Revisiting Generalizability in Deepfake Detection: Improving Metrics and Stabilizing Transfer

Sarthak Kamat
University of California, Berkeley
sartk@berkeley.edu

Shruti Agarwal
Adobe Research
shragarw@adobe.com

Trevor Darrell
University of California, Berkeley
trevor@eecs.berkeley.edu

Anna Rohrbach
University of California, Berkeley
anna.rohrbach@berkeley.edu

Abstract

"Generalizability" is seen as the hallmark quality of a good deepfake detection model. However, standard out-of-domain evaluation datasets are very similar in form to the training data and lag behind the advancements in modern synthesis methods, making them highly insufficient metrics for robustness. We extend the study of transfer performance of three state-of-the-art methods (that use spatial, temporal, and lip-reading features respectively) on four newer fake types released within the last year. Depending on the artifact modes they were trained on, detection methods fail in different scenarios. On diffusion fakes, the aforementioned methods get 96%, 75%, and 51% AUC respectively, whereas on talking-head fakes, the same methods get 80%, 99%, and 92% AUC. We compare various methods of combining spatial and temporal modalities through joint training and feature fusion in order to stabilize generalization performance.

We also propose a new, randomized algorithm to synthesize videos that emulate diverse, visually apparent artifacts with implausibilities in human facial-structure. By testing deepfake detectors on highly randomized artifacts, we can measure the level to which detection networks have learned a strong model for "reality", as opposed to memorizing subtle artifact patterns.

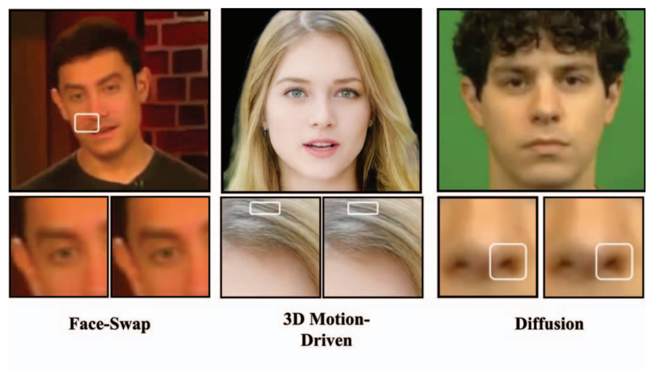

Figure 1. **Artifacts can manifest very differently depending on the type of deepfake.** Notice that the Face-Swap fake (left) from [27] has a discolored spot on the right cheek, which can be detected with a typical spatial classifier. While the frame-level quality of the 3D-motion driven fake [8] is high, temporal detection methods with strong attention mechanisms between adjacent frames should be able to pick up the sudden change in shape of the highlighted hair strand. The diffusion synthesized video [41] has an unnatural blotting of the shadow under the left nostril across two frames.

1. Introduction

Deepfakes are artificially generated videos of humans, created using deep neural network-based generators. While these videos are often visually impressive, their potential for deception and misuse is significant [53]. The threat of deep-fake videos has prompted significant research into techniques for detection, which often involves training binary classifiers to detect manipulated image sequences on a large set of deep-fakes and measuring their performance on various benchmarks. While one can trivially achieve good results on the training set using out-of-the-box image and video classification models (such as XceptionNets, ResNets, and EfficientNets) [38], achieving good generalization when evaluating on deep-fakes and other types of manipulation that are outside the training domain is considerably more challenging [20, 23]. To overcome this, recent methods introduce inductive biases by restricting the training architecture to features along a single modality [60, 13] or patch-wise constraints [10] to prevent the network from picking up on easy-to-detect artifacts that are specific to the deep-fakes in the training dataset. This is a counter-intuitive result: one would imagine that accessing multiple cues and artifact modes would yield superior generalization capabilities; yet, leading detection methods only generalize when they abandon certain detection modes entirely [60, 40]. When artifacts get sparse, overly relying on hand-picked cues has its costs. Newer fakes, for example, can be produced by neurally rendering the image from scratch, rather than modifying a target video on a pixel-level using face-swap and warping, without the distributional inconsistencies between the inside and the outside of the face that existing detectors rely on [57] [5].

We conduct evaluations of state-of-the-art detection methods on four such newer, neural rendering deep-fake creation methods [41, 8, 11, 54] along with the common high quality benchmark in Celeb-DF [28], and find that methods can perform well on one fake type and fail dramatically on another. We showcase the variance in artifact modes using anecdotal examples from our proposed evaluation suite in Figure 1. We explore the complementarity of the spatial and temporal modalities as a means to stabilize generalization performance across a variety of fakes in Section 5. Importantly, we note that since prior works have not handcrafted their methods with the fakes from our evaluation-suite in mind, the results in our paper act as a legitimate *test-set* record of how state-of-the-art methods generalize to unseen manipulation types [41, 8, 11, 54].

