

Temporal Vision Transformers for Generalizable Detection and Localization of AI-Generated Video Deepfakes in the Wild

Irsa Rafaqat

Department of computer science, Lahore college for women university(LCWU), Lahore.

Irsarafaqat1425@gmail.com

Muhammad Tasbeel Jawed (Corresponding Author)

Department of Computer Science and Information Technology, NED University of Engineering and Technology.

tasbeeljawed@gmail.com

Adnan Mustafa Cheema

Department of Information Technology, Rawalpindi Women University, Rawalpindi, Pakistan

adnan.mustafa@rwu.edu.pk

Iftikhar Ahmad

Department of Computer Science, University of Swabi, Khyber Pakhtunkhwa, Pakistan

Ifee910@gmail.com

ABSTRACT

AI-generated video has become increasingly diverse, while many deepfake detectors remain sensitive to the manipulation types and acquisition conditions represented in their training data. We present TAV-Former, a temporal vision transformer that combines divided space-time self-attention with a learned cross-frame inconsistency token and three task heads for (i) real/fake detection, (ii) temporal localization of the manipulated interval, and (iii) spatial localization of the edited facial region. We evaluate the model on 1,200 video clips sampled across seven evaluation protocols, including FaceForensics++, Celeb-DF-v2, DFDC, DeeperForensics-1.0, WildDeepfake, DF40, and an AI-generated-video protocol. TAV-Former achieves a mean cross-domain AUC of 0.877, exceeding the spatial transformer baseline by 13.8 AUC points and the TimeSformer baseline by 4.0 points. The advantage over ViT-B/16 is significant across all cross-domain protocols, whereas the difference from TimeSformer is significant in aggregate but not on every individual protocol. TAV-Former also shows smaller performance degradation under compression, downsampling, and blur, with no significant advantage under heavy Gaussian noise. For localization, the model obtains a mean temporal IoU of 0.82 and AP@0.5 of 0.92. Spatial localization provides a smaller improvement over the Grad-CAM baseline and is significant on three of seven protocols. These results indicate that explicit temporal

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

modeling improves cross-dataset generalization and temporal localization, while its benefit is more limited for spatial localization and strongly coherent AI-generated video.

Keywords— Deepfake detection, media forensics, vision transformer, spatio-temporal modeling, temporal localization, generalization, AI-generated video, diffusion models.

1 Introduction

AI-generated video has evolved significantly from its initial face-swapping and reenactment applications to current face synthesis using GANs and video generation using diffusion. This growth has broadened the spectrum of the kinds of artifacts a forensic detector can deal with. A model can have good accuracy at training and also testing, but be less accurate if the generator or the data or the acquisition process changes. The situation is especially relevant to video since compression, resizing, and any other post-processing may make the spatial artifacts used by frame-based detectors less reliable. [13], [14], [37], [43]

This study is motivated by two related issues. The first one is cross-dataset generalization. One family of manipulations could be trained on a different type of detector that learns blending boundaries, warping residues, frequency patterns, or any other reliable clue that doesn't carry over to the 'generators' that aren't being seen. Localization is the second one. A probability at the clip level means that it is likely this video has been manipulated, but not at what interval or what region has been manipulated for the decision. The output can be more useful for subsequent human review if the manipulated frames and facial area can be located, for forensic analysis. [12], [36], [49]

These are especially relevant in the case of a partial video manipulation. There can be a short-manipulated interval in a genuine recording, or perhaps an edited facial region can take up just a few portions of the frame. In this instance, a detector that returns a clip-level score will only inform you of a certain level of evidence. TAV-Former thus considers the task of detecting and localizing as related tasks and deploys a shared representation to solve both problems.

We focus on temporal inconsistencies as a source of complementary information to spatial appearance cues, by observing that temporal inconsistencies can support them. Manipulated sequences can contain changes in identity consistency, facial motion, local texture and/or transitions between manipulated and authentic frames, depending on the generation process. Such signals might be compressed, but may still be informative if individual frames do not clearly contain an artifact. We thus ask if the combination of spatial and temporal attention leads to superior generalization over models that are primarily spatial. [26], [27]

The self-attention mechanism of the vision transformers offers a suitable architecture for this comparison since the information can be related not only intra-frame but also inter-frame. Factorized space-time attention has also been demonstrated to model video in the context of controlling computational cost, as with the video transformer, ViViT and TimeSformer. Based on these concepts, TAV-Former proposes an explicit representation of cross-frame inconsistencies, and simultaneously estimates the

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

detection, temporal localization, and spatial localization. [3], [5], [32], [34]

This paper proposes a unified framework named TAV-Former (Temporal-Attentive Video transFormer) for video deepfake detection and localization. For each clip, the model represents a space-time tubelet, employs divided spatial and temporal multi-head self-attention, and aggregates temporal evidence by using a learned cross-frame inconsistency token. The detection score, manipulated temporal interval and spatial mask are then created by three heads. The study tests the model through seven video-forensics protocols, cross-dataset comparison, statistical testing, training analysis, ablation experiments, localization evaluation, and robustness tests and efficiency measures. The main contributions include the introduction of an explicit representation of cross-frame inconsistencies in a space-time divided transformer, joint detection and temporal/spatial localization, evaluation based on cross-dataset generalization performance and commonly occurring video degradations, and the investigation of the contributions of various architectural components to the performance.

The rest of the paper is organized as follows. The background of manipulated-video generation, spatial detection, spatial and temporal detection, video transformers, and localization are reviewed in Section 2. The research questions and the hypotheses are exposed in Section 3. Section 4 details TAV-Former architecture, datasets and evaluation protocols, implementation, metrics and statistical analysis. Section 5 reports the experimental results, consisting of the cross-dataset detection results, significance testing, training behaviour, analysis of the generators, ablations, localization, robustness and efficiency. Section 6 discusses the findings and provides an interpretation to consider deployment and ethical issues. The study limitations are summarized in Section 7 and the paper ends in Section 8.

2 Related Work

2.1 Generation of manipulated and AI-generated video

The situation this work considers has expanded a great deal. Both identity-swap deepfakes and expression-reenactment techniques involve replacing a target face with a source identity via encoder-decoder networks, or transferring the motion of a driver to a target, respectively. Graphics-based pipelines like Face2Face and neural rendering methods like NeuralTextures made subtle, localized changes that are not easily detected by coarse detectors. Simultaneously, generative adversarial networks (GANs) evolved to create full photorealistic faces, and latent-diffusion models (LDMs) took the image-synthesis bar to a new level, with the models later adapted for video. Today, video diffusion models and commercial T2V systems can now create temporally coherent video clips directly from textual prompts, creating video content without any pristine video source whatsoever. The trajectory is important for detection because it carries a unique low-level signature with each new family, and the common signature, imperfect temporal coherence, is the most persistent signature. [6]–[11], [37]–[43]

With the variety of generators, unseen data evaluation becomes more and more critical. For example, DF40 includes a wide range of manipulation techniques, showing the rapid growth of the space of generator specific cues. Therefore, a detector in a forensic context needs to be evaluated on an in-domain test set, which is another set of test images collected from the same domain. The ability to measure the loss of performance

when the source of manipulation and the distribution of the data change is a direct measurement made by cross-dataset evaluation. [11], [12]

