced classification and using time masking and random erasing [115].

Representation learning ablation. For stage 1 of our method, we propose to learn temporally dense representations without contrasting negatives. Here, we test our choices against alternatives. We train the network with all combinations of the following settings: dense/global representations, with/without negatives, and with/without a predictor network. For global representation learning, we average-pool the output of the backbone networks, and use MLPs for the projector and predictor. To employ negatives, we use a queue of 65,536 samples and use the InfoNCE loss [88] with a temperature of 0.07. Note that global learning with negatives is similar to cross-modal contrastive learning used in *e.g.*, [76,82]. The predictor network used for global learning is an MLP and for dense learning a one-block transformer. Global learning with negatives and a predictor is analogous to the recent image representation learning method MoCo v3 [23], but for cross-modal learning. More information can be found in the appendix.

In Table 6, we show accuracy scores on FaceShifter and DeeperForensics after training on FaceForensics++. We see that dense representations lead to significantly better performance than global. Also, consistent with the original BYOL method, we find that without negatives and without a predictor the outcome is *representation collapse* [48]. Without negatives and with global representations, no collapse was observed (with a predictor), but we had trouble achieving competitive performance. This may have to do with optimisation difficulties encountered without contrastive learning, since the subsequent inclusion of negatives yielded better results. However, adding negatives *does not* seem to help when we use *dense* learning (and a predictor).

Effect of number of real samples. Next, we vary the number of LRW samples in both stages of our method to see the effect on generalisation. As a baseline, we also consider simply treating the problem as an imbalanced classification task, *i.e.*, training the model with logit adjustment (but without our proposed method). We can see in Figure 5 that RealForensics benefits from a large number of real samples. Moreover, although generalisation for the baseline does increase with more real samples, the increase is significantly less than for RealForensics.

Using a different auxiliary dataset. Here, we use the VoxCeleb2 dataset [27] for our extra real samples. It con-

Method	FSh	DFo
RealForensics (ours)	**97.1**	**97.1**
only CSN	82.1	83.1
stage 1 + finetune	95.0	95.2
w/o logit adjustment	95.7	96.4
w/o time masking	96.1	95.9
w/o random erasing	96.3	96.3

Table 5. **Framework ablation.** Accuracy scores (%) on FaceShifter (FSh) and DeeperForensics (DFo) after training on FaceForensics++. Refer to subsection "Framework ablation" for a discussion. Best results are in **bold**.

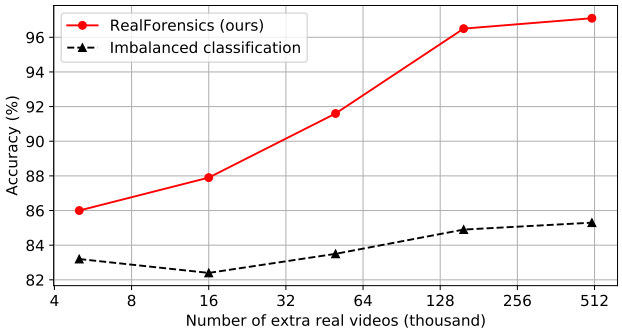

Figure 5. **Effect of number of real samples**. Accuracy scores (%) as a function of the number of real samples from LRW, in log-scale. We show results for our method as well as a baseline where we treat the task as an imbalanced classification one. We average the accuracy of FaceShifter and DeeperForensics after training on FaceForensics++.

Settings			Accuracy (%)	
Global/Dense	Negatives	Predictor	FSh	DFo
Global	✗	✗	n/a	n/a
Global	✗	✓	70.7	74.1
Global	✓	✗	87.9	88.6
Global	✓	✓	87.9	89.1
Dense	✗	✗	n/a	n/a
Dense	✗	✓	**97.1**	**97.1**
Dense	✓	✗	94.0	95.7
Dense	✓	✓	96.4	96.8

Table 6. **Representation learning ablation.** We ablate different components of our representation learning stage (stage 1). Note that "n/a" means that representation collapse was observed in stage 1. Refer to subsection "Representation learning ablation" for a discussion. Best results are in **bold**. Default setting is highlighted.

tains about 1 million videos of talking faces with various identities. We train with the same hyperparameters as for LRW. The AUC results (in %) on CelebDF-v2, DFDC, FaceShifter, and DeeperForensics after training on FaceForensics++ are 82.9, 78.9, 99.3, and 98.8, respectively. This suggests that competitive results can be obtained with minimal tuning using a different dataset.