Still, as such creation methods get incorporated into out-of-domain evaluation datasets, it is possible to design newer inductive biases that can saturate performance on these benchmarks too. Towards this, we propose the use of a *simulated generalizability evaluation* (*SGE*), where we simulate spatial and temporal deepfake artifacts in videos of human faces with a Markov process. We argue that a sufficiently generalizable detection method should be able to identify these artifacts, since they reflect implausibilities in facial structures that can accompany unseen manipulation types. Our design of *SGE*s is modeled off prior work [40, 23] that use synthetic data to train deepfake detectors, but with modifications to produce richer temporal artifacts and spatial localization.

2. Related Work: Deepfake Synthesis

2.1. Facial Retargetting

Synthesizing deepfakes has historically been formulated as a face retargetting problem. Early methods used alpha and poisson blending schemes to replace the pixels in the target image [49]. Eventually, to deal with pose and posture in the target video, Thies *et al*. adopted explicit 3D models or texture maps to estimate facial motion from RGB video and transfer it to a target face [45, 44]. The landscape for facial retargetting changed dramatically with generative models such as GANs that reapplied similar ideas from previous facial retargetting and face-swap methods using an adversarial classifier for hyperrealism [21] [34] [32]. Li *et al*. developed *FaceShifter* to incorporate the crucial lighting details of the scene into facial retargetting to enhance the realism of the rendered videos [22]. *Wav2Lip* changes the formulation of the facial retargetting, conditioning on an input audio sequence, rather than facial movement, to get around structural differences between the source and target [35].

2.2. Neural Rendering

Recently, neural rendering-based deepfakes have become a popular way to generate deepfakes with fewer artifacts. While these methods are still conditioned on pose or input audio, they do not modify a target video sequence to produce the deepfake. These methods use generative models and volumetric representations to render the video sequence from scratch. Recently, Shen *et al*. and Stypulkowski *et al*. use denoising diffusion models to auto-regressively sample frames of the target speaker conditioned on source audio and a single image [39, 41]. Other methods use normalizing-flows [47], and GANs to do motion-transfer, also using a single target image [54, 46]. *MegaPortraits* uses super-resolution networks to generate particularly high resolution talking heads [8].

Recent work has shown that hybrid graphics pipelines too, potentially including the use of 3D models, also allow for more realistic facial expressions and movements [43]. Ji *et al*. and Gurunani *et al*. decompose the audio-conditioned generation problem into predicting facial landmark trajectories, and then rendering them with either a neural transfer or computer graphics model [18] [11]. Liu *et al*. [30] use Neural Radiance Fields to train implicit representation networks for each scene. Due to their novelty, neural rendering methods have not been represented in deepfake detection datasets. This can be a problem in measuring generalizability of detection methods, since their artifact modes diverge significantly from facial-retargetting methods.

3. Related Work: Deep-fake Detection

3.1. Detection: Early Approaches

While supervised face-forgery detection methods initially focused on relatively shallow convolutional neural networks [2, 1], works such as [12, 33, 56, 31, 38, 58, 61, 33, 37, 50] found success training unconstrained deep end-to-end networks that implicitly learn to detect low-level textural artifacts. These methods have now been shown to be highly unstable, with dramatic drops in performance on unseen fake types [4], video compression and perturbations [13], as well as adversarial attacks [17].

3.2. Detection Methods Today

Recent work has stepped away from training models end-to-end without constraints on their representational capacity. High-level semantic methods have shown superior generalization abilities compared to low-level techniques, with some of the best performance on highly compressed videos. They focus on specific features such as blinking and head pose [55, 19, 25], biological and neural patterns [51, 16], and the readability of lip-movements (*Lip-Forensics*) [13]. Among these, *LipForensics* achieves strong generalization performance and is considered a benchmark for measuring face-forgery detection methods. They use a frozen lipreading network as an encoder, which they feed into a temporal classification head. Prashnani *et al*. use the same architecture, but replace RGB lip-region inputs with hand-crafted frequency-domain features [36]. Zheng *et al*. [60] observe that temporal inconsistencies in generators are more transferable between manipulation-types than spatial artifacts, and modify the convolutional kernel size of 3D ResNet-50 to 1x1 along the spatial axes. Another recent method, Guan *et al*. [10], uses a modified version of a vision transformer [7], analyzing temporal inconsistencies along independent 16x16 patch sequences.