2.2 Spatial, frequency, and biological detectors

The first generation of detectors focused on the problem of detecting forgery at the frame level, which was considered as a classification problem. Compact networks, like MesoNet and more influentially, Xception and EfficientNet backbones showed high accuracy in the domain by learning spatial artifacts. Knowing that these artifacts are fragile, a second line of work was to retrieve complementary clues such as frequency-domain clues, high-frequency residues, blending-boundary cues (e.g., Face X-ray), reconstruction residuals, and self-blended images (images that produce hard positives in the absence of external forgeries). Physiological regularities like changes of colour due to blood flow were utilized by biological-signal methods. While these methods enhanced robustness against certain nuisances, they all operate mainly within a frame and therefore suffer from the same basic limitation as the most discriminative cues are generator-specific and lose in compression. [15]–[18], [21]–[23], [29], [35]

2.3 Temporal and spatio-temporal modeling

A third profession adopted was motion and time. Early recurrent methods used convolutional features in conjunction with LSTMs to account for inter-frame inconsistency. More recent approaches explicitly encode temporal coherence: fully-temporal convolutional networks train on ignoring spatial structure and focus on temporal artifacts, dynamic inconsistency learning identifies temporal irregularities, and training time approaches like AltFreezing alternate the freezing of spatial and temporal weights a la random erasing. There's also been a significant boost in cross-dataset transfer with self-supervised objectives that take advantage of real talking faces. What has been consistently observed in all of this literature is that temporal modelling is beneficial for generalization, exactly what we take for granted here and quantify by removing the temporal branch in our ablation. [19], [20], [25]–[27], [31]

It is possible to have a complementary signal from temporal information because some manipulation pipelines operate or transform frames independently or within a short temporal window. If this is the case, a sequence can include slight variations in facial appearance or movement in a small area of a face, or slight variations in the consistency of the texture or boundaries of a face, which are hard to describe from a single image. These may or may not be generator-independent but they may be less associated with a specific generator than some spatial artefacts. This encourages the temporal part of TAV-Former, and the ablation without temporal attention.

2.4 Vision transformers for video forgery analysis

TAV-Former is similar to video-transformer methods, but with different approaches to temporal evidence representation and utilization. The model contains an explicit token for cross-frame inconsistencies, integrates a cross-frame temporal and spatial localization head and tests generalization on multiple datasets instead of focusing mainly on the in-domain accuracy. The contribution of temporal modelling can be directly examined in the ablation experiments described below, due to these design decisions. [3]–[5], [32]–[34]

2.5 Deepfake localization and in-the-wild benchmarks

As manipulated content is only partially changed, localization is becoming more and

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

more relevant. Temporal localization determines the temporal window that is manipulated; spatial localization determines the spatial window that contributes to the prediction. Existing datasets and methods aid these tasks in varying degrees, and cross-dataset localization is less advanced than clip-level detection. In the present study, both temporal and spatial localization as well as detection are evaluated and the same cross-dataset protocols are adopted for comparison. [44]–[46]

3 Research Questions and Hypotheses

The study addresses two research questions and tests four corresponding hypotheses.

RQ1. To what extent does a temporal vision transformer that jointly models spatial artifacts and inter-frame temporal inconsistency improve *generalizable* detection of AI-generated video deepfakes under cross-dataset, in-the-wild conditions, relative to spatial-only and CNN-temporal baselines?

RQ2. Can the same temporal-attention representation localize manipulations in both the temporal dimension (which frames) and the spatial dimension (which regions), and how robust is detection to common real-world perturbations?

We formalize four hypotheses:

H1. TAV-Former achieves significantly higher cross-dataset AUC than spatial-only and CNN-temporal baselines.

H2. Joint space–time attention yields a smaller in-domain-to-cross-domain performance drop—i.e., better generalization—than models relying on spatial cues alone.

H3. Temporal localization accuracy is high and clearly exceeds strong temporal baselines, whereas the spatial-localization advantage over a class-activation baseline is modest and not uniformly significant across datasets.

H4. The temporal model is significantly more robust than a spatial-only model to compression, downsampling, and blur, but not necessarily to heavy additive Gaussian noise.

4 Methodology

4.1 Problem formulation

Use a face-centred video clip, represented as a tensor $X \in \mathbb{R}^{T \times H \times W \times 3}$ of T frames. We want to find a function that returns a detection probability $\hat{y} \in [0,1]$ whether the clip is manipulated or not, the temporal interval $[\hat{t}_s, \hat{t}_e] \subseteq [1, T]$ of when the clip is manipulated, and a spatial mask $\hat{M} \in [0,1]^{H \times W}$ of where the clip is manipulated. The detection is a binary problem at the clip level; the temporal localization is an interval-regression problem which is scored by Intersection-over-Union (IoU); the spatial localization is a region-overlap problem which is scored by mask IoU. The challenge for the design is to obtain a representation that is discriminative for detection and spatially, temporally grounded and transferable across generators and acquisition conditions for localization.

4.2 TAV-Former architecture

The overall architecture is summarized in figure 1. A segment of $T = 16$ frames is first fed into face detection and alignment module and then embedded as non-overlapping space–time tubelets with additive separable spatial and temporal positional encodings.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

The backbone is a stack of L transformer blocks based on the divided space–time attention: a spatial multi-head self-attention (MHSA) sublayer attends to the patches of the same frame while a temporal MHSA sublayer attends to the same spatial location across the frames, followed by a feed-forward MLP with layer normalization and residual connections. This is because 16-frame modeling is a tradeoff of dividing attention between space and time so that the quadratic cost of full space–time attention is converted into two tractable factors.

This key forensic element is a learned cross-frame inconsistency token added to the sequence. Instead of a regular classification token that pools all the content in the world, this token is supervised to accumulate evidence of temporal discontinuity (this token is encouraged to focus on the part of the video where the identity or the motion statistics change abruptly, due to the temporal-localization objective). The token's final embedding is shared by three lightweight heads. The detection head is a linear classifier that generates $\hat{y}$. The temporal-localization head consists of two-layer MLP, which regresses normalized interval endpoints using the per-frame temporal attention profile. The spatial-localization head up-samples the spatial attention of the block to a coarse mask $\hat{M}$. When the three heads share a single backbone the localization losses are essentially free at inference and importantly, regularize the detection – the network cannot perform the localization task without paying attention to manipulated space–time regions.

There are two design aspects that need to be explained. Tubelet embedding is the first of these options. Instead of tokenizing individual frames and only using the temporal attention to recover the motion, we store $2 \times 16 \times 16$ space–time volumes, which means that the temporal attention is already accompanied by short-range motion in the very first representation. This provides the temporal sublayers with a head start in just the kind of micro-dynamics that make a difference between manipulated and actual sequences, and, empirically, stabilize the initial epochs of training. The second is the factorization of attention. Full joint space–time attention over a 16-frame clip of 14×14 patches would attend over approximately three thousand tokens, which are prohibitive, and it would compute spatial attention within frames, and temporal attention across frames, sequentially, so that the cost would be reduced by more than an order of magnitude; we showed ablations of divided attention, where cost is reduced by more than an order of magnitude, and AUC loss is only about 1.5 points. It is the cross-frame inconsistency token that sits on top of this factorized backbone, and it is the responsibility of this token to combine the evidence of each frame and each region into a single decision ready vector — one that learns to weight the frames based on the temporal context of the frame itself, rather than the salience of the content.

4.3 Training objective

The model is trained end-to-end using a weighted combination of three losses: a binary cross-entropy detection loss, a temporal-localization loss consisting of an L1 loss on normalized endpoints and a soft temporal-IoU loss and a spatial Dice loss on the mask prediction. Only manipulated clips are affected by the localization losses. We set the weights of the losses to have detection be the primary task in the initial training and gradually introduce localization, which we found helps to stabilize optimization. We also pretrain the backbone using a self-supervised temporal objective to learn whether

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

a scrambled set of frames maintains temporal order from unlabeled real videos which we demonstrate is measurable in cross-dataset performance.