6. Limitations / Societal Impact

The strong generalisation of RealForensics comes at a cost of higher computational demands during training than methods that do not use auxiliary datasets; however, this is not the case at test-time. Moreover, our detector takes videos as input, and thus does not work for single images. Despite state-of-the-art accuracy, we also observe that when our network produces wrong predictions, they are often confidently wrong, so the probabilities outputted by the model should be interpreted with care. This issue of *model calibration* is common in deep learning models [51], including forgery detectors; thus, an important future direction would be to apply methods in calibration literature [51] to detectors.

Although the purpose of research on forgery detection is to protect society, there are a few concerns that should be kept in mind. For example, pointing out the flaws in current face forgeries could facilitate the development of even better fake videos in the future. This, however, is less of an issue for methods that do not target a *specific* cue, such as RealForensics. Further, it is not prudent for a deployed system to rely exclusively on a single detection method. For greater effectiveness, it should employ an ensemble of independent approaches.

7. Conclusion

In this paper, we propose RealForensics, an approach that uses large amounts of unlabelled real data to detect fake videos. We have shown that our method simultaneously achieves strong cross-manipulation generalisation and robustness to common corruptions. We hope our study encourages future research on leveraging real faces for robust forgery detection.

Acknowledgements. We thank Konstantinos Vougioukas for fruitful discussions. This work has been supported in part by Meta Platforms through research funding made available directly to Imperial College London (project P93445: Cross-modal learning of emotions). Alexandros Haliassos was financially supported by an Imperial President's PhD Scholarship. All training, testing, and ablation studies have been conducted at Imperial College.

References

[1] Deepfakes. https://github.com/deepfakes/faceswap. [Accessed: 2020-11-12]. 5

[2] Faceswap. https://github.com/MarekKowalski/FaceSwap. [Accessed: 2020-11-12]. 5

[3] 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. IEEE, 2018. 1, 2

[4] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020. 14

[5] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 660–661, 2020. 3

[6] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR workshops*, volume 1, 2019. 2

[7] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019. 3

[8] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 3

[9] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 3

[10] Yuki M Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *arXiv preprint arXiv:2006.13662*, 2020. 3

[11] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019. 3

[12] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[13] 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*, pages 5–10, 2016. 2

[14] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 16

[15] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 3

[16] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 3

[17] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 3

[18] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 3

[19] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European Conference on Computer Vision*, pages 103–120. Springer, 2020. 1, 2, 6, 7, 13

[20] 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, 2020. 3

[21] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[22] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 3, 4, 5, 17

[23] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 7

[24] Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Bo Zhang. Magdr: Mask-guided detection and reconstruction for defending deepfakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9014–9023, 2021. 1

[25] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 2

[26] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 439–447, 2020. 3

[27] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 5, 7

[28] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016. 5

[29] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 2, 3, 14

[30] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3965–3969. IEEE, 2019. 3, 14, 15

[31] Soo-Whan Chung, Hong Goo Kang, and Joon Son Chung. Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision. *arXiv preprint arXiv:2004.14326*, 2020. 3

[32] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, pages 159–164, 2017. 2

[33] Davide Cozzolino, Andreas Rossler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15108–15117, 2021. 2

[34] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 1, 2

[35] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020. 1, 2

[36] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 16

[37] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 1, 2, 5, 6, 16

[38] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Identity-driven deepfake detection. *arXiv preprint arXiv:2012.03930*, 2020. 2

[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[40] Mengnan Du, Shiva Pentyala, Yuening Li, and Xia Hu. Towards generalizable forgery detection with locality-aware autoencoder. *arXiv e-prints*, pages arXiv–1909, 2019. 2

[41] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7890–7899, 2020. 2

[42] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021. 3

[43] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, pages 3247–3258. PMLR, 2020. 2

[44] Sheldon Fung, Xuequan Lu, Chao Zhang, and Chang-Tsun Li. Deepfakeucl: Deepfake detection via unsupervised contrastive learning. *arXiv preprint arXiv:2104.11507*, 2021. 1, 2

