Video Forensics by Audio-Visual Anomaly Detection

Chao Feng   Ziyang Chen   Andrew Owens

University of Michigan, Ann Arbor

Abstract

We present a video forensics method that works by spotting inconsistencies in the video's audio and visual streams. We train an anomaly detection model on a feature set that captures the temporal alignment between video frames and sound. At test time, we then flag videos with unusual sequences of features. We show that we can obtain several effective feature sets for this problem, using networks that perform self-supervised audio-visual synchronization. We also show that the anomaly detection problem can be posed as finding low-probability sequences of these features using an autoregressive model. Despite being trained entirely on real videos, our model successfully detects speech videos that have undergone a variety of visual and audio manipulations. It also obtains strong generalization ability, such as to unseen spoken languages and to video frames that have been edited with common postprocessing operations.

1. Introduction

Supervised learning underlies today's most successful methods for media forensics. Yet these approaches' reliance on large collections of manipulated examples limits generalization, since it is extremely challenging to construct datasets that capture all of the types of manipulation that one might encounter in the wild. A longstanding goal in the forensics community is to design methods that, instead, learn to detect manipulations using cues discovered by analyzing large amounts of *real* data through self-supervision [26, 46].

We propose a method that identifies manipulated video through *audio-visual anomaly detection*. Our model learns the synchronization between the visual and audio streams of a video, solely from training on real video. Then, at test time, it then flags videos whose audio-visual signals are unlikely to co-occur.

One might expect that this problem could be posed as simply detecting out-of-sync examples, such as by finding cases in which a speaker's mouth does not open precisely at the onset of a spoken word. Unfortunately, however, videos
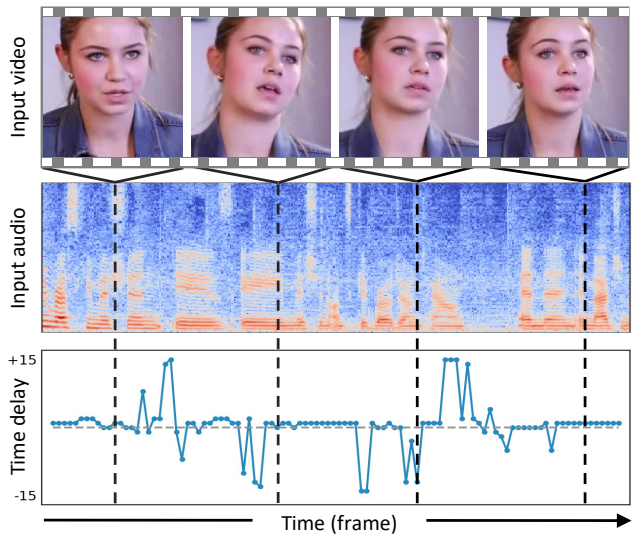

Figure 1. **Audio-visual anomaly detection.** We detect fake video by spotting anomalies in audio-visual synchronization. For example, in one variation, we estimate the temporal misalignment between each video frame and the audio stream, producing a sequence of time delays. Our model learns the distribution of these sequences, and flags videos with unusual sequences.

in the wild are often "naturally" misaligned due to errors in encoding or recording [2, 22].

Instead, we pose the problem as performing anomaly detection on *synchronization features*, representations that convey the temporal alignment between vision and sound. We evaluate several such features, each extracted from a model that has been trained to align audio and visual streams of a video [18]. In Figure 1, we show one such feature set: the amount of time that each video frame appears to be offset from its corresponding sound. To detect anomalies, we fit an autoregressive generative model [82, 97] to sequences of synchronization features extracted from real videos, and flag videos that it assigns low probability.

A key advantage of formulation is that it does not require any manipulated examples for training. It also does not require the speakers in the test set also be present at training time. This is contrast to previous audio-visual forensics ap-

proaches, which either require finetuning on datasets of manipulated video [39], or which are based on verifying that the speaker's voice matches previously seen examples [24].

We evaluate our model on videos that have manipulated a person's speech and face, using datasets of lip-synced and audio-driven face reenactment videos. Our model obtains strong performance on the FakeAVCeleb [52] and KoDF [59] dataset, despite the fact that it is trained entirely on real examples obtained from other video datasets. Our model generalizes to other spoken languages without retraining and obtains robustness to a variety of postprocessing operations, such as compression and blurring. We show through our experiments that:

- Video forensics can be posed as an audio-visual anomaly detection problem.
- Synchronization features convey information about video manipulations.
- Our model can successfully detect fake videos, while training solely on real videos.
- We generalize to many types of image postprocessing operations and to speech videos from spoken languages not observed during training.

2. Related Work

Audio-visual forensics. In early work, Malik and Farid [69] detected audio manipulations by finding inconsistencies in reverberation. Recent work has focused on detecting manipulated speech videos using audio-visual inconsistencies. Several approaches have directly trained audio-visual networks through supervised learning, using labels indicating whether a video is manipulated [19, 71]. A variety of methods have recently used audio-visual self-supervision for pretraining supervised models, which are finetuned with "real or fake" labels. Zeng *et al.* [103] used local and global contrastive learning methods to learn video features. Haliassos *et al.* [39] jointly solved a negative-less contrastive learning problem [37] and a forensics task. Zhou and Lim [110] used audio-visual synchronization to obtain pretraining, and proposed a dataset for audio-visual deepfake detection[1]. Other work [40] pretrains using lip-reading data. In contrast to these methods, our approach is trained entirely using *real* data and does not require any labels or examples of fake videos. Other work has used speaker verification [24] and phoneme-viseme mismatches [7] to detect fake videos and it also detects face swap manipulations, which preserve the synchronization between modalities. In contrast, our approach detects misaligned images and sounds and does not require that examples from the speaker be present in the training set.

Audio-visual representation learning. A variety of methods have been proposed to learn audio-visual representation from videos via self-supervision. Researchers have leveraged the natural semantic correspondence in the videos between frames and audio tracks [10, 73, 103] to learn multi-modal features and applied them to downstream tasks such as sound localization [9, 44, 72]. Other work studies temporal synchronization between audio and visual signals to learn audio-visual features [22, 57, 76], which can be used for active speaker detection [8, 56, 93], source separation [35, 68, 108], lip reading [5, 67, 70] and so on. Our method uses the off-the-shelf audio-visual synchronization model to perform anomaly detection.

Visual face forensics. A major focus of the forensics field has been on the problem of detecting manipulated videos of human faces. In recent years, a variety of visual face manipulation datasets are proposed, such as FaceForensics++ [85], VideoForensicsHQ [33] and $FFIW_{10K}$ [109]. Meanwhile, many methods are proposed to detect synthetic contents to fight against their potential threats. Some work [12, 38, 61] has proposed to use hand-crafted features to capture inconsistent visual or JPEG artifacts. Other work has proposed to use deep learning to inspect specific artifacts, such as blending [60], frequency domain [30, 34, 81], or texture [65]. A variety of methods have studied the generalization between detection classifiers [16, 98].