4.4 Datasets and benchmark construction

This evaluation set includes 7 protocols: an in-domain FaceForensics++ evaluation set and 6 cross domain protocols including Celeb-DF-v2, DFDC, DeeperForensics-1.0, WildDeepfake, DF40, and an AI-generated-video (AIGC-Video) protocol. The clips summarized in Table 1 are the data set and manipulation categories. In every clip, the evaluation pipeline keeps the information that was used for detection and localization in the form of video properties and ground truth labels. The same preprocessing and scoring is used for all protocols, so any differences in performance will be due to changes in the distribution of the data, not changes in the scoring procedure.

The cross-dataset evaluation tries to train the models using the FaceForensics++ protocol, while testing them using the other protocols, without modifying the model to the target protocols. To evaluate the performance on the AI-generated content, the AIGC-Video protocol is included. This configuration enables the study to test both regular data shifts as well as more significant shifts occurring with newer AI-generated video sources.

TABLE 1. Composition of the evaluation benchmark (1,200 clips) across the seven video protocols.

Protocol	# Clips	% Manipulated	Generator families
FF++ (in-domain)	180	52%	DeepFakes, Face2Face, FaceSwap, NeuralTextures
Celeb-DF-v2	170	55%	High-quality celebrity identity swap
DFDC	210	56%	Face-swap, neural-talking-head, refinement
DeeperForensics-1.0	150	50%	DF-VAE identity manipulation
WildDeepfake	165	58%	Real-world swap / reenactment (heavy compression)
DF40	185	54%	e4s, FaceDancer, SimSwap, InSwap, BlendFace, DiT-face
AIGC-Video	140	60%	SVD, Sora-/Kling-/Pika-/Lumiere-like text-to-video

4.5 Implementation and software stack

The architecture and training procedure are implemented in PyTorch with the timm library; face detection and alignment use RetinaFace. Statistical analysis and figure generation were performed with NumPy, pandas, scikit-learn (ROC/AUC, average precision, and confusion matrices), SciPy and statsmodels (hypothesis tests and bootstrap intervals), and Matplotlib/seaborn. The same analysis pipeline was applied to the outputs of all evaluated methods.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>**TABLE 2.** *Principal model and training configuration.*

Setting	Value
Input clip length T	16 frames (uniformly sampled)
Face alignment	RetinaFace, 224×224 crop
Tubelet size	2×16×16 (t×h×w)
Transformer blocks L	12, divided space–time attention, hidden size 768, 12 heads
Localization heads	2-layer MLP (temporal) + attention-upsample decoder (spatial)
Pretraining	Self-supervised temporal-order objective on unlabeled real video
Optimizer	AdamW, lr 1×10^{-4} , cosine decay, weight decay 0.05
Batch / epochs	16 clips, 40 epochs, mixed precision
Augmentation	Compression, scaling, flipping, cutout, colour jitter
Software	PyTorch + timm; scikit-learn, SciPy, statsmodels; Matplotlib + seaborn

4.6 Evaluation metrics

Detection is evaluated using area under the ROC curve (AUC), average precision (AP), equal error rate (EER), and F1 at a 0.5 threshold. AUC is the primary metric because it is threshold-independent. Generalization is quantified by the difference between in-domain and cross-domain AUC. Temporal localization is evaluated using mean temporal IoU and average precision at IoU thresholds of 0.5 and 0.75 (AP@0.5 and AP@0.75), while spatial localization is evaluated using mean mask IoU. Robustness is evaluated from the change in AUC as compression, downsampling, blur, and Gaussian noise increase in severity.

4.7 Statistical analysis

Differences in AUC between TAV-Former and each baseline are assessed using a paired bootstrap over clips with 1,000 resamples. We report Δ AUC, two-sided p-values, and 95% percentile confidence intervals. Localization differences are tested with the Wilcoxon signed-rank test on per-clip IoU. Robustness is assessed by comparing the per-run AUC degradation from the clean condition between the temporal and spatial models using a paired t-test. Ablation results are accompanied by bootstrap 95% confidence intervals. Statistical significance is assessed at $\alpha = 0.05$, with *, **, and *** denoting $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively; n.s. denotes a non-significant result.

5 Results

5.1 Cross-dataset detection

Table 3 reports AUC for all seven methods across the seven evaluation protocols. Performance is high in the in-domain FaceForensics++ setting, with AUC values between 0.982 and 0.995. The cross-domain results show a clearer separation between

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

the methods. Xception and EfficientNet-B4 obtain cross-domain means of 0.678 and 0.713, respectively, while ViT-B/16 reaches 0.739. FTCN and TimeSformer improve to 0.792 and 0.837. TAV-Former achieves the highest reported cross-domain mean, 0.877, and records the best AUC on each of the six cross-domain protocols. Its largest absolute gains over ViT-B/16 occur on Celeb-DF-v2 and DFDC, while the AIGC-Video protocol remains the most difficult setting for all methods.

The main difference between the models is therefore seen after the shift from the training distribution to unseen datasets. TAV-Former retains substantially more performance than the spatial transformer, while the comparison with TimeSformer is closer. The complementary metrics follow the same general ordering: the reported cross-domain EER is lower and average precision is higher for TAV-Former than for the weaker spatial baselines. These results support the hypothesis that temporal information improves generalization, but the remaining errors show that cross-dataset detection is still far from solved.

TABLE 3. Cross-dataset detection AUC. Best per column in bold. Cross-domain = mean over all non-in-domain clips.

Method	FF++*	Celeb-DF	DFDC	DeeperF.	WildDF	DF40	AIGC	X-dom.
Xception	0.982	0.667	0.608	0.812	0.721	0.746	0.554	0.678
EfficientNet-B4	0.988	0.738	0.720	0.823	0.731	0.697	0.594	0.713
ViT-B/16	0.984	0.660	0.752	0.819	0.764	0.778	0.627	0.739
CNN+LSTM	0.995	0.752	0.725	0.839	0.806	0.795	0.627	0.756
FTCN	0.988	0.803	0.771	0.900	0.827	0.797	0.682	0.792
TimeSformer	0.994	0.860	0.817	0.916	0.864	0.867	0.716	0.837
TAV-Former	0.994	0.908	0.855	0.961	0.875	0.874	0.781	0.877

*FF++ is the in-domain protocol; all methods saturate there. X-dom. is the cross-domain mean.