[45] Ipek Ganiyusufoglu, L Minh Ngô, Nedko Savov, Sezer Karaoglu, and Theo Gevers. Spatio-temporal features for generalized detection of deepfake videos. *arXiv preprint arXiv:2010.11844*, 2020. 2

[46] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8059–8068, 2019. 3

[47] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 3

[48] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 2, 3, 4, 7, 16

[49] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3473–3481, 2021. 1

[50] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018. 1

[51] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 8

[52] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Supplementary material for lips don't lie: A generalisable and robust approach to face forgery detection. 14

[53] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5039–5049, 2021. 1, 2, 5, 6, 7, 13, 14, 15, 16, 17

[54] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3, 17

[55] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 17

[56] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4360–4369, 2021. 13

[57] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020. 3

[58] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *arXiv preprint arXiv:1906.12340*, 2019. 3

[59] Byeongho Heo, Sanghyuk Chun, Seong Joon Oh, Dongyoon Han, Sangdoo Yun, Gyuwan Kim, Youngjung Uh, and Jung-Woo Ha. Adamp: Slowing down the slowdown for momentum optimizers on scale-invariant weights. *arXiv preprint arXiv:2006.08217*, 2020. 5

[60] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 101–117, 2018. 2

[61] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 4, 5

[62] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2889–2898, 2020. 1, 5, 6, 7, 14, 16

[63] Sohail Ahmed Khan and Hang Dai. Video transformer for deepfake detection with incremental learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1821–1828, 2021. 1

[64] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv preprint arXiv:1807.00230*, 2018. 3

[65] Pavel Korshunov and Sébastien Marcel. Speaker inconsistency detection in tampered video. In *2018 26th European signal processing conference (EUSIPCO)*, pages 2375–2379. IEEE, 2018. 3

[66] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6458–6467, 2021. 2

[67] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5083, 2020. 1, 5, 6, 16

[68] 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. 1, 2, 5, 6, 7, 13

[69] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. 2

[70] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018. 2, 5

[71] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 6

[72] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020. 1, 5, 6, 16

[73] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 772–781, 2021. 2

[74] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[75] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16317–16326, 2021. 2

[76] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. *arXiv preprint arXiv:2009.09805*, 2020. 3, 7

[77] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Contrastive learning of global and local audio-visual representations. *arXiv preprint arXiv:2104.05418*, 2021. 3

[78] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323. IEEE, 2020. 6, 14

[79] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *European Conference on Computer Vision*, pages 667–684. Springer, 2020. 1, 2, 6, 13

[80] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar.

Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 5, 7, 16

[81] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2823–2832, 2020. 1, 3

[82] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021. 3, 7

[83] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–88, 2018. 2, 3

[84] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8427–8436, 2018. 3

[85] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019. 2, 5, 6

[86] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Byol for audio: Self-supervised learning for general-purpose audio representation. *arXiv preprint arXiv:2103.06695*, 2021. 3

[87] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 3

[88] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 7

[89] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019. 4

[90] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, Joao F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020. 3

[91] Pedro O Pinheiro, Amjad Almahairi, Ryan Y Benmalek, Florian Golemo, and Aaron Courville. Unsupervised learning of dense visual representations. *arXiv preprint arXiv:2011.05499*, 2020. 3

[92] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pages 86–103. Springer, 2020. 1, 2

[93] Adrià Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Patraucean, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. *arXiv preprint arXiv:2103.16559*, 2021. 3

[94] 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. 1, 2, 5, 6, 7, 13, 16

[95] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3(1), 2019. 6, 7, 13

[96] Abhinav Shukla, Stavros Petridis, and Maja Pantic. Does visual self-supervision improve learning of speech representations for emotion recognition. *IEEE Transactions on Affective Computing*, 2021. 2

[97] Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3609–3618, 2021. 1

[98] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 5

[99] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 5

[100] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 3

[101] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019. 4, 6, 7, 14, 16, 17

[102] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14923–14932, 2021. 2

[103] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 5, 16

[104] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*, 2019. 2

[105] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020. 1, 2, 6, 7, 13

[106] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised

visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 3

[107] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 3

[108] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019. 2

[109] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 3

[110] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 17

[111] Jian Zhang, Jiangqun Ni, and Hao Xie. Deepfake videos detection using self-supervised decoupling network. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 2