3.3. Training on Synthetic Data

In parallel, there has been an effort to increase the diversity of the training set by augmenting or replacing the training-set fakes entirely with fake emulation schemes. Li *et al*. [26] reproduces face warping artifacts that appear on GAN-fakes, whereas [24, 59] focus on source-target blending region artifacts associated most commonly with face-swaps. [40] modifies these algorithms to produce self-blended images that use a single-image as both the source and the target. It outperforms state-of-the-art supervised methods on uncompressed fakes and cross-dataset generalization. This method, however, is purely image-based and could suffer dramatically when conditions around blending-boundaries in manipulations reduce (See Figure 5). Further, it does not consider temporal or multi-modal inconsistencies that can be valuable signal for even better generalization.

4. Improving Generalizability Metrics

4.1. Methods being Evaluated

We evaluate the following state-of-the-art methods with publicly available code-bases:

1. **Temporal**: Zheng *et al*. [60] use a 3D Res-Net 50 [14], and modify it to reduce the spatial kernel to a 1x1. This is then received by multi-layer transformer network [48] with the class token as proposed in [6]. The class token is then linearly projected to predict the final logit.

2. **Self-Blended Images**: Shiohara *et al*. train an EfficientNet-b4 [42] on a synthetic image blending scheme using Sharpness-Aware Minimization, a second-order optimization method [40].

3. **Lip-Forensics** Haliassos *et al*. [13] use a frozen ResNet-18 [15] pre-trained on lip-reading, and feed it to a MS-TCN based temporal classification head [9]. This is designed to capture unnatural movement in the lip-region.

4.2. Higher Quality Deepfakes

Typical generalizability paradigms involve training on four fakes types in the FaceForensics++ [38], and are then measured on standard datasets such as DFDC [3], FFIW [62], or Celeb-DFv2[29]. However, solely using these datasets to measure out of domain robustness is not sufficient since they are predominantly comprised of facial-retargetting deepfakes which are similar in domain to the training fakes, even if they come from different generators.

Further, since these datasets have existed for a while, one runs the risk of "overfitting to the metric". It is possible that the handcrafted methods that measure to be generalizable are relying on a single artifact mode that is shared between the the training domain and out-of-domain datasets. To test this hypothesis, we re-evaluate transfer accuracy by complementing the Celeb-DFv2 dataset with four other unseen manipulation types that are released within the last year.

We detail the manipulation types below:

1. **Face-Swap (2020)**

   We use 340 samples from test set of Celeb-DFv2 [27] to evaluate *generalizability* on Face-Swap fakes. The fakes are generated using an undisclosed Deep-Fake algorithm with additional post-processing to remove otherwise clearly visible color and frequency related artifacts. The results from our re-evaluation are on Figure 2.

2. **MegaPortraits (2022)**

   We use 48 samples from the test-set output of the method from [8]. This method transfers the expression from the source video onto a target image. To encode the appearance of the target frame [8] predict volumetric features, a global descriptor, and an appearance encoder. In parallel, they predict the motion representations from the driving video, including head motions and latent descriptors. This in turn outputs a 3D warping operations that map the current expression to a canonical space, and then re-warp it into the target expression. Figure 3 plots the logit distribution of state-of-the-art methods on this synthesis method.

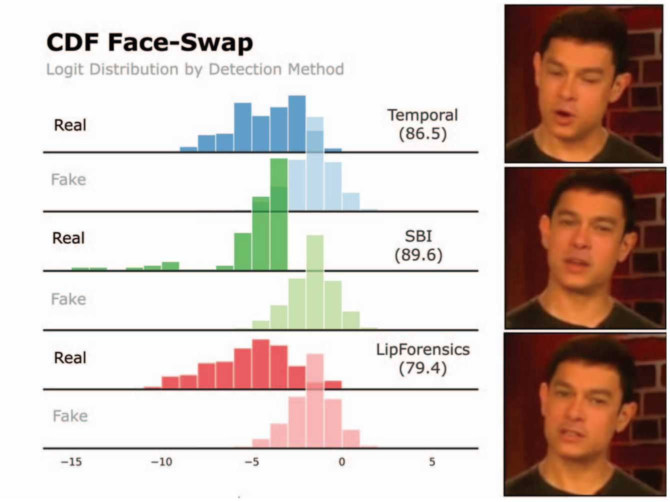

Figure 2. **Logit Distribution on Face-Swap Fakes**: We use a fixed bin width histogram to plot the predicted logit distribution ($\ln \frac{P_{\text{fake}}}{1-P_{\text{fake}}}$) using the models from [60], [40], and [13] and the fake-set from [29]. AUC scores are parenthesized for reference.