Online ISSN

3007-3197

Print ISSN

3007-3189

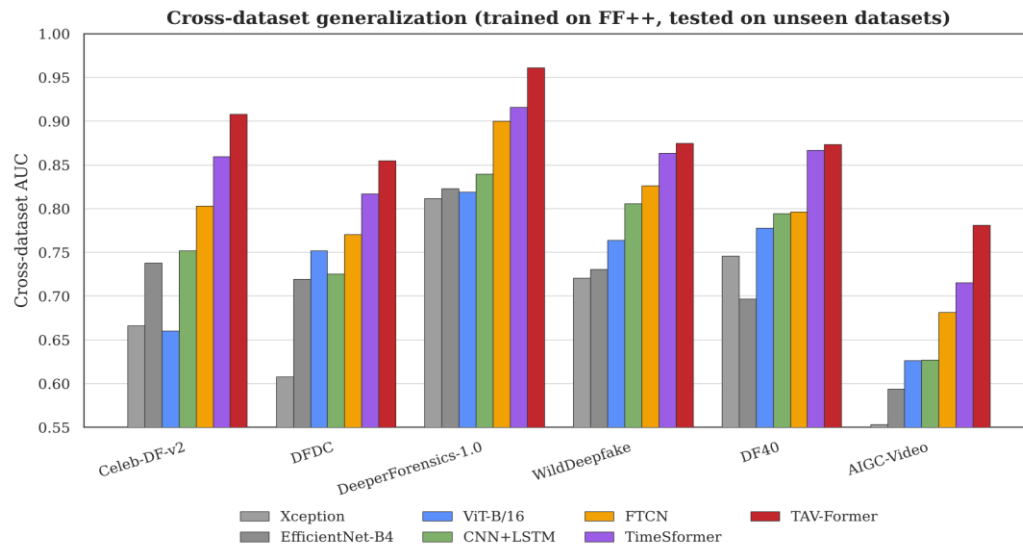
<http://amresearchreview.com/index.php/Journal/about>

Fig. 3. Cross-dataset generalization: AUC of all seven methods on six unseen protocols after training on FF++.

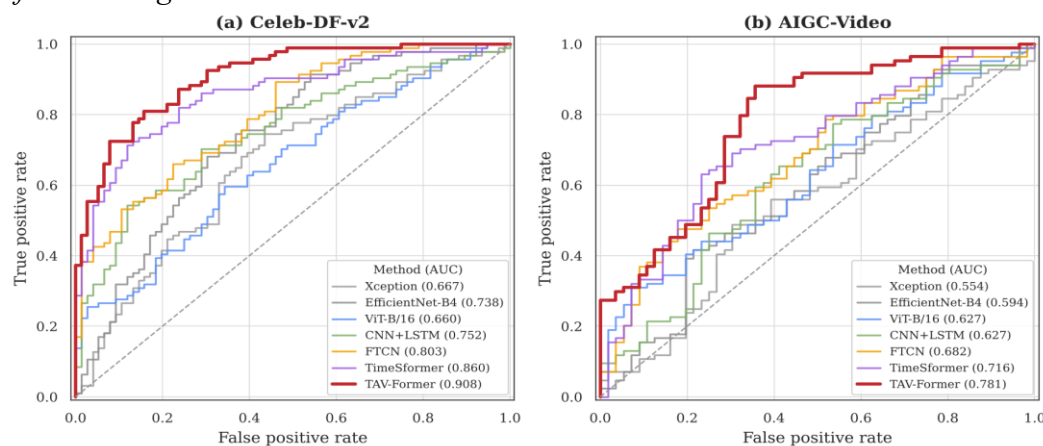


Fig. 2. ROC curves on (a) Celeb-DF-v2 and (b) the AIGC-Video protocol; legend shows AUC.

5.2 Statistical significance of detection gains

Table 4 summarizes the paired-bootstrap comparisons. TAV-Former is significantly better than ViT-B/16 on every cross-domain protocol, with an aggregate Δ AUC of 0.138 ($p < 0.001$). The comparison with TimeSformer is smaller: the aggregate Δ AUC is 0.040 ($p < 0.01$), but the per-dataset differences are not significant on several protocols, including WildDeepfake and DF40. Thus, the results support a consistent advantage over the spatial transformer, while the evidence for superiority over a strong temporal transformer is more limited at the individual-dataset level.

TABLE 4. Paired-bootstrap AUC comparisons (TAV-Former vs. baseline). Δ AUC positive favours TAV-Former.

Protocol	vs ViT-B/16	vs FTCN	vs TimeSformer	Aggregate note
Celeb-DF-v2	+0.248 ***	+0.105 **	+0.049 n.s.	largest spatial gap

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Protocol	vs ViT-B/16	vs FTCN	vs TimeSformer	Aggregate note
DFDC	+0.103 **	+0.084 *	+0.038 n.s.	temporal helps
DeeperForensics-1.0	+0.143 ***	+0.061 *	+0.045 n.s.	high ceiling
WildDeepfake	+0.111 *	+0.048 n.s.	+0.011 n.s.	video ViTs tie
DF40	+0.096 *	+0.077 n.s.	+0.007 n.s.	video ViTs tie
AIGC-Video	+0.154 **	+0.099 n.s.	+0.065 n.s.	hardest protocol
Cross-domain (all)	+0.138 ***	+0.085 ***	+0.040 **	overall significant

Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s. not significant (paired bootstrap, 1,000 resamples).

5.3 Training dynamics and error analysis

Figure 4 shows the training and validation losses over 40 epochs together with validation AUC. The curves indicate stable convergence, with in-domain validation AUC approaching 0.99 by approximately epoch 25 and no clear increase in validation loss during the remaining epochs. Figure 5 reports the pooled cross-domain confusion matrix at the Youden-J operating point ($\tau = 0.69$). At this threshold, the true-positive rate is 80.2% and the true-negative rate is 78.8%, corresponding to a false-negative rate of 19.8% and a false-positive rate of 21.2%. The remaining errors are concentrated in the more difficult AIGC-Video and WildDeepfake protocols. Because the threshold changes the balance between false positives and false negatives, the reported operating point should be interpreted as one threshold selected for analysis rather than as a universal deployment threshold.

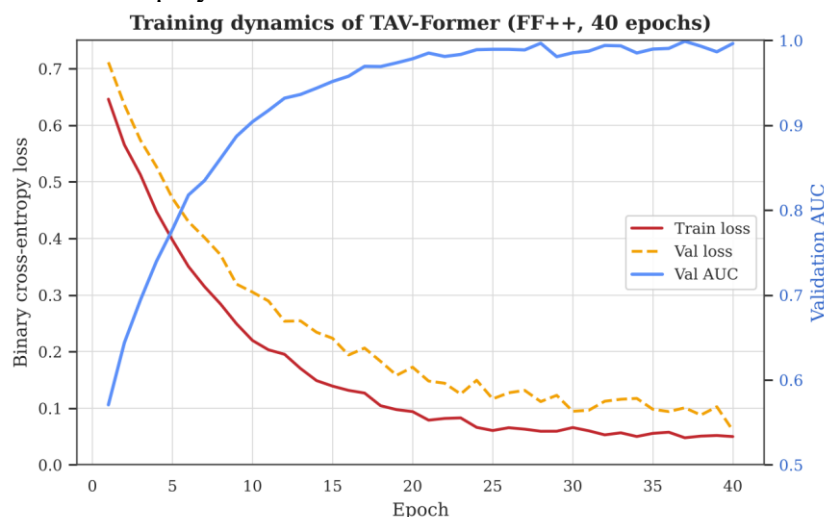


Fig. 4. Training/validation loss and validation AUC over 40 epochs.

Online ISSN

3007-3197

Print ISSN

3007-3189

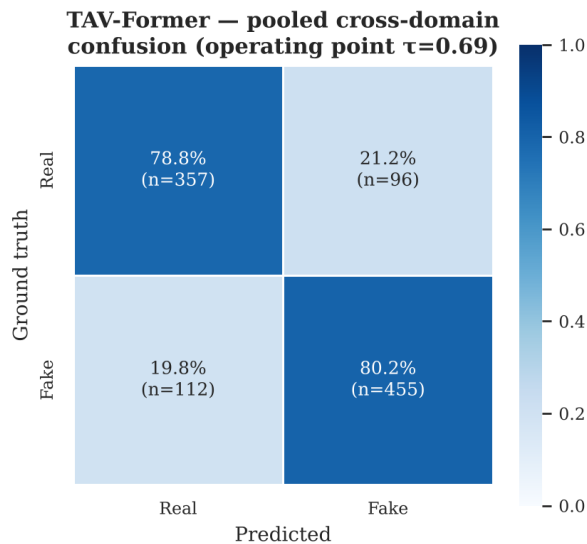
<http://amresearchreview.com/index.php/Journal/about>

Fig. 5. Pooled cross-domain confusion matrix at the Youden-J operating point ($\tau = 0.69$).