[112] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2185–2194, 2021. 1, 2

[113] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15023–15033, 2021. 2

[114] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15044–15054, 2021. 1, 2, 5, 6, 7, 13, 15

[115] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020. 4, 7

[116] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839. IEEE, 2017. 1

[117] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021. 3, 5, 6

[118] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z Li. Face forgery detection by 3d decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2929–2939, 2021. 1

Method	Accuracy (%)			AUC (%)		
	Raw	c23	c40	Raw	c23	c40
Xception [94]	99.0	97.0	89.0	99.8	99.3	92.0
CNN-aug [105]	98.7	96.9	81.9	99.8	99.1	86.9
Patch-based [19]	**99.3**	92.6	79.1	**99.9**	97.2	78.3
Two-branch [79]	—	—	—	—	99.1	91.1
Face X-ray [68]	99.1	78.4	34.2	99.8	97.8	77.3
CNN-GRU [95]	98.6	97.0	90.1	**99.9**	99.3	92.2
LipForensics [53]	98.9	98.8	94.2	**99.9**	99.7	98.1
FTCN [114]	—	**99.1**	—	—	**99.8**	98.3
RealForensics (ours)	**99.3**	**99.1**	**96.1**	**99.9**	**99.8**	**99.5**

Table 7. **In-distribution performance.** Accuracy and AUC scores on the test set of FaceForensics++ (FF++) after training on FF++. We repeat experiments for the dataset's three compression types: raw (no compression), c23 (mild compression), and c40 (strong compression. Best results are in **bold**.

	Ours	LipForensics [53]	FTCN [114]
ForgeryNet	**71.8**	66.7	57.3

Table 8. **Generalisation to ForgeryNet.** AUC scores (%) on the val set of ForgeryNet after training on FF++. Best results are in **bold**.

A. More Experiments

A.1. In-distribution performance

Although our approach has been developed for cross-manipulation generalisation and robustness, for completeness we present results for in-distribution performance in Table 7. For each compression level (raw, c23, c40), we train on the training set and show results on the corresponding test set. We are on par with the state-of-the-art in the no/low compression regime, while outperforming the other methods on the more compressed data.

A.2. Generalisation to ForgeryNet

In Table 8, we provide results on generalisation performance to the newly-released ForgeryNet dataset [56]. We compare our model with the publicly-available LipForensics and FTCN models (all trained on FF++). RealForensics significantly outperforms both.

A.3. Detailed analysis of robustness

Following [53], we present more detailed results on robustness by plotting AUC as a function of corruption severity (see Figure 7). On average, RealForensics deteriorates less abruptly as severity increases than other methods, with especially noteworthy results on video compression, which is ubiquitous on social media. We also highlight our significantly higher results over LipForensics on block-wise

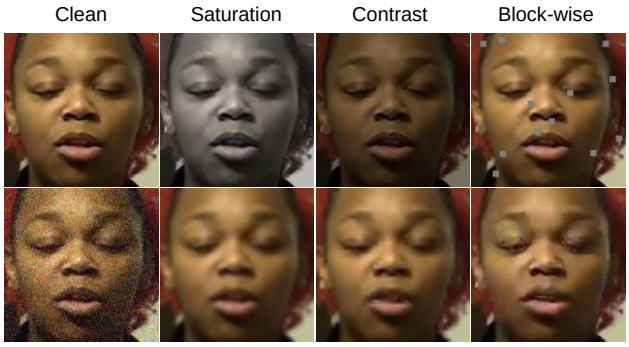

Figure 6. **Examples of corruptions.** A clean frame from a real FaceForensics++ video along with the same frame but corrupted with various perturbations. For more information on this set of corruptions, see [62].

Crop	Acc (%)		AUC (%)
	FSh	DFo	FS
Full face	**97.1**	**97.1**	**97.1**
Mouth	95.5	95.0	88.9

Table 9. **Full face versus mouth.** Accuracy and AUC scores when training on full faces and mouth crops. We test on FaceShifter (FSh) and DeeperForensics (DFo) after training on FaceForensis++ (FF++). We also test on FaceSwap (FS) after training on the remaining three FF++ types. Best results are in **bold**. Default setting is highlighted.

distortions (*i.e.*, occlusions), which are likely influenced by our method's use of the whole face rather than solely the mouth. For example, in some cases the mouth may be occluded while other parts of the face are not.