Anomaly detection. A variety of methods have learned a distribution, then flagged unusual examples. These examples are often considered anomalies [64, 89, 102, 111] or outliers [79, 87], and are used as part of open-set recognition [55, 104]. We formulate video forensics as the task of detecting anomalies, using a feature set that conveys information that would be hard for a forger to create. There have been a variety of methods proposed for learning this distribution, such as GAN discriminators [36, 55, 64, 79, 87, 89, 102, 111], flow-based models [104], and autoregressive models [91]. Similarly, our model is based on an autoregressive generative model [11, 82], since they have achieved strong performance at modeling complex distributions. Other work addresses goals similar to anomaly detection by creating methods that model uncertainty [62] or that perform outlier exposure [29, 43, 86]. Some work [13, 25, 27, 46, 49, 53, 77] has used special-purpose anomaly detection methods for image/video forensics. Other work [32, 45, 99, 106] uses supervised learning to find anomalous patterns. In contrast, our method builds the likelihood function entirely on real videos and views low-probability examples as fake.

3. Method

We formulate the problem of detecting manipulated videos as an anomaly detection problem. We model the distribution of audio-visual examples, then flag examples that have low probability. If we were to fit a model on the raw

---

[1]Their dataset is not publicly available.

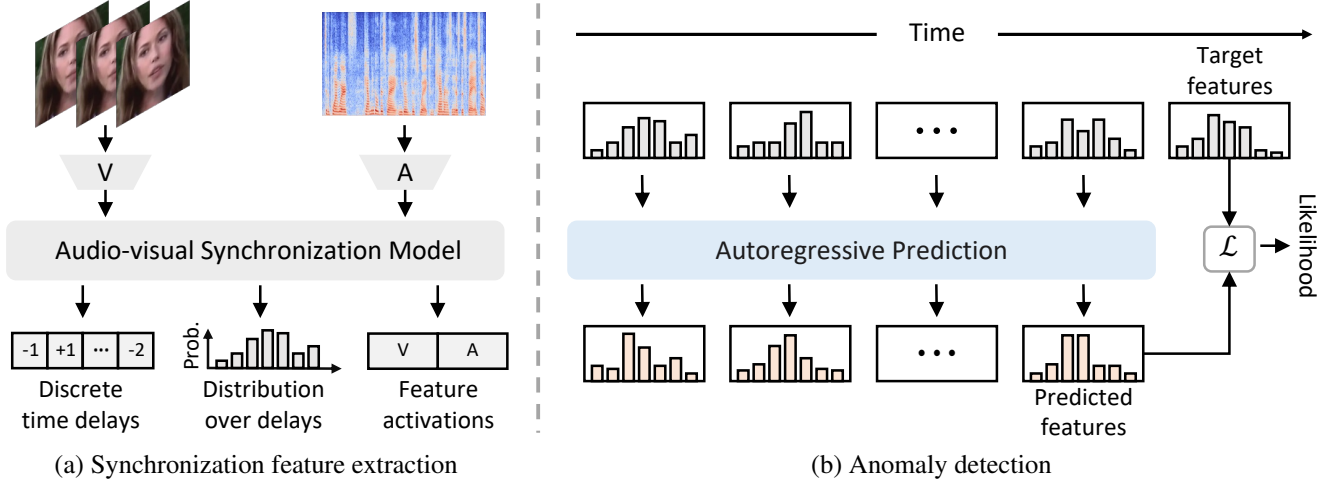|	
(a) Synchronization feature extraction	(b) Anomaly detection

Figure 2. **Audio-visual anomaly detection model.** (a) We extract a feature set from an audio-visual synchronization network: the number of frames of delay between video frames and sound, the distribution over delays at each frame, and feature activations from the audio and visual subnetworks. (b) We train an autoregressive Transformer model to assign probabilities to synchronization features. At test time, we flag low probability examples.

data, then this would be a very challenging learning task. Instead, we learn the distribution over a feature set that conveys subtle properties that are unlikely to be accurately captured in manipulated video.

3.1. Estimating audio-visual synchronization

We obtain our feautres from a network that performs audio-visual synchronization [18, 22, 23, 76]. We use the model of Chen *et al.* [18]. We learn a function $\phi(V_i, A_j)$ that indicates how likely video clip V_i temporally co-occurs with audio clip A_j. We estimate the synchronization score $S(i, j)$ of all audio-visual pairs in a temporal window as in Eq. (1):

$$S(i, j) = \frac{\exp\left(\phi(V_i, A_j)\right)}{\sum_{k=i-\tau}^{i+\tau} \exp\left(\phi(V_i, A_k)\right)}, \quad (1)$$

where τ is maximum time difference between two streams, and $\phi(V_i, A_j) = h\left(g_v(V_i), g_a(A_j)\right)$ is calculated using late fusion by a visual encoder g_v, audio encoder g_a and the fusion module h. We also interpret $S(i, j)$ as synchronization probability. We maximize the synchronization of true audio-visual pairs (V_i, A_i) using the InfoNCE loss [75]:

$$\mathcal{L}_{sync} = -\frac{1}{T} \sum_{i=1}^{T} \log S(i, i), \quad (2)$$

for a video of length T. We provide details about the architecture and training procedure in Appendix D.

After training, we can use the learned model to obtain a feature set for anomaly detection. For example, we can use the rows of S, which provide a probability distribution over possible alignments between video clips and audio clips.

3.2. Audio-visual anomaly detection

We use our learned model to obtain a feature set for anomaly detection. We learn the distribution of these features on a training set of real videos. Then at test time, videos with low probability will be flagged as potential fakes. We now explore two key design decisions that go into such a system: what feature set to use, and how the distribution is learned.

Given features for each frame, we learn a distribution $p_\theta(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$. We generally use autoregressive models to learn this distribution, given their success in modeling complex distributions [14, 101]. These models take the form:

$$p_\theta(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N) = \prod_{i=0}^{N-1} p_\theta(\mathbf{x}_{i+1} | \mathbf{x}_1, \cdots, \mathbf{x}_i). \quad (3)$$

We train a model $\hat{\mathbf{x}}_{i+1} = f_\theta(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i)$ that estimates the features of the next frame, given all of the features from the previous frames. Maximizing the log probability can be posed as minimizing a per-frame loss, L:

$$\mathcal{L} = \sum_{i=1}^{N} L(\hat{\mathbf{x}}_i, \mathbf{x}_i). \quad (4)$$

We now describe different formulations of the loss function L, the feature representation $\mathbf{x}_i$. In each case, we implement f_θ as a Transformer [97].

Discrete time delays. We first consider a simple model that uses discrete *time delay* estimates as our feature representation, following the success of autoregressive models for fitting discrete data [31, 84, 95]. For every video

3

frame, we estimate how far ahead (or behind) it appears to be from the audio signal. For each frame, we set $\mathbf{x}_i$ to be the time delay with the highest probability, *i.e.*, $x_i = \arg\max_j(S(i,j))$. We then set L to be cross entropy loss between the ground truth and predicted time delay. This amounts to solving a categorization problem with $2\tau + 1$ possible labels for each frame.