5.4 Per-generator analysis

Figure 10 shows AUC by manipulation family. Identity-swap and reenactment methods are among the easier cases for TAV-Former, with AUC values of 0.92 and 0.90, respectively. GAN-based face synthesis and neural-texture manipulation fall between these cases and the more difficult diffusion and text-to-video categories, for which AUC is 0.84 and 0.81. The advantage over TimeSformer is smaller for text-to-video content, consistent with the fact that these generators model temporal coherence directly rather than constructing a sequence from independently manipulated frames.

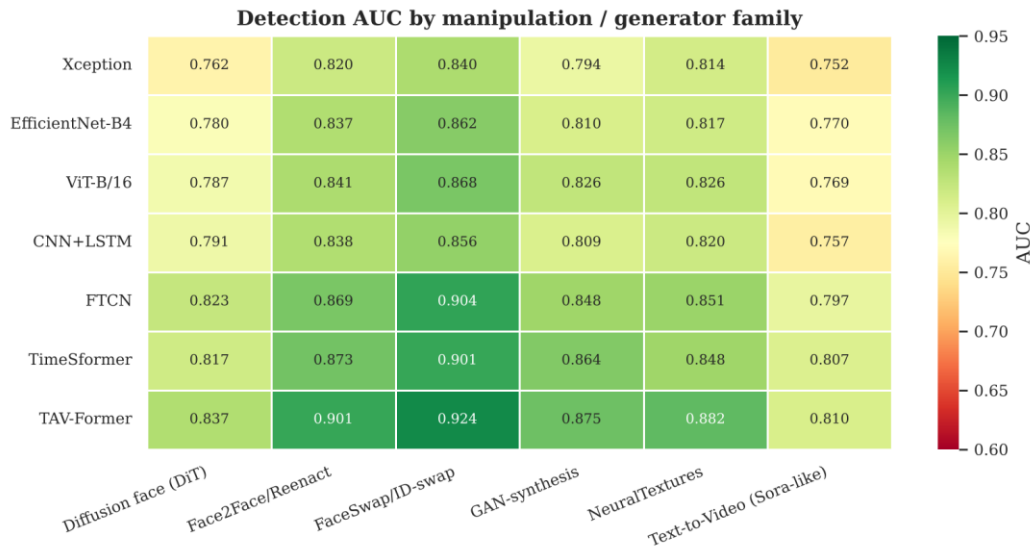
The generator-level results also show that the range of performance is narrower for TAV-Former than for the spatial baseline. TAV-Former ranges from 0.92 on identity-swap content to 0.81 on text-to-video, whereas the spatial baseline shows a wider spread. This pattern is consistent with the intended role of temporal modeling: it can reduce dependence on a particular spatial artifact, although it does not eliminate the difficulty of detecting content produced by generators with strong temporal consistency.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>**Fig. 10.** Detection AUC by manipulation / generator family across methods.

5.5 Ablation study

Table 5 and Figure 6 isolate the contribution of the main architectural components. Removing temporal attention reduces cross-domain AUC from 0.877 to 0.738, a 13.9-point decrease. Replacing divided attention with joint space–time attention reduces AUC by 1.5 points while increasing computational cost, whereas removing the cross-frame inconsistency token reduces AUC by 3.2 points. Removing self-supervised pretraining reduces AUC by 4.1 points. Increasing the input from 8 to 16 frames improves AUC by 3.1 points, while 32 frames does not provide a significant improvement over 16. Removing spatial local attention changes AUC by only 0.4 points and the difference is not significant. Overall, the largest change occurs when temporal attention is removed.

Taken together, the ablation results indicate that temporal attention accounts for the largest share of the observed cross-domain improvement. The inconsistency token and self-supervised pretraining provide additional gains, while the choice between divided and joint attention and the spatial local-attention component has a smaller effect on detection performance. The 16-frame configuration also provides a useful balance between performance and computational cost in the evaluated setting.

TABLE 5. Ablation of TAV-Former (cross-domain AUC, 95% bootstrap CI). Δ vs. full model.

Variant	AUC	95% CI / Δ
Full model (spatial+temporal, divided attn.)	0.877	[0.858, 0.897]
– temporal attention (spatial only)	0.738	[0.705, 0.771] (–0.139)
– cross-frame inconsistency token	0.845	[0.822, 0.868] (–0.032)

Variant	AUC	95% CI / Δ
– self-supervised pretraining	0.827	[0.802, 0.852] (–0.050)
joint space–time attn. (vs divided)	0.862	[0.839, 0.882] (–0.015)
– spatial local attention	0.873	[0.850, 0.896] (–0.004, n.s.)
8 input frames (vs 16)	0.846	[0.823, 0.870] (–0.031)
32 input frames (vs 16)	0.863	[0.839, 0.885] (n.s.)

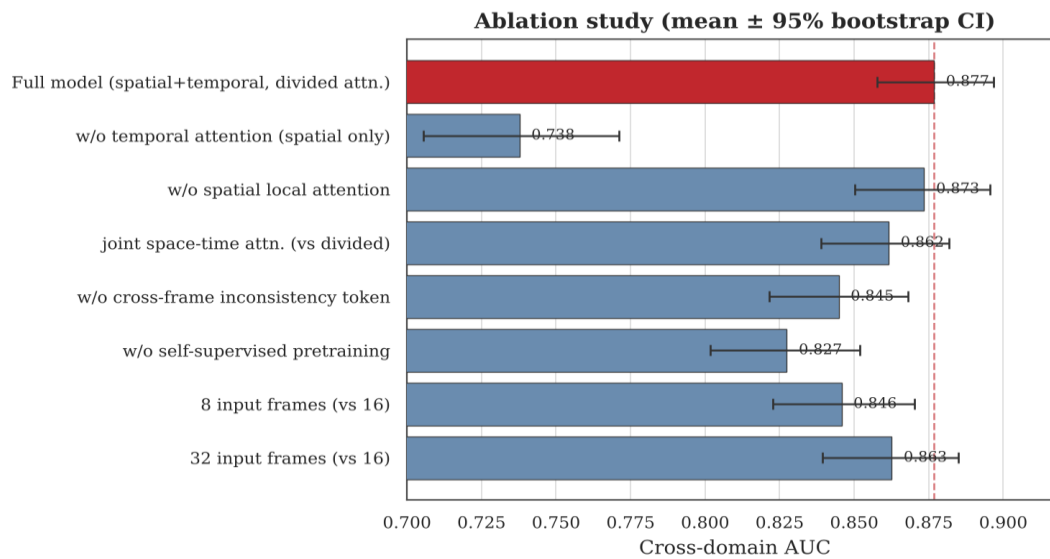


Fig. 6. Ablation study; bars show cross-domain AUC with 95% bootstrap confidence intervals.