A.4. More ablations

Full face versus mouth. In the main text, we argue that focusing only on the mouth region, like LipForensics [53], may be suboptimal for performance. We validate this by training (for both stages 1 and 2) on mouth crops and comparing the performance with the default setting. As shown in Table 10, our method consistently benefits from using the full face rather than the mouth, which was not observed for LipForensics [52]. This may be due to the cross-modal prediction task being more general than lipreading. For example, the video network is encouraged to retain information about the eyes to better model expression (which correlates with audio); on the other hand, a model trained to perform lipreading may focus predominantly on the mouth region.

Effect of clip size. Table 9 shows the effect on generalisation when changing the video clip size (default is 25 frames per clip). We observe that generalisation improves with clip size, up to a point.

Clip size (# frames)	5	10	15	20	25	30
DeeperForensics	88.2	95.0	96.1	96.4	97.1	**97.4**
FaceShifter	87.9	93.4	95.4	95.7	**97.1**	96.7

Table 10. **Effect of clip size.** Accuracy (%) as a function of the clip size. We test on FaceShifter and DeeperForensics after training on FaceForensis++. Best results are in **bold**.

Backbone	FSh	DFo
CSN	**97.1**	**97.1**
ResNet+MS-TCN	94.0	95.7

Table 11. **Backbones.** Accuracy scores (%) on FaceShifter (FSh) and DeeperForensics (DFo) after training on FaceForensics++. We show results for two different backbones. Best results are in **bold**. Default setting is highlighted.

Different backbone. Our default video backbone is a CSN network [101]. In Table 11 we also show generalisation results for ResNet+MS-TCN [78], used in [53]. We significantly outperform LipForensics with the same backbone and auxiliary dataset (compare with Table 3 in the main text), without requiring any auxiliary labels.

Projector and predictor. We propose in the main text to use a single linear layer as our projector and a shallow transformer as the predictor. In Table 12, we show generalisation results when using different types of projectors/predictors. Since we output dense representations, the linear layers in the MLPs can be thought of as convolutional layers with kernel size 1. We use a learning rate of 3×10^{-4} when employing MLP predictors, as we found it to perform best in that setting.

Notably, we observe that using a transformer improves results over the MLP variant. This suggests that allowing the predictor to model temporal dynamics can benefit representation learning for our task. Further, in Table 13 we show results for a 1-block and a 2-block transformer predictor. We find that the 1-block variant performs slightly better.

Different contrastive baselines. As mentioned in the main text, self-supervised methods that aim to learn representations for lipreading tend to contrast samples from the same video to achieve invariance to identity [4, 29, 30]. Here, instead of our proposed non-contrastive approach, we apply the strategy of the audiovisual method Perfect Match [30] for stage 1 of our method. For fair comparison, we use the same backbones as for RealForensics. We follow the instructions from the paper for implementation. In particular, the inputs to the video and audio backbones

