FACIAL REGION-BASED ENSEMBLING FOR UNSUPERVISED TEMPORAL DEEPFAKE LOCALIZATION

Nesryne Mejri[1], *Pavel Chernakov*[1], *Polina Kuleshova*[1], *Enjie Ghorbel*[1,2], *Djamila Aouada*[1]

Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg[1], Luxembourg
Cristal Laboratory, National School of Computer Sciences, University of Manouba[2], Tunisia
firstname.lastname@uni.lu

ABSTRACT

This paper addresses the challenge of temporal deepfake localization. Instead of classifying entire videos as real or fake, the goal is isolating forged frames in untrimmed videos that might be partially manipulated. Recently, few deepfake localization methods have emerged. They are mostly supervised, therefore relying on costly annotations and suffering from a lack of generalization to unseen manipulations. As an alternative, we propose reformulating deepfake localization as an unsupervised time-series anomaly detection problem. Hence, to investigate the relevance of the proposed formulation, recent state-of-the-art techniques in anomaly detection for time-series are evaluated in the context of deepfake localization. To avoid using large architectures, geometric representations, e.g., facial landmarks, are used as input. Moreover, a facial-region based ensembling strategy is introduced for a better modelling of localized deepfake artifacts. Experiments performed on the ForgeryNet dataset demonstrate the effectiveness of the proposed ensembling method and highlight the suitability of the suggested formulation.

Index Terms— Unsupervised Anomaly Detection, Time-series, Temporal Deepfake Localization.

1. INTRODUCTION

The rise of deepfake technology, involving the creation of realistic facial media using Deep Neural Networks (DNN), calls into question the credibility of digital content [1, 2]. One major risk is the misuse of these manipulated data for spreading misinformation. Consequently, the development of effective deepfake detection methods has become crucial. Current deepfake detection strategies [3, 4] generally rely on binary classification, focusing on the prediction of one label for an entire video. Hence, these approaches simplify the problem by assuming that forged videos are temporally segmented. This, however, hinders their application in a real-world scenario, especially if real-time performances are required. A more plausible approach would be to localize deepfakes in an untrimmed video stream that can be locally forged. Recently, few methods have been proposed for temporal localization [5, 6, 7, 8]. The latter are trained in a supervised manner, thereby inheriting two major shortcomings. First, a large set of annotated data is needed, which can be costly and hard to obtain. Second, as discussed in [3, 9], overfitting issues can occur causing a poor generalization to unseen manipulations.

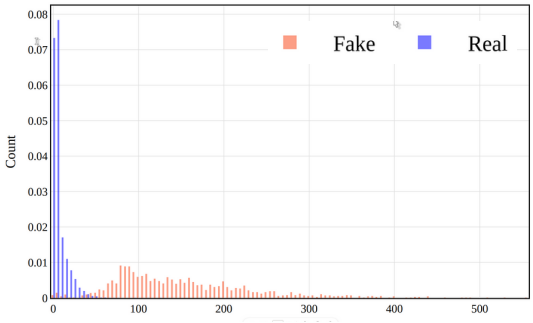

Fig. 1. Normalized histograms of the standard deviation of landmark displacements extracted from fake and real videos from ForgeryNet [6].

Inspired by recent works on unsupervised deepfake detection [10, 9], we propose to reformulate the problem of deepfake localization as an unsupervised anomaly detection problem in multivariate time-series. In other words, we suggest learning a time-series model using only real videos and considering out-of-distribution frames as deepfakes at inference. Specifically, we represent each video by the position trajectories of facial landmarks. These trajectories can be seen as a multivariate time-series, which can be prone to temporal inconsistencies in the case of forged videos. As experimentally demonstrated in Fig. 1, a noticeable discrepancy exists between the standard deviation of landmark displacements of real and fake videos. Furthermore, such a geometric representation has the advantage of being low-dimensional, resulting in more compact models. It is also universal across all datasets as it is robust to illumination changes and image content, thereby reducing overfitting risks. Hence, we propose

This work is supported by the Luxembourg National Research Fund, under the BRIDGES2021/IS/16353350/FaKeDeTeR and UNFAKE, ref.16763798 projects, and by Post Luxembourg.

979-8-3503-9015-5/24/$31.00 ©2024 IEEE

to study the suitability of recent time-series anomaly detection for the concrete use case of landmark-based deepfake localization. Furthermore, we propose a simple, yet effective, region-based ensembling strategy for deepfake localization relying on autoencoder (AE) architectures. Extensive experiments and analysis demonstrate the relevance of the proposed formulation as well as the introduced ensembling methods, suggesting a promising direction for future research in deepfake temporal localization.