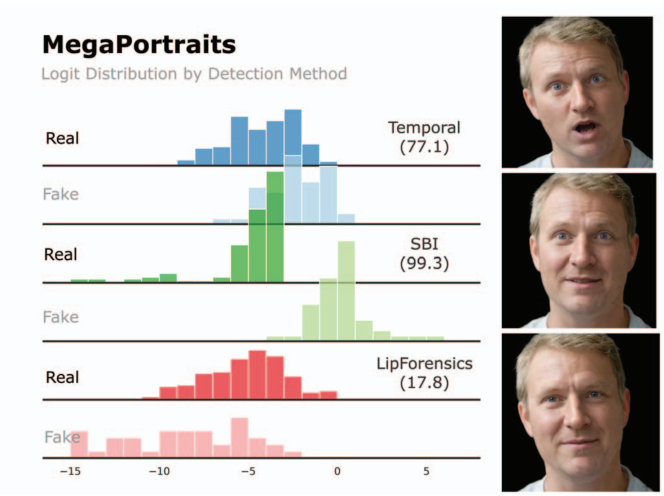

Figure 3. **Logit Distribution on Mega-Portrait Fakes**: We use a fixed bin width histogram to plot the predicted logit distribution ($\ln \frac{P_{\text{fake}}}{1-P_{\text{fake}}}$) using the models from [60], [40], and [13] and the fake-set from [8]. AUC scores are parenthesized for reference.

3. **Diffusion (2023)**

   Diffusion methods originally popularized for text-to-image-synthesis have been successfully repurposed for talking head generation. Stypulkowski *et al*. uses an auto-regressive diffusion model that samples frames conditioned on input audio and an image of the target speaker [41]. The results from evaluating on 820 test-set outputs are in Figure 4.

4. **Speech Conditioned Face-Vid2Vid (2022)**

   Face-Vid2Vid was originally proposed by [52], where they synthesize a talking-head video using the target

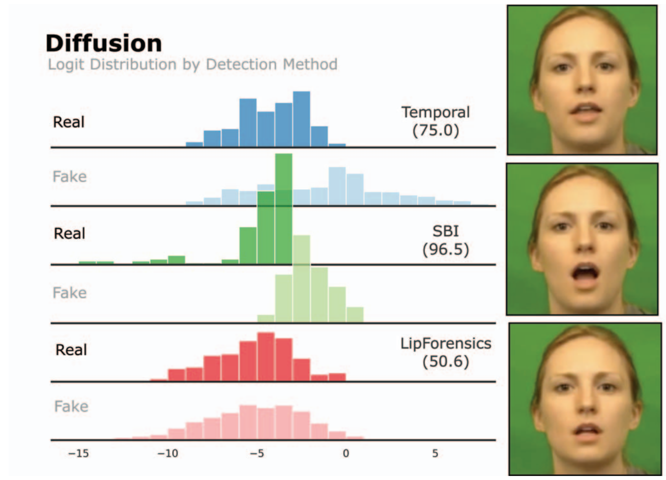

Figure 4. **Logit Distribution on Diffusion Fakes**: We use a fixed bin width histogram to plot the predicted logit distribution ($\ln \frac{P_{\text{fake}}}{1-P_{\text{fake}}}$) using the models from [60], [40], and [13] and the fake-set from [41]. AUC scores are parenthesized for reference.

person's appearance and a driving video. Gurunani *et al*. extend this [11] to be speech conditioned by predicting target landmarks using an LSTM before *Face-Vid2Vid* renders them from projected latents. We evaluate on 100 sample videos (see Figure 5).

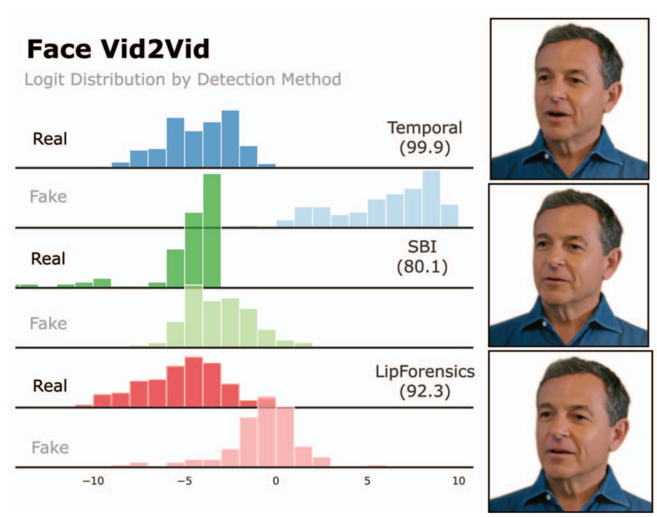

Figure 5. **Logit Distribution on Face-Vid2Vid Fakes**: We use a fixed bin width histogram to plot the predicted logit distribution ($\ln \frac{P_{\text{fake}}}{1-P_{\text{fake}}}$) using the models from [60], [40], and [13] and the fake-set from [11]. AUC scores are parenthesized for reference.

5. **AniFaceGAN (2022)**

   AniFaceGAN is an animatable 3D-aware GAN for multi-view consistent face animation generation [54]. They explicitly formulate deformation fields to synthesize an input image using driving facial motion.