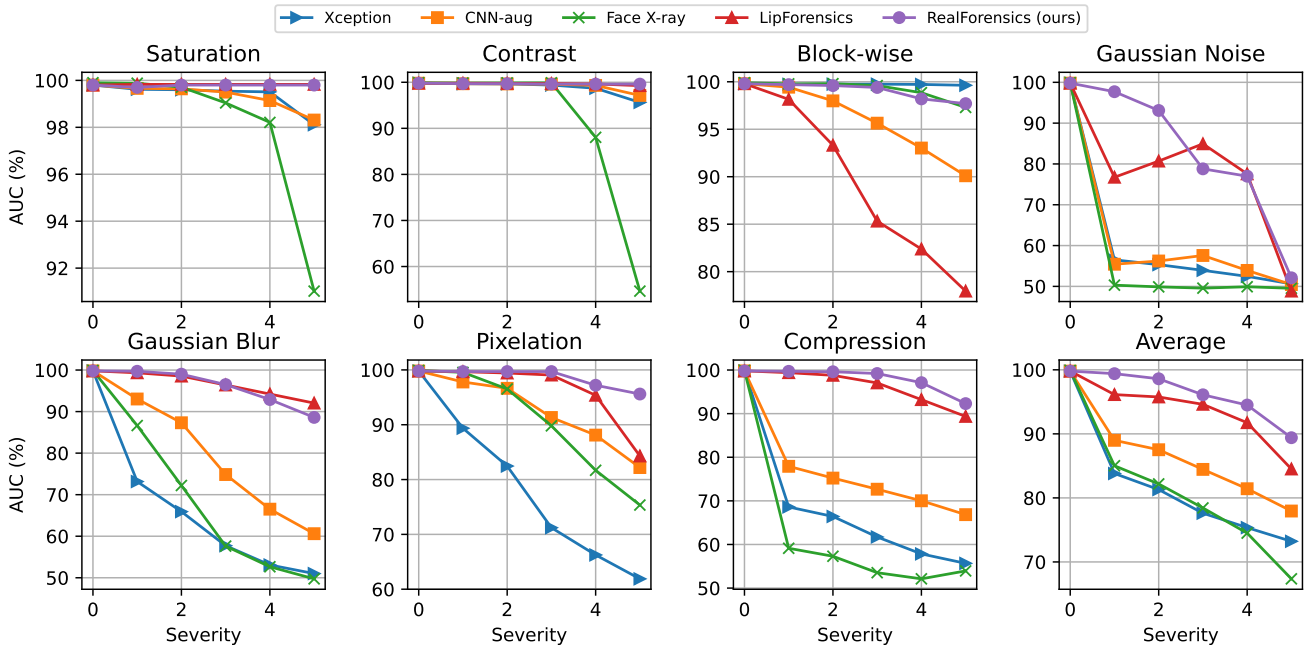

Figure 7. **Robustness to unseen perturbations.** AUC scores (%) on FaceForensics++ samples which have been corrupted by various unseen perturbations of varying severity. We also present the average scores across all perturbations. All methods were trained on FF++ without these corruptions. To avoid visual clutter in the plots, we show results for five representative methods. For more results, see [53] and [114].

Settings		Accuracy (%)	
Projector	Predictor	FSh	DFo
Linear	MLP	91.8	92.9
Linear	Transformer	**97.1**	97.1
MLP	MLP	91.1	92.5
MLP	Transformer	96.1	**97.5**

Table 12. **Projector and predictor.** We test different types of projectors and predictors for the representation learning stage of our method (stage 1), and see how generalisation to FaceShifter (FSh) and DeeperForensics (DFo) is affected after training on FaceForensics++. Refer to subsection "Projector and predictor" for a discussion. Best results are in **bold**. Default setting is highlighted.

# blocks	FSh	DFo
1	**97.1**	**97.1**
2	96.8	96.4

Table 13. **Number of transformer blocks.** Accuracy scores (%) on FaceShifter (FSh) and DeeperForensics (DFo) after training on FaceForensics++. We show results for a 1-block and a 2-block transformer predictor. Best results are in **bold**. Default setting is highlighted.

Method	FSh	DFo
PMatch	91.4	87.9
PMatch++	91.8	90.2
RealForensics (ours)	**97.1**	**97.1**

Table 14. **Different contrastive baselines.** Accuracy scores (%) on FaceShifter (FSh) and DeeperForensics (DFo) after training on FaceForensics++. We show results by employing the learning strategy of Perfect Match [30] (PMatch) for stage 1 of our method. We also use a symmetrised version of Perfect Match, which we call PMatch++. Best results are in **bold**.

are 5-frame video clips and 20-frame log mel spectrograms. Each network yields a single feature (via a temporal pooling layer). Then, for a single video feature, a contrastive loss is employed to match it to its aligned audio feature while repelling misaligned ones from the same video. We found that symmetrising this loss by additionally adding the loss corresponding to the reversal of the roles of the video and audio features yielded improvements; we refer to this variant as Perfect Match++. The results in Table 14 suggest that our proposed method, which does not target identity invariance, is better suited for forgery detection.