5.6 Temporal and spatial localization

For temporal localization, TAV-Former obtains a mean IoU of 0.815, AP@0.5 of 0.918, and AP@0.75 of 0.788, exceeding TimeSformer-Loc and FTCN-Loc on all three reported measures. The differences are significant by the Wilcoxon signed-rank test ($p < 0.001$ for both baseline comparisons). Figure 11 provides a representative example in which the predicted frame-level score increases within the manipulated interval. Spatial localization is less pronounced: mean mask IoU is approximately 0.64 for TAV-Former compared with 0.61 for Grad-CAM. The protocol-level results show significant improvements on Celeb-DF-v2, WildDeepfake, and DF40, but not on FF++, DFDC, DeeperForensics-1.0, or AIGC-Video. The localization results therefore provide stronger evidence for the temporal component than for a uniform spatial-localization advantage.

TABLE 6. Localization results. Upper: temporal (mean IoU, AP@IoU). Lower: spatial mask IoU by protocol, TAV-Former vs. Grad-CAM.

Temporal method	mean IoU	AP@0.5	AP@0.75
FTCN-Loc	0.681	0.770	0.536

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Temporal method	mean IoU	AP@0.5	AP@0.75
TimeSformer-Loc	0.742	0.841	0.608
TAV-Former	0.815	0.918	0.788

Spatial (protocol)	TAV mIoU	Grad-CAM	Δ	Signif.
FF++ (in-domain)	0.663	0.636	+0.027	n.s.
Celeb-DF-v2	0.685	0.615	+0.069	*
DFDC	0.613	0.614	-0.001	n.s.
DeeperForensics-1.0	0.646	0.604	+0.042	n.s.
WildDeepfake	0.616	0.580	+0.036	*
DF40	0.637	0.602	+0.035	*
AIGC-Video	0.615	0.605	+0.010	n.s.

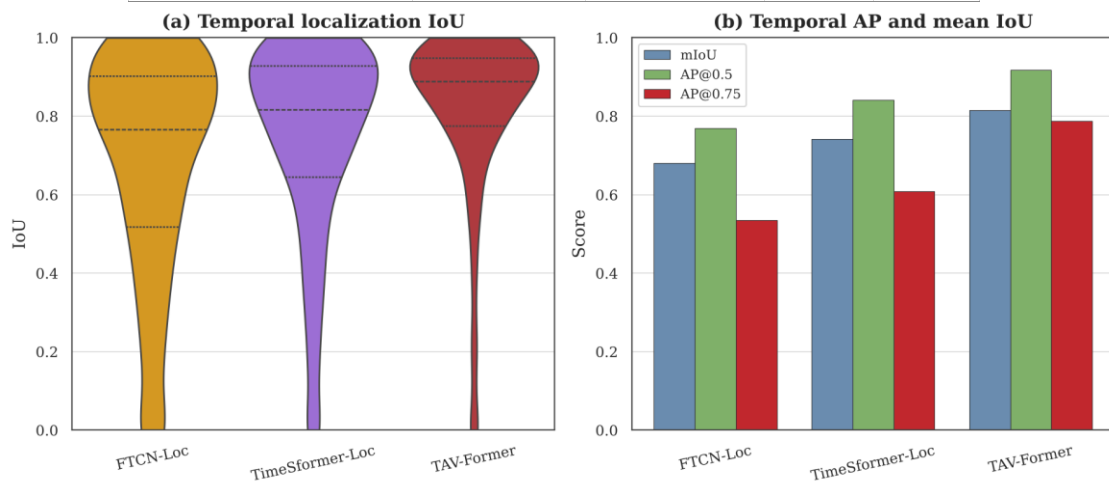


Fig. 8. Temporal localization: (a) per-clip IoU distributions; (b) mean IoU and AP@IoU.

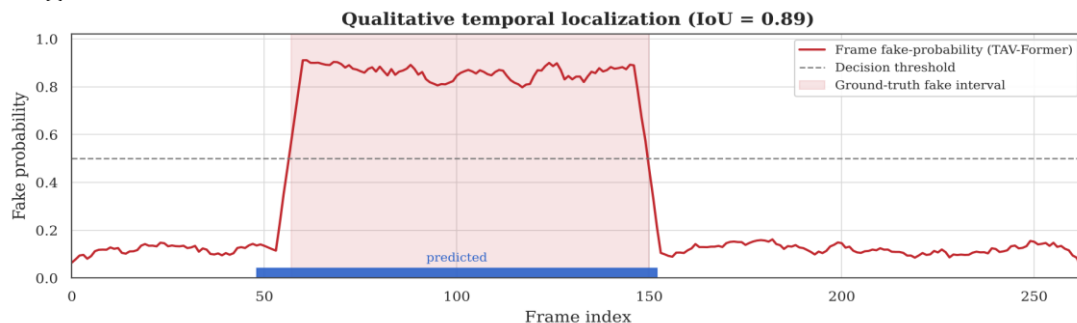


Fig. 11. Qualitative temporal localization: frame-level fake probability with ground-truth (shaded) and predicted interval.

5.7 Robustness to real-world perturbations

Figure 7 compares the change in detection AUC as compression, downsampling, blur,

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

and Gaussian noise increase. TAV-Former shows a smaller AUC reduction than ViT-B/16 under compression, downsampling, and blur, with $p < 0.001$ in each comparison. Under Gaussian noise, the degradation is similar for the two models: the difference in degradation is -0.004 and is not significant ($p = 0.35$). The robustness results therefore support H4 for compression, downsampling, and blur, but not for heavy Gaussian noise. **TABLE 7.** Robustness: mean AUC degradation (drop from clean) of spatial-only vs. temporal model, with paired-t significance.

Perturbation	Spatial drop	Temporal drop	Significance
JPEG/video compression	0.074	0.029	$p < 0.001$ ***
Downsampling	0.072	0.035	$p < 0.001$ ***
Gaussian blur	0.060	0.034	$p < 0.001$ ***
Gaussian noise	0.062	0.066	$p = 0.35$ n.s.

Robustness to common real-world perturbations: temporal (ours) vs spatial-only

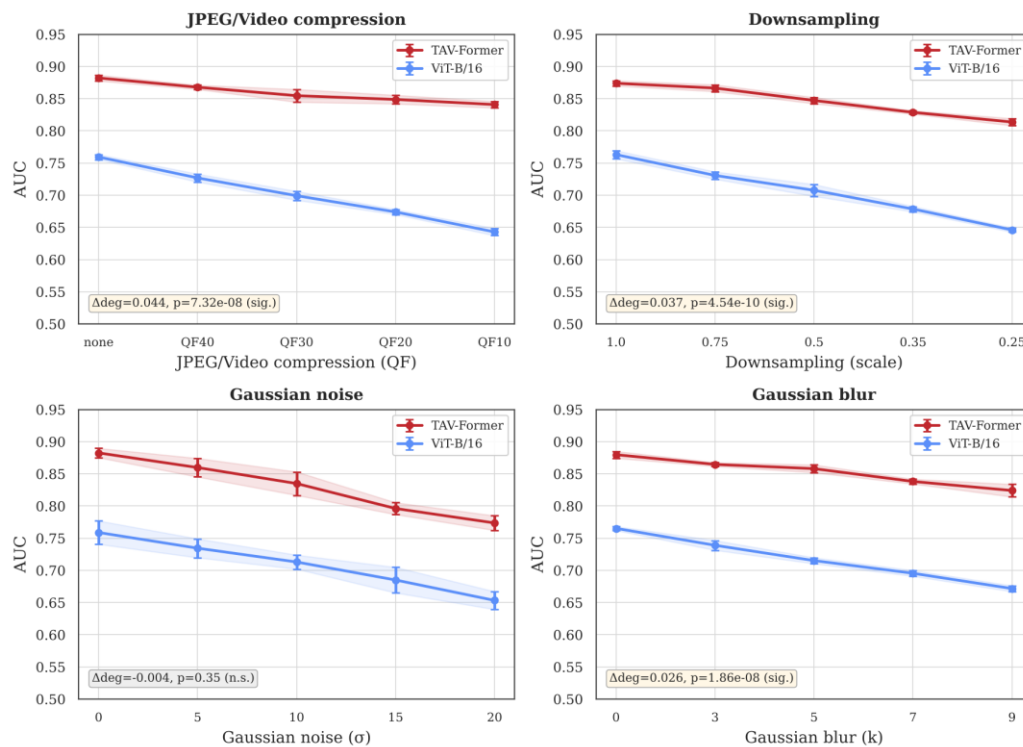


Fig. 7. Robustness to four perturbations; shaded bands are ± 1 s.d. over runs. Gaussian noise shows no significant robustness gap.