We evaluate detection methods on 112 samples and plot logit histograms in Figure 6.

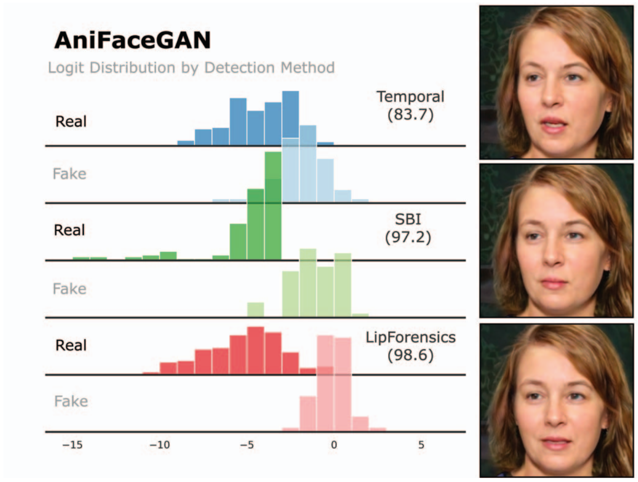

Figure 6. **Logit Distribution on AniFaceGAN Fakes**: We use a fixed bin width histogram to plot the predicted logit distribution ($\ln \frac{P_{\text{fake}}}{1-P_{\text{fake}}}$) using the models from [60], [40], and [13] and the fake-set from [54]. AUC scores are parenthesized for reference.

4.3. Simulated Generalizability Evaluation

We propose the use of a *simulated generalizability evaluation* (*SGE*), where we simulate spatial and temporal deepfake artifacts using a randomized algorithm. These artifacts reflect implausibilities in facial structures that can accompany unseen manipulation types can be detected with a human eye. We detail the process of generating *SGEs* below:

- Let M be a binary mask representing the sub-region of a face in a given frame of a video. To generate a synthetic training sample, we apply a random affine transform θ_M to M to obtain a distorted mask M'. Similarly, we apply a random affine transform θ_I to the entire frame I and blend the resulting image I' with the original image I using M' as the blending mask:

$$M' = \text{Distort} \circ \text{Affine}(M; \theta_M)$$
$$I' = (M' \odot \text{Distort} \circ \text{Affine}(I; \theta_I)) + ((1 - M') \odot I)$$

  where $\odot$ denotes element-wise multiplication.

- Note that, we also post-process masks and images with separate constant distortion operations to the mask as well as the images in the entire frame sequence. For the mask, this involves a combination of feathering of the borders and an elastic distortion, whereas for the image, only an elastic distortion is randomly used.

- To produce a synthetic video, we repeat this process for each subsequent frame of the video, with the twist that the parameters for the random affine transform are chosen to stay the same with probability p, or to be randomly re-selected with probability $(1 - p)$. Additionally, with probability q, we do not blend the next frame and instead use the actual image I_{t+1} as is. In other words, if $\theta_M^{(t)}$ and $\theta_I^{(t)}$ represent the affine transforms applied to the mask and image respectively, for the t-th frame, then we have:

$$\theta_M^{(t)} = \begin{cases} \theta_M^{(t-1)}, & \text{w.p. } p \\ \text{Rand}\theta_M(), & \text{w.p. } (1-p) \end{cases}$$

$$\theta_I^{(t)} = \begin{cases} \theta_I^{(t-1)}, & \text{w.p. } p \\ \text{Rand}\theta_I(), & \text{w.p. } (1-p) \end{cases}$$

- While such a set-up emulates sudden flickering artifacts well, it is not well suited to smoother changes over time. In order to reproduce those artifacts, we post-process with an additional randomization of sometimes linearly interpolating between affine transformation matrices. Figure 7 shows a sample rollout from our *SGE* method.

5. Stabilizing Generalizability Performance with Multi-modal Detection

5.1. Spatio-Temporal Detection

An observable outcome of the potpourri of hand-crafted training methods is that *generalizability* is not just a scalar metric. Depending on which set of manipulated fakes you evaluate on, vastly different detection methods outperform one another or fail inexplicably. We hypothesize that a single spatio-temporal architecture can achieve more uniform transfer performance since it has access to more artifact modes. However, this is not a trivial task, since large spatio-temporal networks broadly to tend to overfit on dataset-specific artifacts. We analyze if it is possible to get the best of both words with intermediate and late fusion schemes. We describe the architectures below, with illustrations in Figure 8.