Distributions over delays. While discrete time delays are straightforward to represent in the model, they discard important information, such as when there is ambiguity in the delay. We therefore propose a model that directly predicts the entries of the time delay distribution. We set the features $\mathbf{x}_i$ to be the rows of S, i.e. the probability of each possible delay, and use cross entropy loss:

$$L(\hat{\mathbf{x}}_i, \mathbf{x}_i) = -\sum_{j}^{2\tau+1} x_{i,j} \log(\hat{x}_{i,j}). \quad (5)$$

We constrain the predictions made by our model f_θ to sum to 1, via a softmax.

Audio-visual network activations. The feature activations within the audio-visual synchronization network convey information about the time delay. We, therefore, ask whether these activations can be directly used as features for anomaly detection. We concatenate the representations of the visual and audio subnetworks, g_v and g_a. To provide a straightforward comparison with the time delay distribution model, we reduce the dimensionality of the features by projecting them onto the top $2\tau + 1$ principal components, following other work in autoregressive models of features [83]. We use squared distance as our loss: $L(\hat{\mathbf{x}}_i, \mathbf{x}_i) = \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2$.

4. Results

We evaluate the different variations of our model on a variety of video forensics tasks.

4.1. Implementation details

Synchronization model. Following Chen *et al*. [18], we use ResNet-18 2D+3D [41, 42] as the visual encoder, using 5 frames frames (25 fps) as input. The audio encoder uses VGG-M [17] and extracts features from 0.2s audio clips (16kHz). We fuse audio and visual data using a Transformer that has 3 standard Transformer encoder blocks [97], 4 attention heads, and 512 channels. We train using the cropped faces provided by each dataset. Please see Appendix D for details.

Anomaly detection model. We use a decoder-only autoregressive Transformer [31, 63, 82] to learn the distribution over synchronization features. We use 2 decoder blocks [97], each with 16 attention heads with 256 channels.

For models that use time delay, we set the maximum delay to be $\tau = 15$ frames, resulting in the distribution $S_i \in \mathbb{R}^{31}$ for each video frame. We use sequences of length $N = 50$ from 2.0s video.

Hyperparameters. All videos are resampled to 25 fps and 16kHz mono audio. We represent audio segments as the mel spectrogram of size 21×80 via Short-Term Fourier Transform (STFT) with 80 mel filter banks, the hop length of 160, and the window size of 320. Please see more details in Appendix D.

4.2. Dataset

We train our model on real, unlabeled speech video, and evaluate it on forensics datasets.

Training datasets. We train our models on Lip Reading Sentences 2 (LRS2, 97k videos) [3] and Lip Reading Sentences 3 (LRS3, 120k videos) [4]. The videos in each contain tightly cropped face tracks. We divide each dataset into 3 splits and train the audio-visual synchronization model and the autoregressive model on different splits.

Evaluation datasets. We evaluate on two video forensics datasets, spanning several different types of manipulations that change the speech and face of a human speaker. **FakeAVCeleb** [52], which is derived from Vox-Celeb2 [20]. This dataset contains 500 real videos and 19,500 fake videos manipulated by Faceswap [58], FS-GAN [74], and Wav2Lip [80], and fake sounds that are generated by SV2TTS [47]. The examples in the dataset contain different combinations of these manipulations. We use the dataset's provided face crops. We sample 2400 videos (400 real videos and 2000 fake videos) as train/val splits and 600 videos (100 real videos and 500 fake videos) as test split. We note that our method does not use any videos from train/val splits, since it is trained from another dataset (LRS2 or LRS3). Second, we evaluate on **KoDF** [59], a large-scale Korean-language deepfake detection dataset. It contains 62,166 real videos and 175,776 fake videos, where fake videos are generated by 6 synthesized methods: FaceSwap [1], FSGAN [74], DeepFaceLab [78], FOMM [90], ATFHP [100] and Wav2Lip [80]. We extract faces by using face detection [105] and alignment [15].

4.3. Evaluation methods

Following common practice [16, 39, 40, 52, 59, 60, 81, 85, 98, 107], we evaluate using average precision (AP) and AUC. These evaluation metrics are widely used for cross-dataset generalization and unsupervised models since they avoid the need to threshold the predictions. We compare our approach to both supervised and self-supervised methods. Unless otherwise stated, we use time delay distributions as our feature set (Sec. 3.2).

4

	Method	Modality	Pretrained dataset	Category												
				RVFA		FVRA-WL		FVFA-FS		FVFA-GAN		FVFA-WL		AVG-FV	
				AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
Supervised	Xception [85]	$\mathcal{V}$	ImageNet [28]	–	–	88.2	88.3	92.3	93.5	67.6	68.5	91.0	91.0	84.8	85.3
	LipForensics [40]	$\mathcal{V}$	LRW [21]	–	–	**97.8**	**97.7**	99.9	99.9	61.5	68.1	98.6	98.7	89.4	91.1
	AD DFD [110]	$\mathcal{AV}$	Kinetics [51]	**74.9**	**73.3**	97.0	97.4	99.6	99.7	58.4	55.4	**100.**	**100.**	88.8	88.1
	FTCN [107]	$\mathcal{V}$	–	–	–	96.2	97.4	**100.**	**100.**	77.4	78.3	95.6	96.5	92.3	93.1
	RealForensics [39]	$\mathcal{V}$	LRW [21]	–	–	88.8	93.0	99.3	99.1	**99.8**	**99.8**	93.4	96.7	**95.3**	**97.1**
Unsupervised	AVBYOL [37, 39]	$\mathcal{AV}$	LRW [21]	50.0	50.0	73.4	61.3	88.7	80.8	60.2	33.8	73.2	61.0	73.9	59.2
	VQ-GAN [31]	$\mathcal{V}$	FFHQ [50]	–	–	52.7	53.8	49.0	51.2	63.9	61.1	52.0	54.4	54.4	55.1
	Ours	$\mathcal{AV}$	LRS2 [3]	62.4	71.6	**93.6**	**93.7**	95.3	95.8	94.1	94.3	93.8	94.1	94.2	94.5
	Ours	$\mathcal{AV}$	LRS3 [4]	**70.7**	**80.5**	91.1	93.0	91.0	92.3	91.6	92.7	91.4	93.1	91.3	92.8

Table 1. **Manipulation detection on FakeAVCeleb.** We report AP scores (%) and AUC scores (%), following the evaluation protocol of Haliassos et al. [40], in which supervised methods are evaluated on unseen manipulation types (unsupervised methods are not trained with labels). We report results with combinations of real/fake video/audio, using different generation algorithms. We report the average performance over four fake video (FV) categories in AVG-FV. We retrained all models, since there is no standard dataset split.