Visual-only representation learning. Although it is natu-

Type	FSh	DFo
Visual	92.9	89.7
Audiovisual	**97.1**	**97.1**

Table 15. **Visual versus audiovisual representation learning.** Accuracy scores (%) on FaceShifter (FSh) and DeeperForensics (DFo) after training on FaceForensics++. We compare visual-only with audiovisual representation learning (using BYOL-style training) for stage 1 of our method. Best results are in **bold**. Default setting is highlighted.

ral to use the correspondence between the visual and auditory modalities to capture information related to facial behaviour and appearance, we present here some preliminary results on using only the visual modality in the representation learning stage. To this end, we extend BYOL to the video setting by using a single student-teacher pair. As is the case for the cross-modal task, the network outputs temporally dense representations, and we use a transformer for the predictor. We apply the augmentations proposed in [48] to each frame, consistently across the whole video. The results in Table 15 indicate that our proposed cross-modal task strongly benefits generalisation, likely because audiovisual correspondence provides a richer signal for encoding natural facial movements and expressions. We leave for future work the investigation of more effective video augmentations that could further improve the visual-only baseline.

B. Further Implementation Details

B.1. Preprocessing

We use RetinaFace [36][3] for face detection and a 2-D FAN network [14][4] to extract 68 facial landmarks. For each frame, we take the mean landmarks around a 12-frame window to reduce motion jitter and then affine warp to LRW's mean face based on eight stable points.

B.2. Dataset details

We provide further details on the used datasets. The licenses of all datasets permit their use for research purposes.

FaceForensics++ [94] (FF++). We use the dataset from the official webpage[5]. We use the provided train/validation/test splits, which include 720 training, 140 validation, and 140 test videos, respectively.

FaceShifter [67]. We use the dataset (at compression c23) from the FF++ webpage. Its real videos come from FF++. Note that we do not treat FaceShifter as part of FF++, consistent with the original paper [94].

[3] https://github.com/biubug6/Pytorch_Retinaface
[4] https://github.com/1adrianb/face-alignment
[5] https://github.com/ondyari/FaceForensics

DeeperForensics [62]. We use the dataset from the official webpage[6]. Its real videos also come from FF++ c23.

CelebFD-v2 [72]. We use the dataset from the official webpage[7].

DFDC [37]. We use a subset of the dataset from the official webpage[8]. This subset was used in [53] and features single-subject videos for which the face and landmark detectors did not fail (since many videos have been subjected to extreme perturbations).

B.3. Architecture and training details

Supervised loss details. As described in Section 3.2 of the main text, we use a cosine classifier for our supervised head and also employ logit adjustment [80] to address data imbalance. Given the (average-pooled) output e of the backbone network and the weight vector w of the supervised head's linear layer, the normalised score of a sample's "fakeness" *during training* is given as

$$p = \frac{1}{1 + e^{-\left(s \frac{w \cdot e}{\|w\|_2 \|e\|_2} + \log \frac{\pi}{1-\pi}\right)}}, \tag{5}$$

where $s = 64$ scales the cosine similarity, as in e.g., [103], and π is the prior probability of a sample being fake, as described in [80]. We set π to be the ratio of fake samples to the batch size. We found using cosine similarity (*i.e.*, normalising the feature and weight vectors) yielded slight improvements; the ablation on logit adjustment is given in Table 5 of the main text. The supervised loss $\mathcal{L}_s(\mathcal{D}; \theta_b, \theta_s)$, introduced in Section 3.2 of the main text, is simply the standard binary cross entropy acting on these scores.

Random masking. We apply random erasing to video frames with probability 0.5, scale of $(0.02, 0.33)$, and ratio of $(0.3, 3.3)$. Moreover, we randomly erase a random number of video frames, ranging from 0 to 12, a random number of audio frames, ranging from 0 to 48, and a random number of mel filters, ranging from 0 to 27. This is applied with probability 0.5.

Backbones. Our video backbone is a modified Channel-Separated Convolutional Network (CSN) [101], chosen for its high accuracy in video action recognition [101] in conjunction with its relatively low parameter count. Unlike the original architecture, we set the temporal strides to 1 for all layers, thus preserving the temporal dimension. See Table 16 for more information.

[6] https://github.com/EndlessSora/DeeperForensics-1.0/tree/master/perturbation
[7] https://github.com/yuezunli/celeb-deepfakeforensics
[8] https://ai.facebook.com/datasets/dfdc

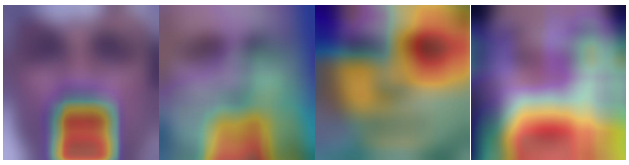

Figure 8. **Occlusion sensitivity analysis.** Occlusion sensitivity examples for FaceForensics++ types. The faces have been blurred to preserve anonymity.