1. **Joint Training**: We modify the architecture from [60], by restoring the spatial kernel size from the original 3D Res-Net architecture [14]. We then concatenate a learnable class token along the time axis and add

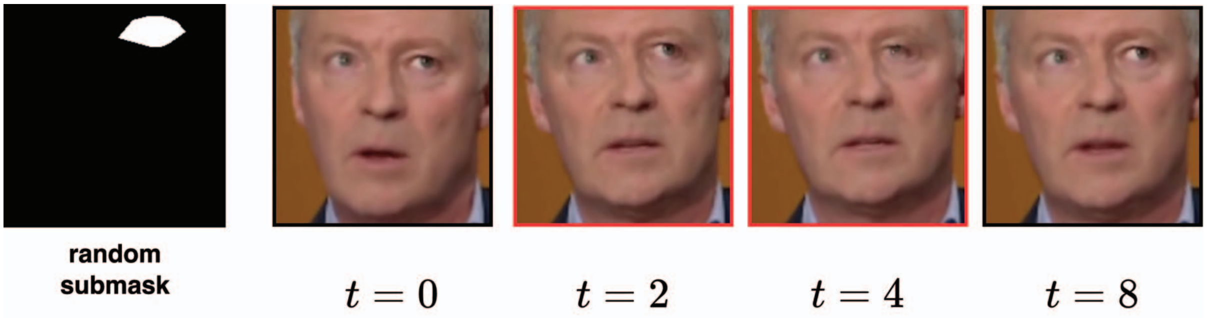

Figure 7. **A sample rollout from our *SGE* evaluation method.** The frames with the red border are manipulated. Notice that artifacts can range from subtle to highly visible: in the second frame there is a slight change in eye position in a way that is inconsistent with previous frames. A generalizable detector should be able to pick up on these randomized artifacts, since the glitch in the sub-mask could not have been plausible in a real human video.