Supervised methods. For supervised methods, we train several state-of-the-art detectors on the two datasets: 1) **Xception** [85]: a popular baseline for forensics detection; 2) **LipForensics** [40]: a detector is built on high-level semantic embeddings of mouth and targets irregularities in mouth movements; 3) **AD DFD** [110]: a multimodal detector with audio and video branches, utilizes audio-visual synchronization signal implicitly for detection; 4) **FTCN** [107]: a video forensics detector leverages temporal incoherence to boost generalization capability; 5) **RealForensics** [39]: it first pretrains the network by audio-visual BYOL [37] framework and then finetunes the pretrained model and forensics datasets by multi-task learning to obtain robust and general face forgery detection. Since there is no standard split for our forensics tasks, we retrain each model.

Self-supervised methods. Since we are not aware of any existing methods that consider self-supervised speech video forensics, we adapt two existing methods to the task. First, we consider an audio-visual contrastive learning model, which we call **AVBYOL**, that learns to determine whether the visual and audio streams of a video do (or do not) match, an approach that has been used as a part of other audio-visual forensics models [24, 39]. We adapt the model of Haliassos *et al.* [39], which uses BYOL [37] to learn a joint audio-visual embedding for pretraining. Instead of pretraining, we directly use the model's audio-visual similarity score to flag fake examples. Second, we use an off-the-shelf generative model, **VQGAN** [31], for anomaly detection. VQGAN converts an image into a sequence of discrete codes, then uses an autoregressive Transformer to learn the distribution of codes. We use the code's log likelihood, averaged over each video frame, for anomaly detection.

4.4. Evaluation

In real-world scenarios, the deployed detectors are expected to recognize fake videos manipulated by unseen techniques. Thus, following the standard procedure used in [39, 40, 107, 110], we conduct the experiment to evaluate the cross-manipulation generalization ability of our model on the FakeAVCeleb dataset [52] which videos are manipulated in various ways. Since our approach and other self-supervised baselines learn from real, unmanipulated videos and perform zero-shot fake video detection, all the fake videos during evaluation are considered as manipulated by unseen methods.

We split FakeAVCeleb dataset [52] into five categories based on the manipulation methods and manipulated modalities: 1) **RVFA**: real video with fake audio by SV2TTS [47]; 2) **FVRA-WL**: real audio with fake video by Wav2Lip [80]; 3) **FVFA-WL**: fake video by Wav2Lip [80], and fake audio by SV2TTS [47]; 4) **FVFA-FS**: fake video by Faceswap [58] and Wav2Lip [80], and fake audio by SV2TTS [47]; 5) **FVFA-GAN**: fake video by FaceswapGAN [74] and Wav2Lip [80], and fake audio by SV2TTS [47]. For the supervised methods, we hold out the evaluated category and train the models on the four remaining categories. Note that some approaches are only able to detect the manipulation on a certain modality, we do not report their performance on the categories with the manipulation only on the other modalities.

We show our results in Tab. 1. Our method substantially outperforms both self-supervised methods AVBYOL [37, 39] and VQGAN [31] on each category by a large margin. More importantly, our method works on par with or outperforms some supervised methods on certain categories, especially FVFA-GAN, even though our method does not use any labeled supervision or fake examples.

Method		Modality	KoDF	
			AP	AUC
Supervised (transfer)	Xception [85]	$\mathcal{V}$	76.9	77.7
	LipForensics [40]	$\mathcal{V}$	89.5	86.6
	AD DFD [110]	$\mathcal{AV}$	79.6	82.1
	FTCN [107]	$\mathcal{V}$	66.8	68.1
	RealForensics [39]	$\mathcal{V}$	**95.7**	**93.6**
Unsupervised	AVBYOL [37, 39]	$\mathcal{AV}$	74.9	78.9
	VQ-GAN [31]	$\mathcal{V}$	49.0	50.0
	Ours	$\mathcal{AV}$	**87.6**	**86.9**

Table 2. **Generalization to Korean speech.** AP scores (%) and AUC scores (%) are reported on KoDF dataset [59]. Supervised methods are trained on FakeAVCeleb dataset [52] and transferred. Ours is trained on LRS2 [3]. Best results are in **bold**.

Moreover, our method has quite consistent performances and it can achieve AP over 90% on the most of categories. While Xception [85], LipForensics [40], AD DFD [110] and FTCN [107] work well on 75% of the settings, there are settings where performance collapses to near-chance (*e.g.*, AD DFD [110] on FVFA-GAN). Interestingly, the two self-supervised baselines struggle to detect fake videos, perhaps because both models do not necessarily capture the subtle information that would be needed to detect manipulations. In addition, VQGAN [31] compresses the visual signal using a codebook, which might drop the artifact clues and harm the detection performance. Moreover, our model trained on LRS2 [3] works on par with the one trained on LRS3 [4], indicating that our method's performance is not tied to a single training set.

Cross-dataset generalization. We also evaluate the generalization capability of our model by evaluating it on the KoDF dataset [59], following [39,40,107,110]. We focus on the audio-driven synthesis examples in the dataset, where videos are manipulated by ATFHP [100] or Wav2Lip [80]. We train the supervised models on FakeAVCeleb [52] to evaluate their generalization ability. Many of these training videos share the same technique used in KoDF for synthesis [80]. As the results are shown in Tab. 2, our approach obtains a comparable performance to many supervised methods. Although our system is trained on the English speech datasets, it still generalizes to KoDF [59] dataset of Korean speech, perhaps because it learns low-level lip motion cues that are broadly useful.

Qualitative results. We visualize the ground truth and predicted time delay distributions generated by our autoregressive continuous time delay model (Sec. 3.2) in Fig. 3. We use the four main manipulation categories from FakeAVCeleb dataset [52]. For each one, we display a heat map indicating the predicted time delay, using a model that obtains the ground truth delays from the previous frames as
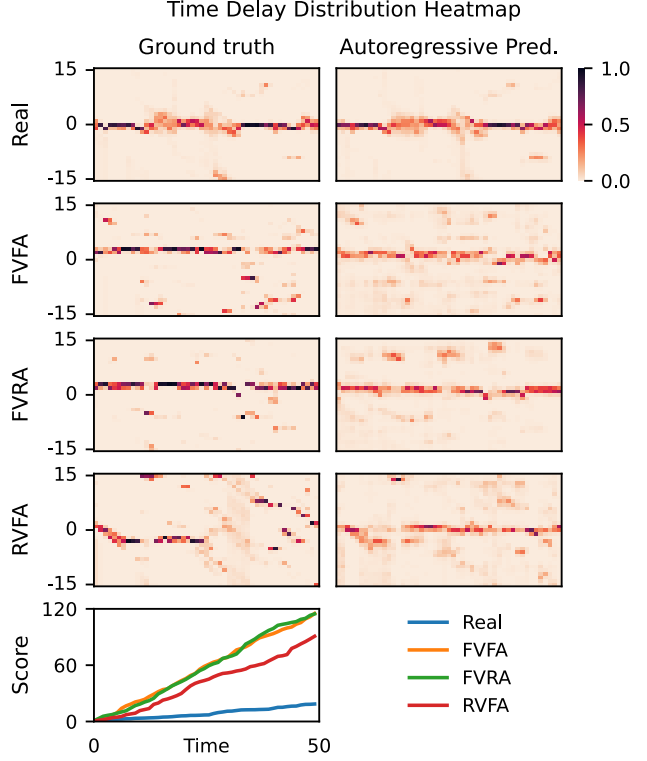

Figure 3. **Time delay predictions for real vs. fake examples.** We visualize the time delay distributions from the synchronization model and predicted results generated by the autoregressive model for four random samples from different categories. Synchronization probabilities are in a range from 0 to 1. We show the predictions of the autoregressive model when feeding it ground truth observations for the previous timesteps. We show cumulative prediction error (indicating the probability of being fake) for each sample over time steps in the last row.

input. We also plot the cumulative prediction loss (Eq. 4) over time. From Fig. 3, we can see that our autoregressive model accurately predicts the ground truth for real video, which results in a lower score. For fake videos, we can find clear differences between ground truth and predicted time delay distribution, leading to higher prediction loss.