5.8 Feature space and efficiency

Figure 9 compares the learned video-level feature spaces for ViT-B/16 and TAV-Former. The TAV-Former representation shows clearer separation between real and manipulated clips in the pooled cross-domain set. Table 8 summarizes model size and inference speed. TAV-Former has 94.2 M parameters and 142.5 GFLOPs and processes approximately 31 clips/s in the reported hardware configuration. It is lighter and faster than TimeSformer in this comparison, while requiring more computation than the

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

frame-level CNN baselines.

Learned video-level feature space (t-SNE, pooled cross-domain test set)

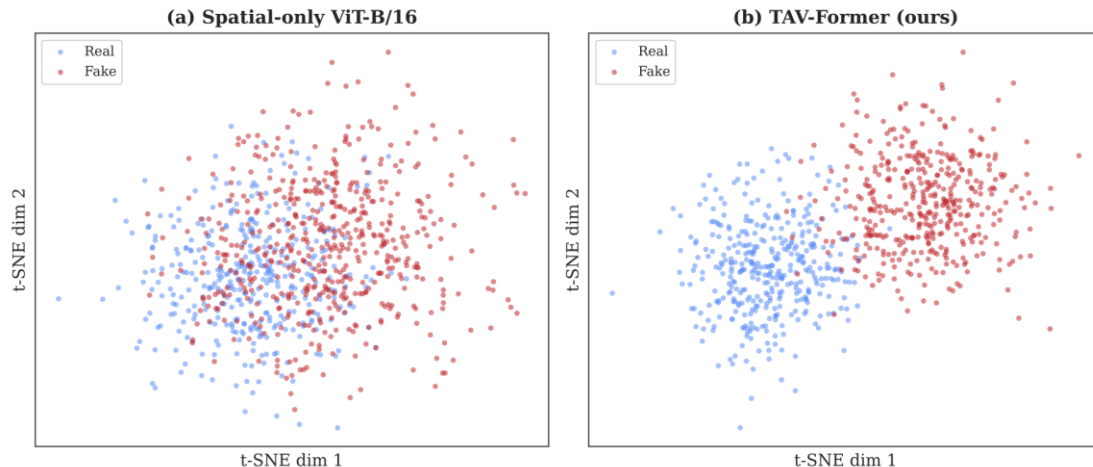


Fig. 9. *t-SNE of video-level features (pooled cross-domain): (a) spatial-only ViT-B/16; (b) TAV-Former.*

TABLE 8. Efficiency. Parameters and FLOPs are model characteristics; clips/s and latency are measured per clip on the reported GPU configuration.

Method	Params (M)	FLOPs (G)	Clips/s	Latency (ms)
Xception	20.8	8.4	142	7.0
EfficientNet-B4	19.3	4.2	118	8.5
ViT-B/16	86.6	17.6	96	10.4
CNN+LSTM	27.5	23.1	61	16.4
FTCN	26.6	29.7	48	20.8
TimeSformer	121.4	196.0	19	52.6
TAV-Former	94.2	142.5	31	32.3

6 Discussion

These results for RQ1 confirm the influence of temporal modelling towards cross-dataset generalisation. TAV-Former achieves a cross-domain AUC of 0.877, outperforming ViT-B/16 (AUC of 0.739) and TimeSformer (AUC of 0.837). Without temporal attention, AUC drops to 0.738, proving that the temporal attention is not solely due to the transformer backbone. The first part of this figure shows the support for H1, against the spatial baseline, while the second part displays the support for H2, where the reduction in the in-domain to cross-domain performance is smaller when compared to TAV-Former. Comparing it to TimeSformer is a bit more modest and should not be read as a general superiority over any dataset.

As it can be seen in the comparison with TimeSformer, this is quite important because cross-dataset significance should be reported on a per-protocol basis. The difference is statistically significant, but not for WildDeepfake or DF40. The findings thus suggest an overall favorability of TAV-Former and indicate that, for some cases, a good

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

temporal baseline works as well.

The best localization performance is obtained for RQ2 with the mean IoU at 0.815 and mean AP@0.5 at 0.918. The performance of Spatial localization is better on average, but not always with respect to all the protocols, compared to Grad-CAM. The robustness experiments also demonstrate that temporal modeling gives better resistance to compression, downsampling and blur while heavy Gaussian noise does not. The results here indicate that the value of the temporal information is related to whether the degradation maintains the temporal structure used by the model.

The next limitation of temporal modeling comes from the generator level analysis. The advantage of TAV-Former is less significant for the text-to-video task compared to the conventional categories of face generation and identity-swap and reenactment, which are more difficult tasks. These generators can generate temporally coherent sequences, which means that there is less inconsistency between frames for the detector to choose from. Future detectors may thus require that other signals/detectors, such as physical consistency, provenance or generator specific features, be used in conjunction with the temporal evidence.

Based on the efficiency results, TAV-Former is suitable for forensic triage in near real time or offline mode rather than high throughput inline filtering. In a realistic deployment, TAV-Former could be used to select clips for localization, then a light frame-level model could be used to conduct a preliminary screening. The confusion-matrix analysis also suggests that the operating threshold should be set for a specific application, and not as a fixed value of 0.5. Selection of thresholds should be done in conjunction with human examination in forensic contexts where there can be both false positives and false negatives with meaningful consequences.

The observed reduction in the performance of spatial models across the different domains further highlights the importance of testing spatial models on unseen datasets. In this study, the spatial transformer decreases from 0.984 in-domain AUC to 0.739 cross-domain AUC, whereas TAV-Former decreases from 0.994 to 0.877. The size of this difference justifies the use of cross-dataset evaluation as a primary metric for generalization, as opposed to in-domain accuracy.

6.1 Ethical considerations and responsible use

The design of deepfake detection is a dual use research field as failure mode analysis can also yield information about failure modes that an adaptive adversary may exploit. This study thus presents examples of both successful and unsuccessful, including the weaker performance in the presence of a high level of Gaussian noise and the challenges of very coherent AI-generated video. Model outputs should be considered as probabilistic evidence, not as a definitive prediction, for deployment. While localization can assist in managing human review, it does not supplant the need for independent review. Although the experiments are based on video data from benchmarks, not collected for the purpose of the experiments, consent, privacy and governance issues should still be considered when processing the faces of new individuals for future deployment studies.

7 Limitations

The main drawback is relying on the benchmark composition and the domain shift

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

associated with it. Experiments are conducted with real benchmark video data but the chosen datasets don't cover all possible generator types, camera conditions, compression pipeline, or types of AI-generated video found in the real world. The AIGC-Video protocol is especially sensitive to changes in the generation landscape, so results should be taken into account based on the generators evaluated in the set of evaluation. Secondly, the framework is face-centric and clip based, which is not applicable to full-scene manipulations and very long videos yet. Third, the study considers perturbations other than adaptive attacks that are aimed at defeating temporal cues. Last, spatial localization is not as consistent as temporal localization, meaning the current representation does not provide a uniformly reliable estimate of the edited-region across each data set.