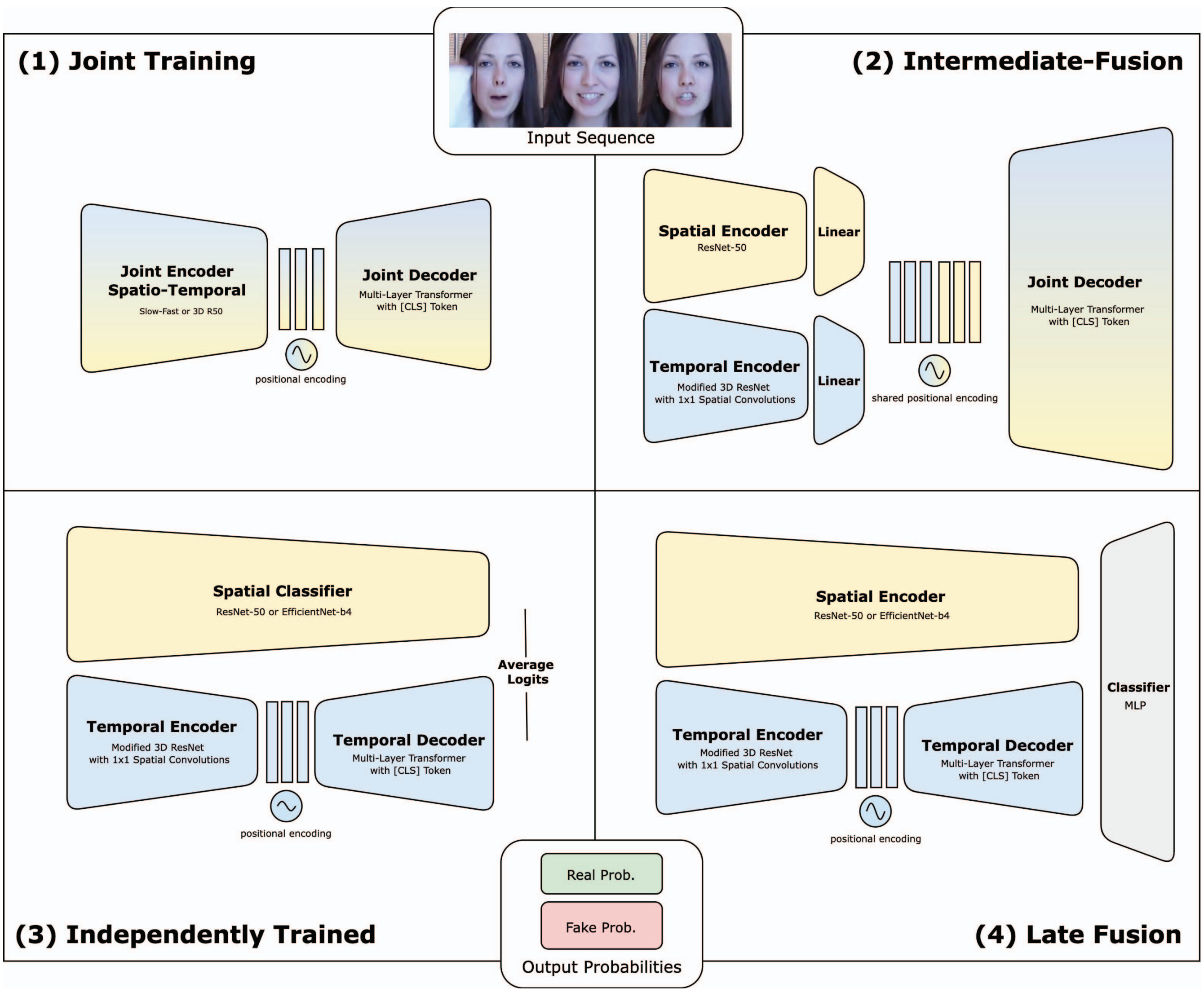

Figure 8. **Spatio-Temporal Architectures:** We illustrate our ablations for combining spatial and temporal features above. We pre-train the spatial and temporal encoders on self-blended images and on FF++ respectively.

	FF++	CDF [29]	V2V* [11]	MP* [8]	DIFF* [41]	GAN* [54]	SGE*	**AVG**
SBI + EfficientNet-b4 [40]	100.0	89.6	80.1	**99.3**	96.5	**97.2**	**94.0**	92.8
Lip Forensics [13]	100.0	82.4	92.3	17.8	50.6	98.6	93.2	68.3
Temporal (FTCN) [60]	99.9	86.5	**99.9**	77.1	75.0	83.7	88.0	85.0
LTTD [10]	100.0	89.3	-	-	-	-	-	89.3
Joint S-T (#1)	99.9	47.2	88.6	84.5	49.5	54.6	22.1	57.8
S-T Intermediate Fusion (#2)	99.9	78.2	90.0	85.0	55.6	59.0	71.2	64.1
S-T Ensemble (#3)	99.9	**90.7**	**99.9**	81.0	93.9	87.9	**94.0**	91.2
S-T Late Fusion (#4)	99.9	89.8	**99.9**	97.9	**99.4**	93.4	93.7	**95.7**

Table 1. **AUC Scores when Evaluating SOTA Methods**: The categories with the asterisk are our proposed evaluation metrics. Methods in italics are our spatio-temporal fusion methods. [10] was not evaluated since the codebase was not publicly available and we were not able to reproduce their results with our own code. We also note the limitation that *SGE* is only an upper-bound metric for generalizability, since it is possible for a classifier to achieve high classification performance on algorithmically generated psuedo-fakes, but fail to classify fakes generated by a deep-network. This appears to be the case for [13], which performs well on *SGE*, but poorly on *MegaPortraits* for instance.

a learnable positional embedding. Then, we linearly project the classification token to compute the classification logits.

2. **Intermediate-Fusion** We combine two Res-Net encoders, one with convolutions only along the spatial axes, and another with convolutions only along the temporal axis. These encodings are then projected into a shared latent space with per-modality encoders, and an asymmetric channel-wise dropout scheme. The full-token dropout prevents the network from binding to a specific artifact from a single modality. Finally, once the features are in the same latent space, we add shared learnable positional embeddings. The aforementioned transformer decoder learns to combine features from each modality, and outputs the target logits. We freeze the encoders after pre-training.

3. **Independent Training (Ensemble)** We ensemble the independently trained methods from [60] and [40] with equal weight.

4. **Late-Fusion** We freeze the methods from [60] and [40], and project the final embedding vectors from both methods with a three-layer perceptron. To prevent the classifier from picking up on only one of the modalities, we employ a paired modality dropout, i.e. when predicting on fakes we replace a modality input with the corresponding real video with a certain probability. We freeze the spatial and temporal encoders after pre-training on self-blended images and FF++ HQ.

5.2. Analysis

Our results in 1 show that the best performance overall was attained by the late-fusion model. Unlike the single modality detection methods, our late fusion method generalizes universally across fake types, either performing the highest or comparable to the highest AUC score on all fake types. We hypothesize that poor generalization performance in setups 1 and 2 were caused by direct spatial supervision. Prior works [10] conduct experiments describing how training on deepfake datasets with strong convolutions allows deep-fake detectors to cheat by picking subtle, unseen artifacts, rather than generalizable cues. Since the spatial encoder was frozen till the very last layer in setup 4, and we included modality dropout to prevent overfitting to purely the spatial domain, this method of combining did not suffer from the same generalizability issues. Overall, we show that relatively simple adjustments to existing methods can help existing methods generalize better out-of-domain.

The results in table 1 also generally support the validity of our simulated generalizability evaluation (*SGE*) method. Methods with higher *SGE* scores tend to also have high average evaluation score across the five out-of-domain SOTA methods considered.