4.5. Robustness to unseen perturbations

When the fake video is redistributed, it may undergo many types of postprocessing that result in corruption, making detection more difficult. Thus, it is important for forensics models to be robust to the types of postprocessing operations they may encounter in the wild. Following [39, 40, 107], we use the set of visual perturbations proposed in [48]: 1) Color saturation change; 2) Block-wise distortion; 3) Color contrast change; 4) Gaussian blur; 5) Gaussian noise; 6) JPEG compression; 7) Video compression rate change. We set the intensity levels from 0 to 5
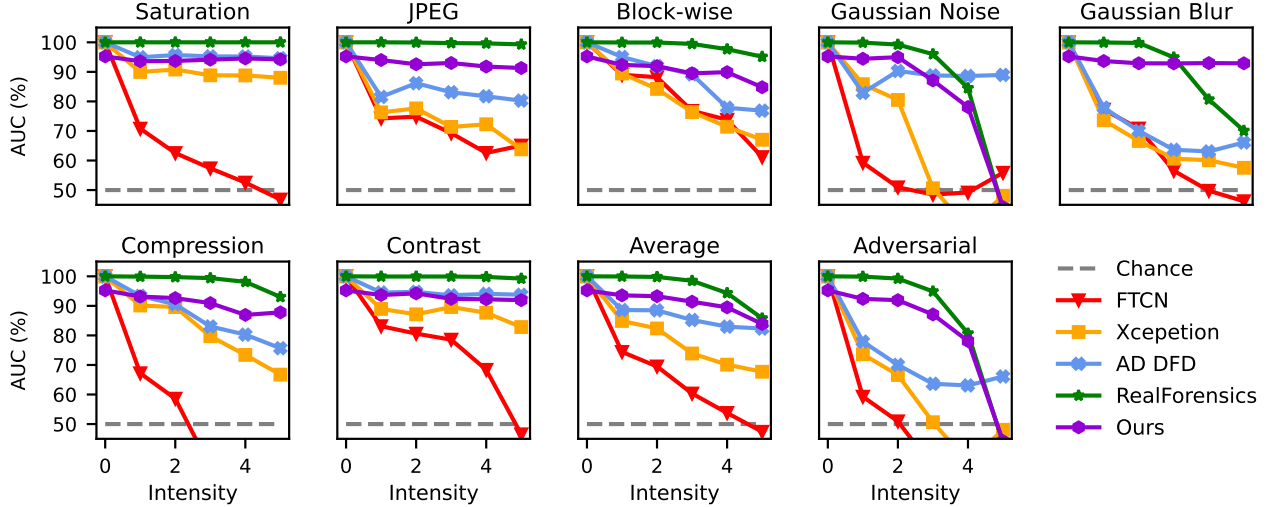
Figure 4. **Robustness to unseen perturbations.** AUC scores (%) of different detectors as a function of perturbation intensities. There are 6 intensity levels in total from [48]. "Average" represents the average over 7 perturbations under each intensity. "Adversarial" means we pick the worst performance across 7 perturbations under each intensity.

for each perturbation.

We compare our model with four supervised methods XceptionNet [85] FTCN [110], AD DFD [110] and RealForensics [39]. The results are reported in Fig. 4. Our proposed self-supervised model is overall more robust to unseen visual perturbations on average compared with these supervised methods, with the exception of RealForensics [39]. This is also true when we consider "worst case" performance, by taking the minimum performance over all types of augmentation of a given intensity level. Interestingly, we obtain this performance even though our model is trained in a very different way from other works, suggesting that the feature set continues to convey useful information to the anomaly detection model, even in the presence of significant corruption.

4.6. Feature set analysis

We evaluate the effectiveness of different feature sets used by our anomaly detection model. As described in Sec. 3.2, we start with **discrete time delays** as our feature representation and optimize the model with cross entropy loss. Then, we use **continuous time delay distributions** for representations instead, where we optimize models with different objective functions: 1) Soft CE: we use the time delay distribution as target (akin to a "soft" label) and use cross entropy loss (Sec. 3.2); 2) CE: we map each distribution into one-hot encoding as target by using $\arg\max$ and employ cross entropy loss; 3) BCE: we use the distribution as the target while treating each synchronization score $S(i, j)$ (Sec. 3.1) within the same time step independently. We use the sigmoid function and binary cross entropy loss to train the model. We also use our network's **fea-**

Model	Feature set	$\mathcal{L}$	AVG-ALL		AVG-FV	
			AP	AUC	AP	AUC
Bayes	-	-	73.1	85.1	72.4	86.0
	discrete delay	CE	80.8	86.5	80.0	86.6
	distribution	CE	84.8	87.9	90.3	92.2
	distribution	BCE	78.6	83.4	80.5	84.8
Ours	distribution	Soft CE	**87.8**	**90.0**	**94.2**	**94.5**
	activation-AV	MSE	86.5	87.1	91.5	91.9
	activation-V	MSE	–	–	77.6	85.9
	discrete prob.	–	83.4	86.9	88.6	91.1

Table 3. **Feature set analysis.** AP (%) and AUC (%) are reported on FakeAVCeleb [52] when using different feature sets. Best results are in **bold**. AVG-ALL means the average over all categories. AVG-FV represents the average over four fake video categories.

ture activations as in Sec. 3.2: 1) audio-visual feature activations (activation-AV); 2) visual-only feature activations (activation-V).

We also compare with a simple model based on **Naive Bayes** and discrete time delays. This model assumes that each frame's time delay is independent, and obtains a probability for the entire sequence by multiplying the probability of each frame's time delay. This amounts to simply detecting large misalignments, since in practice the Naive Bayes model will assign probability solely based on the magnitude of each delay.