Our audio backbone is a ResNet18 [55]. We modify the temporal strides to match the output size of the video backbone. In particular, the stem subsamples the temporal dimension by 4, after which no further temporal subsampling is performed. See Table 17 for more information.

Details on MLPs used in ablations. In the ablations where we use MLPs for the projector and/or predictor, we follow the design proposed in [22], as we found it to perform well. Thus, the projector MLP has 3 layers with hidden dimension 2048, and each layer is followed by batch normalisation (BN); the output layer has no ReLU activation. The predictor MLP has 2 layers with hidden dimension 512 and output dimension 2048, and the output layer has no BN nor ReLU.

Further details on contrastive experiments. We provide more details on the experiments with contrastive learning given in Table 6 of the main text. For dense representation learning, the output of the network consists of 25 embeddings (one for each video frame); we select a random embedding to add to the queue of negative samples. We also use shuffling batch normalisation to prevent the network from cheating on the pretext task [54].

C. Visualisation

We use occlusion sensitivity analysis [110] for visualisation, as in [53]. We systematically occlude, in a sliding-window fashion, parts of the video via random erasing of size $40 \times 40 \times T$ (where T is the number of frames). We record for each occluded pixel the effect that the occlusion has on the model predictions. A heatmap is produced by averaging the output probabilities for each pixel. After normalisation, we overlay the heatmap on the first video frame. We show examples for FaceForensics++ in Figure 8. We see that for NeuralTextures and Face2Face (first two examples), which modify expressions, our network usually focuses on the mouth region. On the face-swapping types, we observe that sometimes the network focuses on the mouth and sometimes on other facial regions.

stage	filters	output size
conv_1	$3 \times 7 \times 7$, stride $1 \times 2 \times 2$	$25 \times 56 \times 56$
pool_1	max, $1 \times 3 \times 3$, stride $1 \times 2 \times 2$	$25 \times 28 \times 28$
res_1	$\begin{bmatrix} 1 \times 1 \times 1, 256 \\ 3 \times 3 \times 3, 64 \\ 1 \times 1 \times 1, 256 \end{bmatrix} \times 3$	$25 \times 28 \times 28$
res_2	$\begin{bmatrix} 1 \times 1 \times 1, 512 \\ 3 \times 3 \times 3, 128 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 4$	$25 \times 14 \times 14$
res_3	$\begin{bmatrix} 1 \times 1 \times 1, 1024 \\ 3 \times 3 \times 3, 256 \\ 1 \times 1 \times 1, 1024 \end{bmatrix} \times 23$	$25 \times 7 \times 7$
res_4	$\begin{bmatrix} 1 \times 1 \times 1, 2048 \\ 3 \times 3 \times 3, 512 \\ 1 \times 1 \times 1, 2048 \end{bmatrix} \times 3$	$25 \times 4 \times 4$
pool_2	global spatial average pool	$25 \times 1 \times 1$

Table 16. **Video backbone architecture.** The architecture of the modified CSN [101] network that we employ for the video backbone. The layers in the bottleneck blocks, shown in brackets, use *depthwise convolutions*. Next to the brackets we give the number of times the blocks are repeated in each stage. The output size is of the form $T \times H \times W$, where T denotes time, H height, and W width. Note that differently from the original architecture [101], we do not subsample the temporal dimension at any stage and also only use spatial pooling at the end, rather than spatio-temporal, since we employ dense learning.

stage	filters	output size
conv_1	7×7, stride 2×2	50×40
pool_1	max, 3×3, stride 2×2	25×20
res_1	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	25×20
res_2	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	25×10
res_3	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	25×5
res_4	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	25×3
pool_2	global frequency average pool	25×1

Table 17. **Audio backbone architecture.** The architecture of our modified ResNet18 [55] network that we employ for the audio backbone. The layers in a residual blocks are in brackets, next to which we give the number of times the blocks are repeated in each stage. The output size is of the form $T \times F$, where T denotes time and F mel filters. Note that differently from the original architecture [55], we do not subsample the temporal dimension at any stage and also only use mel frequency pooling at the end, since we employ dense learning.