6. Implementation Details

We first pre-process all videos by clipping into segments of 32 frames, then find the smallest square region that contains the face, and cropped with a static camera. We perform all training experiments using the train and validation sets from *Face Forensics++* (HQ compression). For evaluations of other methods, we consult the pre-processing steps from their code-bases. When computing prediction logits, we average prediction probabilities for all clips in the video. We use the real set from Celeb-DFv2 [29] for all our evaluations.

For our spatio-temporal experiments in Section 5, we train using Adam Optimizer and a fixed learning rate (0.001) chosen with grid-search on a validation set. We continue training for as many epochs as necessary until the validation loss does not significantly decrease for five consecutive epochs. All reported results are measured on the same out-of-sample test set as prior methods.

7. Conclusion

In this paper, we introduce a new evaluation benchmark for measuring generalizability on a modern, more diverse set of deepfakes. We also evaluate the state-of-the-art methods on our new benchmarks and identify multi-modal architectures to mitigate the variance that comes with highly handcrafted detection methods. We are releasing this evaluation set, along with the code for *SGE* to encourage further research in using randomized artifacts to improve generalizability metrics.

Acknowledgement: We thank the authors of [11], [54], [41], and [8] making the outputs from their methods available for our research. This work was supported in part by DoD including DARPA's SemaFor program.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: A compact facial video forgery detection network. 2018.

[2] 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.

[3] 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.

[4] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

[5] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9468–9478, June 2022.

[6] 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.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[8] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. *arXiv preprint arXiv:2207.07621*, 2022.

[9] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019.

[10] Jiazhi Guan, Hang Zhou, Zhibin Hong, Errui Ding, Jingdong Wang, Chengbin Quan, and Youjian Zhao. Delving into sequential patches for deepfake detection. *arXiv preprint arXiv:2207.02803*, 2022.

[11] Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. Space: Speech-driven portrait animation with controllable expression, 2022.

[12] 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, 2018.

[13] 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.

[14] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017.

[15] 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.

[16] Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400*, 2020.

[17] Shehzeen Hussain, Paarth Neekhara, Cristian Canton Ferrer, Julian McAuley, and Farinaz Koushanfar. Exposing vulnerabilities of deepfake detection systems with robust attacks. *Digital Threats: Research and Practice*, 2021.

[18] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021.

[19] Tackhyun Jung, Sangwon Kim, and Keecheon Kim. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, 2020.

[20] Ali Khodabakhsh, Raghavendra Ramachandra, Kiran Raja, Pankaj Wasnik, and Christoph Busch. Fake face detection methods: Can they be generalized? In *2018 international conference of the biometrics special interest group (BIOSIG)*, pages 1–6. IEEE, 2018.

[21] 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.

[22] 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.

[23] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-ray for more general face forgery detection. 2020.

[24] 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.

[25] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing AI created fake videos by detecting eye blinking. 2018.

[26] 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.

[27] Yuezun Li, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, Seattle, WA, United States, 2020.

[28] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df (v2): a new dataset for deepfake forensics. *arXiv preprint arXiv:1909.12962*, 4, 2019.

[29] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. 2020.

[30] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021.

[31] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92, 2019.

[32] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint arXiv:1804.03447*, 2018.

[33] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. 2019.

[34] 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.

[35] 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. 2020.

[36] Ekta Prashnani, Michael Goebel, and B. S. Manjunath. Generalizable deepfake detection with phase-based motion analysis, 2022.

[37] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. 2020.

[38] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. 2019.

[39] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized talking head synthesis. *arXiv preprint arXiv:2301.03786*, 2023.

[40] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022.

[41] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zieba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. *arXiv preprint arXiv:2301.03396*, 2023.

[42] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[43] Ayush Tewari, Justus Thies, and Michael Zollhöfer. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13081–13090, 2020.

[44] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. 2020.

[45] 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.

[46] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Facegan: Facial attribute controllable reenactment gan. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1329–1338, 2021.

[47] Andrea Valenzuela, Carlos Segura, Ferran Diego, and Vicenç Gómez. Expression transfer using flow-based generative models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1023–1031, 2021.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[49] Hong-Xia Wang, Chunhong Pan, Haifeng Gong, and Huai-Yu Wu. Facial image composition based on active appearance model. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 893–896, 2008.

[50] Junke Wang, Zuxuan Wu, Jingjing Chen, and Yu-Gang Jiang. M2TR: Multi-modal multi-scale transformers for deepfake detection. *arXiv preprint arXiv:2104.09770*, 2021.

[51] 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.

[52] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021.

[53] Mika Westerlund. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 2019.

[54] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. In *Advances in Neural Information Processing Systems*.

[55] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. 2019.

[56] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7556–7566, 2019.

[57] Eric Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *ICCV 2021*, 2021.

[58] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. 2021.

[59] 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.

[60] 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.

[61] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Two-stream neural networks for tampered face detection. 2017.

[62] 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.