Finally, we consider a version of the model that autoregressively predicts the entire distribution of time delays, inspired by autoregressive models, such as PixelCNN [94] that generate images in a raster scan order. We autoregressively predict each element of the 2D matrix $\hat{S}(i, j)$, where $\hat{S}(i, j)$

Synchronization dataset	Auto-regression dataset	AVG	
		AP	AUC
LRS2 [3]	LRS2 [3]	**87.8**	90.0
	LRS3 [4]	85.0	89.6
LRS3 [4]	LRS2 [3]	86.6	89.0
	LRS3 [4]	87.2	**90.3**

Table 4. **Dataset ablation.** AP scores (%) and AUC scores (%) are reported on FakeAVCeleb [52] dataset by using different datasets to train synchronization model or atuoregressive model. Best results are in **bold**.

is created by vector quantizing the entries of the synchronization probability $S(i, j)$ using k-means (see Appendix C for details).

Analysis. We evaluate each variant on FakeAVCeleb [52] and report results in Tab. 3. These results suggest that all formulations achieve performance significantly better than chance, indicating that these feature sets are useful for anomaly detection. As in Tab. 3, the time delay distribution model outperforms the discrete time delay model, suggesting that there is important information conveyed in the probability of unlikely delays. The autoregressive model that uses distribution as input and soft labels (soft CE) performs best, since it forces the output prediction to match the distribution from the synchronization model. Interestingly, the model that uses audio-visual feature activations obtains performance close to that of the soft CE model, suggesting that the networks' audio-visual features convey useful information. Finally, the multimodal activation-AV model significantly outperforms the visual-only activation-V model, suggesting that having access to both modalities is useful for our anomaly detection model.

4.7. Ablation study

Different training dataset. We ask how the choice of dataset affects the quality of the model. To test this, we train our synchronization and autoregressive models on different datasets to analyze the generalization abilities of each component, *i.e.*, training the synchronization model on LRS2/LRS3 and training the autoregressive model on LRS3/LRS2 with the same hyperparameters. As shown in Tab. 4, there is no significant performance change when we train these two components on different combinations of dataset, including when they are trained on the same dataset. This suggests that the distribution of time delay predictions may be stable between these speech video datasets.

Influence of sequence length. To explore the influence of input sequence length for autoregressive model, we sample the same amount of training videos for sequence length N of 10, 20, 30, 40, 50, and 60, and keep other hyperpa-
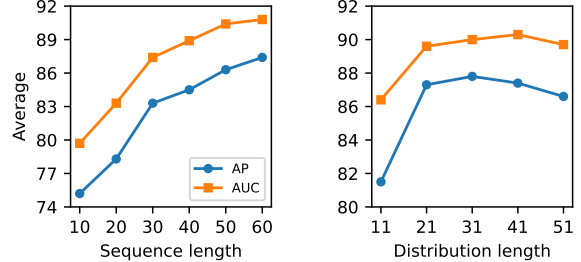

Figure 5. **Hyperparameter ablation.** We evaluate with different input sequence lengths for our autoregressive model on FakeAVCeleb (left), and study the effect of the time delay distribution's maximum temporal offset (right).

rameters the same. We test these models on FakeAVCeleb dataset [52]. Fig. 5 shows that as the sequence length increases, the performance increases with it.

Effect of time delay distribution maximum offset. We also study how the length of time delay distribution would affect the performance of the autoregressive model with distribution over delays. We experiment with maximum offset $\tau \in \{5, 10, 15, 20, 25\}$ resulting in the delay distribution length of $\{11, 21, 31, 41, 51\}$. We test these models on the FakeAVCeleb dataset [52]. Fig. 5 shows that as distribution length increases, the performance first increases, after which point results plateau or slightly decrease. This may be due to the fact that when considering larger ranges of offsets, the distribution is spread over a large number of unlikely possibilities, making important information less apparent after normalization.

5. Conclusion

We have proposed a method for detecting video manipulation by self-supervised anomaly detection. To do this, we create novel feature sets that convey audio-visual synchronization. We then show that fake videos can be detected by flagging examples with unlikely sequences of these features, according to a learned distribution. Our model obtains strong performance on the FakeAVCeleb and KoDF datasets, despite the fact that it was trained only on real video. It also obtains robustness to visual postprocessing operations and to videos containing other spoken languages. We see our work as opening in two directions. The first is in posing forensics as an anomaly detection problem with a self-supervised feature set. While we have proposed one such model, based on autoregressive sequence models, the field of anomaly detection offers many possible future approaches. The second direction is in developing new feature sets that are well-suited to forensics problems, beyond the synchronization features used in this work. We will release code, models, and dataset splits upon acceptance.

Limitations and Broader Impacts. Our work provides methods that can potentially be applied to detecting malicious video manipulations and disinformation. While we have shown that our model is capable of detecting several types of fake video, there may be other techniques that our model fails to detect. In particular, due to the design of our use of synchronization-based features, our model is not well suited to detecting manipulations that leave the synchronization between motion and sound relatively consistent, such as those that change a speaker's appearance without significantly changing the motion of their mouth.

References

[1] Faceswap. https://github.com/deepfakes/faceswap, 2022. 4

[2] Triantafyllos Afouras, Yuki M. Asano, Francois Fagan, Andrea Vedaldi, and Florian Metze. Self-supervised object detection from audio-visual correspondence, 2021. 1

[3] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. In *arXiv:1809.02108*, 2018. 4, 5, 6, 8

[4] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 4, 5, 6, 8

[5] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2143–2147. IEEE, 2020. 2

[6] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, 2020. 14

[7] 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. 2

[8] Juan León Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 265–274, 2021. 2

[9] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 2

[10] Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Advances in Neural Information Processing Systems*, 33:4660–4671, 2020. 2

[11] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000. 2

[12] Tiziano Bianchi and Alessandro Piva. Image forgery localization via block-grained analysis of jpeg artifacts. *IEEE Transactions on Information Forensics and Security*, 7(3):1003–1017, 2012. 2

[13] Luca Bondi, Silvia Lameri, David Guera, Paolo Bestagini, Edward J Delp, Stefano Tubaro, et al. Tampering detection and localization through clustering of camera-based cnn features. In *CVPR Workshops*, volume 2, 2017. 2

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[15] 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. 4

[16] 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. 2, 4

[17] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014. 4

[18] H Chen, W Xie, T Afouras, A Nagrani, A Vedaldi, and A Zisserman. Audio-visual synchronisation in the wild. In *Proceedings of the 32nd British Machine Vision Conference*. British Machine Vision Association, 2021. 1, 3, 4, 14

[19] 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. 2

[20] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *Proc. Interspeech 2018*, pages 1086–1090, 2018. 4

[21] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, 2016. 5

[22] 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. 1, 2, 3

[23] 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

[24] Davide Cozzolino, Matthias Nießner, and Luisa Verdoliva. Audio-visual person-of-interest deepfake detection. *arXiv preprint arXiv:2204.03083*, 2022. 2, 5

[25] Davide Cozzolino and Luisa Verdoliva. Single-image splicing localization through autoencoder-based anomaly detection. In *2016 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2016. 2

[26] Davide Cozzolino and Luisa Verdoliva. Noiseprint: a cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2019. 1

[27] Dario D'Avino, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Autoencoder with recurrent neural networks for video forgery detection. *arXiv preprint arXiv:1708.08754*, 2017. 2