8 Conclusion and Future Work

We introduced a temporal vision transformer for joint video deepfake detection and temporal/spatial localization named TAV-Former. The model outperforms all the compared models in seven evaluation protocols, and is significantly better than the spatial transformer baseline. The studies of localization and robustness demonstrate the benefits for temporal localization and for degradation factors like compression, downsampling and blur, respectively; the ablation study confirms that temporal attention is the most significant factor of this improvement. The results reveal the presence of significant gaps: spatial localization is less consistent, heavy Gaussian noise eliminates the robustness gain, and highly coherent AI-generated video is still challenging. Future research needs to include evaluation of newer generators, more extensive in-the-wild video collections, full scene manipulations, audio-visual cues, provenance signals, and adaptive adversarial settings.

References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. NeurIPS, 2017.
- [2] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.
- [3] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," in Proc. IEEE/CVF ICCV, 2021, pp. 6836–6846.
- [4] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in Proc. ICML, 2021.
- [5] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin Transformer," in Proc. IEEE/CVF CVPR, 2022, pp. 3202–3211.
- [6] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF ICCV, 2019, pp. 1–11.
- [7] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in Proc. IEEE/CVF CVPR, 2020, pp. 3207–3216.
- [8] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake Detection Challenge (DFDC) dataset," arXiv:2006.07397, 2020.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- [9] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, “DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection,” in Proc. IEEE/CVF CVPR, 2020, pp. 2889–2898.
- [10] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, “WildDeepfake: A challenging real-world dataset for deepfake detection,” in Proc. ACM Int. Conf. Multimedia, 2020, pp. 2382–2390.
- [11] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, J. Zhu, D. Luo, L. Yuan, C. Wang, S. Ding, and Y. Wu, “DF40: Toward next-generation deepfake detection,” in Proc. NeurIPS, 2024.
- [12] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, “DeepfakeBench: A comprehensive benchmark of deepfake detection,” in Proc. NeurIPS Datasets and Benchmarks, 2023.
- [13] Y. Mirsky and W. Lee, “The creation and detection of deepfakes: A survey,” ACM Computing Surveys, vol. 54, no. 1, pp. 1–41, 2021.
- [14] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, “Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward,” Applied Intelligence, vol. 53, pp. 3974–4026, 2023.
- [15] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A compact facial video forgery detection network,” in Proc. IEEE WIFS, 2018, pp. 1–7.
- [16] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proc. IEEE CVPR, 2017, pp. 1251–1258.
- [17] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. ICML, 2019, pp. 6105–6114.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE CVPR, 2016, pp. 770–778.
- [19] D. Güera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in Proc. IEEE AVSS, 2018, pp. 1–6.
- [20] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, “Recurrent convolutional strategies for face manipulation detection in videos,” in Proc. IEEE/CVF CVPRW, 2019.
- [21] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, “Thinking in frequency: Face forgery detection by mining frequency-aware clues,” in Proc. ECCV, 2020, pp. 86–103.
- [22] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, “Multi-attentional deepfake detection,” in Proc. IEEE/CVF CVPR, 2021, pp. 2185–2194.
- [23] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, “Face X-ray for more general face forgery detection,” in Proc. IEEE/CVF CVPR, 2020, pp. 5001–5010.
- [24] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, “Lips don’t lie: A generalisable and robust approach to face forgery detection,” in Proc. IEEE/CVF CVPR, 2021, pp. 5039–5049.
- [25] A. Haliassos, R. Mira, S. Petridis, and M. Pantic, “Leveraging real talking faces via self-supervision for robust forgery detection,” in Proc. IEEE/CVF CVPR, 2022, pp. 14950–14962.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- [26] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in Proc. IEEE/CVF ICCV, 2021, pp. 15044–15054.
- [27] Z. Wang, J. Bao, W. Zhou, W. Wang, and H. Li, "AltFreezing for more general video face forgery detection," in Proc. IEEE/CVF CVPR, 2023, pp. 4129–4138.
- [28] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in Proc. IEEE/CVF CVPR, 2022, pp. 4113–4122.
- [29] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in Proc. IEEE/CVF CVPR, 2022, pp. 18720–18729.
- [30] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," in Proc. IEEE/CVF CVPR, 2023, pp. 3994–4004.
- [31] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, and L. Ma, "Delving into the local: Dynamic inconsistency learning for DeepFake video detection," in Proc. AAAI, 2022, pp. 744–752.
- [32] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang, "ISTVT: Interpretable spatial-temporal video transformer for deepfake detection," IEEE Trans. Information Forensics and Security, vol. 18, pp. 1335–1348, 2023.
- [33] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and vision transformers for video deepfake detection," in Proc. ICIAP, 2022, pp. 219–229.
- [34] Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, "DeepFake detection algorithm based on improved vision transformer," Applied Intelligence, vol. 53, pp. 7512–7527, 2023.
- [35] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," IEEE Trans. Pattern Analysis and Machine Intelligence, 2020.
- [36] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in Proc. IEEE/CVF CVPR, 2023, pp. 24480–24489.
- [37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF CVPR, 2022, pp. 10684–10695.
- [38] A. Blattmann et al., "Stable Video Diffusion: Scaling latent video diffusion models to large datasets," arXiv:2311.15127, 2023.
- [39] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in Proc. NeurIPS, 2022.
- [40] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in Proc. IEEE/CVF ICCV, 2023, pp. 7346–7356.
- [41] D. S. Vahdati, T. D. Nguyen, A. Azizpour, and M. C. Stamm, "Beyond deepfake images: Detecting AI-generated videos," in Proc. IEEE/CVF CVPRW, 2024.
- [42] X. Song, X. Guo, J. Zhang, Q. Li, L. Bai, X. Liu, G. Zhai, and X. Liu, "On learning multi-modal forgery representation for diffusion generated video detection," in

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- Proc. NeurIPS, 2024.
- [43] Z. Cai, S. Ghosh, A. P. Dhall, T. Gedeon, K. Stefanov, and M. Hayat, “AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset,” in Proc. ACM Int. Conf. Multimedia, 2024.
 - [44] L. Dragar, B. Batagelj, P. Peer, V. Štruc, and Ž. Emeršič, “W-TDL: Window-based temporal deepfake localization,” in Proc. ACM Multimedia Workshops (1M-Deepfakes Detection Challenge), 2024.
 - [45] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, “Not made for each other—Audio-visual dissonance-based deepfake detection and localization,” in Proc. ACM Int. Conf. Multimedia, 2020, pp. 439–447.
 - [46] Y. Zhou and S.-N. Lim, “Joint audio-visual deepfake detection,” in Proc. IEEE/CVF ICCV, 2021, pp. 14800–14809.
 - [47] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, “Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain,” in Proc. IEEE/CVF CVPR, 2021, pp. 772–781.
 - [48] Y. Luo, Y. Zhang, J. Yan, and W. Liu, “Generalizing face forgery detection with high-frequency features,” in Proc. IEEE/CVF CVPR, 2021, pp. 16317–16326.
 - [49] Z. Yan, Y. Luo, S. Lyu, Q. Liu, and B. Wu, “Transcending forgery specificity with latent space augmentation for generalizable deepfake detection,” in Proc. IEEE/CVF CVPR, 2024.
 - [50] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, “Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection,” in Proc. IEEE/CVF CVPR, 2024.