[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[29] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[30] 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

[31] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3, 4, 5, 6

[32] Jianwei Fei, Yunshu Dai, Peipeng Yu, Tianrun Shen, Zhihua Xia, and Jian Weng. Learning second order local anomaly for general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20270–20280, 2022. 2

[33] Gereon Fox, Wentao Liu, Hyeongwoo Kim, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. Videoforensicshq: Detecting high-quality manipulated face videos. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 2

[34] 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

[35] Ruohan Gao and Kristen Grauman. Visualvoice: Audiovisual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500. IEEE, 2021. 2

[36] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 2

[37] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2, 5, 6

[38] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Eyes tell all: Irregular pupil shapes reveal gan-generated faces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2904–2908. IEEE, 2022. 2

[39] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14950–14962, 2022. 2, 4, 5, 6, 7

[40] 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. 2, 4, 5, 6

[41] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 4

[42] 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

[43] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 2

[44] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[45] Ziheng Hu, Hongtao Xie, Yuxin Wang, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Dynamic inconsistency-aware deepfake video detection. In *IJCAI*, 2021. 2

[46] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. *European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[47] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, 2018. 4, 5

[48] 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. 6, 7, 15

[49] Sri Kalyan Yarlagadda, David Güera, Paolo Bestagini, Fengqing Maggie Zhu, Stefano Tubaro, and Edward J Delp. Satellite image forgery detection and localization using gan and one-class classifier. *arXiv e-prints*, pages arXiv–1802, 2018. 2

[50] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5

[51] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The

kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[52] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. 2, 4, 5, 6, 7, 8, 14

[53] Hasam Khalid and Simon S Woo. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 656–657, 2020. 2

[54] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 14, 15

[55] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2021. 2

[56] Okan Köpüklü, Maja Taseska, and Gerhard Rigoll. How to design a three-stage architecture for audio-visual active speaker detection in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1193–1203, 2021. 2

[57] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018. 2, 14

[58] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685, 2017. 4, 5

[59] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10744–10753, 2021. 2, 4, 6

[60] 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. 2, 4

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

[62] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 2

[63] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*, 2018. 4

[64] Yezheng Liu, Zhe Li, Chong Zhou, Yuanchun Jiang, Jianshan Sun, Meng Wang, and Xiangnan He. Generative adversarial active learning for unsupervised outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1517–1528, 2019. 2

[65] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8060–8069, 2020. 2

[66] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 15

[67] Pingchuan Ma, Brais Martinez, Stavros Petridis, and Maja Pantic. Towards practical lipreading with distilled and efficient models. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7608–7612. IEEE, 2021. 2

[68] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2hear: Active audio-visual source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 275–285, 2021. 2

[69] Hafiz Malik and Hany Farid. Audio forensics from acoustic reverberation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010. 2

[70] 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. 2

[71] 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. 2

[72] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[73] 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. 2

[74] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019. 4, 5

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

[76] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. *European Conference on Computer Vision (ECCV)*, 2018. 2, 3

[77] Daniel Pérez-Cabo, David Jiménez-Cabello, Artur Costa-Pazo, and Roberto J López-Sastre. Deep anomaly detection for generalized face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[78] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, RP Luis, Jian Jiang, et al. Deepfacelab: A simple, flexible and extensible face swapping framework. 2020. 4

11

[79] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. *Advances in neural information processing systems*, 31, 2018. 2

[80] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV 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*, pages 484–492, 2020. 4, 5, 6

[81] 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. 2, 4

[82] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1, 2, 4

[83] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 4

[84] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 3

[85] 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, 4, 5, 6, 7

[86] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. *arXiv preprint arXiv:1906.02694*, 2019. 2

[87] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3379–3388, 2018. 2

[88] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017. 14

[89] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 2

[90] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[91] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, 2017. 2

[92] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 15

[93] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021. 2

[94] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 7

[95] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[96] Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. 14

[97] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 3, 4, 15

[98] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4

[99] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9543–9552, 2019. 2

[100] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020. 4, 6

[101] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 3

[102] Houssam Zenati, Manon Romain, Chuan-Sheng Foo, Bruno Lecouat, and Vijay Chandrasekhar. Adversarially learned anomaly detection. In *2018 IEEE International conference on data mining (ICDM)*, pages 727–736. IEEE, 2018. 2

[103] Zhaoyang Zeng, Daniel McDuff, Yale Song, et al. Contrastive learning of global and local video representations. *Advances in Neural Information Processing Systems*, 34:7025–7040, 2021. 2

[104] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In *European Conference on Computer Vision*, pages 102–117. Springer, 2020. 2

[105] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face

detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017. 4

[106] 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

[107] 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. 4, 5, 6

[108] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[109] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5778–5788, 2021. 2

[110] 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. 2, 5, 6, 7

[111] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018. 2

A. Video Results

We provide some qualitative video results of some random samples from the FakeAVCeleb dataset [52] in `visualization.mp4` with audio. We show "ground truth" outputs from the synchronization model and autoregressive predictions over time. We also show a score indicating the probability that the example is fake (Eq. 4, main paper). We use the time delay distribution as our feature set, and display the predictions as a heat map. Each step $\mathbf{x}_t$ of the predicted outputs were obtained by providing the ground truth features from times $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{t-1}$ into the autoregressive model.

B. Ablation Study

Feature activation. We study the influence of the number of principal components D for the variation of the anomaly detection model that projects the audio-visual feature activations into a lower dimensional space. We set D to 11, 31, 32, 64, 128, 256, 512 and keep other hyperparameters the same. Fig. 6 shows that as the D increases, the accuracy of the anomaly detection model decreases. The reason might be that the higher dimensional prediction problem is also more challenging.

C. Feature Set Variations

We describe more details about building the autoregressive model on some of our feature sets in this section.

Binary cross entropy (BCE) model. We assume that the $2\tau + 1$ possible time delays of each time step are independent, and use BCE loss for probability $S(i, j)$ of each time delay. Thus, we can decompose $p_\theta(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$ as in Eq. (6):

$$p_\theta(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N) = \prod_{i=0}^{N-1} \prod_{q=1}^{2\tau+1} p_\theta\left(x_{i+1,q}|\mathbf{x}_1, \cdots, \mathbf{x}_i\right),$$
(6)

We then maximize $p_\theta\left(x_{i+1,q}|\mathbf{x}_1, \cdots, \mathbf{x}_i\right)$ using BCE loss:

$$L(\hat{\mathbf{x}}_i, \mathbf{x}_i) = -\sum_{j=1}^{2\tau+1} x_{i,j} \log(\hat{x}_{i,j}) + (1-x_{i,j}) \log(1-\hat{x}_{i,j}).$$
(7)

We constrain the prediction $\hat{x}_{i,j}$ to the range of $[0, 1]$ via a sigmoid function.

Discrete 2D probability model. The *discrete prob.* model generates the entire 2D frame/time delay matrix autoregressively in "raster scan" order, similar to models such as PixelCNN [88, 96]. We unroll the time delay distribution sequence into 2D grid like a grayscale image as shown in Fig. 7. We use K-means ($K = 8$) clustering to quantize eacn entry and assume each probability $\hat{S}(i, j)$ of time
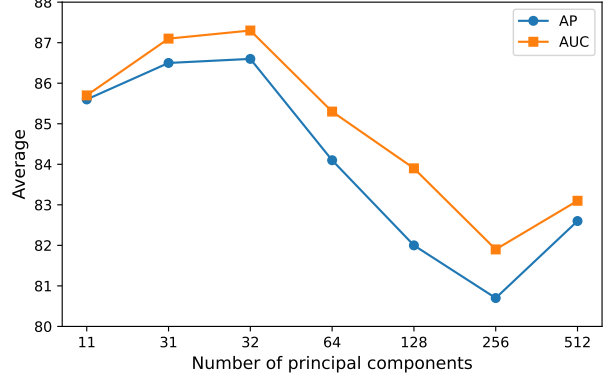

Figure 6. **Number of PCA projections.** We evaluate with different number of principal components for the model that obtains a feature set by projecting the audio-visual feature activations to a lower dimensional space using PCA. We report average results for the metrics in the FakeAVCeleb dataset [52].

delay not only depends on the previous time delay distributions but also the previous probabilities within the same distribution. Then we decompose $p_\theta(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$:

$$p_\theta(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N) = \prod_{i=0}^{N-1} \prod_{q=1}^{2\tau+1}$$
$$p_\theta(x_{i+1,q}|\mathbf{x}_1, \cdots, \mathbf{x}_i; x_{i+1,1}, \cdots, x_{i+1,q-1}).$$
(8)

We utilize similar loss function in PixelCNN [88, 96] to supervise the autoregressive model.

D. Implementation Details

Audio-visual synchronization model. Following prior work [6, 18, 57], we utilize curriculum learning to train the audio-visual synchronization model. Specifically, we use the two-stage training procedure. During the first phase, the negatives are from different videos, while for the second stage, the negatives are sampled within the same videos randomly (the starting time steps are sampled randomly).

Anomaly detection model. We process each the input vector $\mathbf{x}_i$ with an affine transformation, before passing it into the autoregressive model. This projects the input (e.g., a time delay distribution $S_i \in \mathbb{R}^{31}$) to $\mathbb{R}^{256}$. Also, we add an affine transformation for the to project the embedding back into the feature set space, e.g., $\mathbb{R}^{256} \rightarrow \mathbb{R}^{31}$ for the time delay distribution. We use a learnable positional encoding for both the synchronization and autoregressive models.

Hyperparameters. For the synchronization model, we use Adam [54] with a learning rate of 1×10^{-4}, with a batch size of 16 for the first phase and 40 for the second. During the second stage, we sample four 5-frame short clips
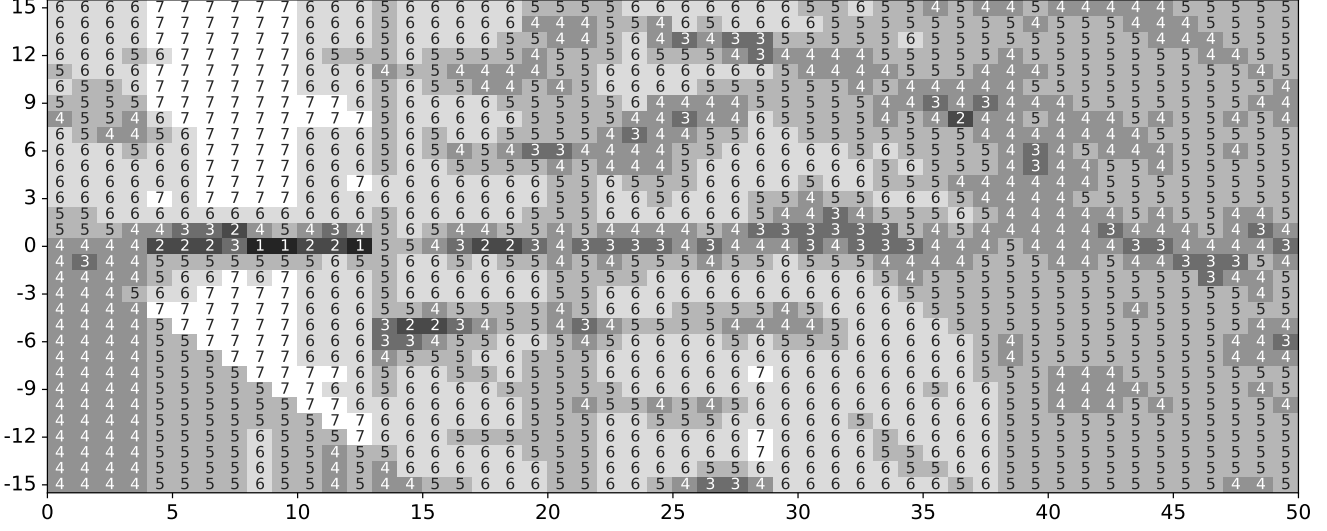
Figure 7. **Visualization of discrete probability grid.** We use K-means clustering to quantize the probability space to convert continuous synchronization probability $S(i,j)$ to discrete probability bin $\hat{S}(i,j)$. Then we build a autoregressive Transformer [97] model on the probability grid.
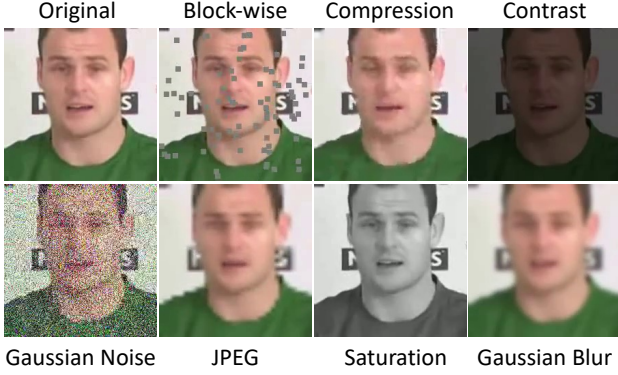

Figure 8. **Visualization of corrupted images.** Examples of the corruptions taken into account at intensity level 5. The set of corruptions is introduced in Jiang et al. [48], and it consists of color saturation, local block-wise distortion, color contrast, Gaussian blur, white Gaussian noise, JPEG compression, and video compression rate change.

per video. For the autoregressive model (anomaly detection model), we use the the Adam optimizer [54] with 1×10^{-3} learning rate, weight decay of 1×10^{-6} and warm-up and cosine learning rate decay [66] strategies. We use batch size of 16 and the dropout [92] rate of 0.1. We train the synchronization model on 8 NVIDIA A40 GPUs, and use a single GeForce RTX 2080 Ti GPU for the autoregressive model.

E. Visualization of perturbed images

We visualize some images used for unseen perturbations robustness test in Fig. 8.