In short, our contributions can be summarized as follows: (1) The formulation of temporal deepfake localization as an unsupervised anomaly detection problem in time-series. (2) An ensemble of lightweight autoencoders focusing on facial regions, trained only on real videos. (3) A comprehensive analysis and comparison of recent anomaly detection techniques on time-series in the context of deepfake localization.

This paper is structured as follows: Section 2 formulates temporal deepfake localization as a time-series anomaly detection problem. Section 3 details the proposed region-based ensembling approach. Section 4 describes the experiments and analyzes the results. Finally, Section 5 concludes this work.

2. UNSUPERVISED ANOMALY DETECTION IN TIME-SERIES FOR DEEPFAKE LOCALIZATION USING GEOMETRIC REPRESENTATIONS

An untrimmed video $\mathbf{V}$ can be defined as a temporally-ordered sequence of T images denoted as $\mathbf{V} = \{\mathbf{I}_t\}_{1 \leq t \leq T}$ with $\mathbf{I}_t \in \mathbb{R}^{h \times w \times c}$ and h, w and c being the height, width and the number of channels of $\mathbf{I}_t$, respectively. We assume that $\mathbf{l} = \{l_t\}_{1 \leq t \leq T}$ corresponds to the ground-truth label vector of $\mathbf{V}$, with $l_t \in \{0, 1\}$ representing the label of $\mathbf{V}$ at an instant t. Note that $l_t = 1$ in the presence of a forgery and $l_t = 0$ otherwise. We denote by $\mathbf{V}_t$ the subsequence of $\mathbf{V}$ formed by $(\mathbf{I}_{t-\tau_1}, \mathbf{I}_{t-\tau_1+1}, ..., \mathbf{I}_t, ..., \mathbf{I}_{t+\tau_2-1}, \mathbf{I}_{t+\tau_2})$ with $\tau_1$and τ_2 two integers defining the position and the size of a sliding window and T_s being its length. The goal of deepfake localization is estimating a function $f : \mathbb{R}^{h \times w \times c \times T_s} :\rightarrow \{0, 1\}$ for all t,

$$f(\mathbf{V}_t) = l_t. \tag{1}$$

Existing localization methods mostly learn f in a supervised manner using deep learning architectures [6, 7, 8]; thereby relying on costly annotations. Moreover, supervision leads to a lack of generalization to unseen manipulations, as this was demonstrated in the context of deepfake detection [3, 9]. To address this issue, we propose reformulating the problem of deepfake localization as an unsupervised anomaly detection task. Thus f can be viewed as a composition of two functions $f = \Phi \circ \Psi$ where $\Psi : \mathbb{R}^{h \times w \times c \times T_s} \rightarrow \mathcal{X}$ models normal time-series and is learned using only real data and $\Phi : \mathcal{X} \rightarrow \{0, 1\}$ is a thresholding function only used at inference.

Another aspect that should be considered is the model size. In fact, existing multivariate time-series anomaly detection architectures have been initially designed for relatively low dimensional data [11]. Hence, directly modelling videos as time-series might results in cumbersome models. As a solution, we propose the use of geometric representations, e.g., 2D facial landmarks. As shown in [12], in the context of deepfake detection, they can be used for obtaining more compact models, while demonstrating more robustness to illumination changes and noise. In other words, Ψ can be defined as $\Psi = \Psi_2 \circ \Psi_1$ such that $\Psi_1 : \mathbb{R}^{h \times w \times c \times T_s} \rightarrow \mathbb{R}^{2 \times n \times T_s}$ maps a video subsequence to its corresponding 2D facial landmark subsequence and $\Psi_2 : \mathbb{R}^{2 \times n \times T_s} \rightarrow \mathcal{X}$.

3. FACIAL REGION-BASED ENSEMBLING FOR UNSUPERVISED DEEPFAKE LOCALIZATION

As unsupervised anomaly detection for time-series approaches were not originally proposed for deepfake temporal localization, they do not explicitly focus on artifact-prone facial regions. Those regions, however, have been proven to be extremely effective in capturing deepfake artifacts [13, 9] from different deepfake generation methods. Hence, as illustrated in Fig. 2, we propose an ensembling strategy of different models that are focused on localized regions. For that purpose, we train an ensemble of K autoencoders, each one trained on a subset of landmarks belonging to a manually-selected facial region such as the mouth or the nose. Then, a voting strategy is applied for building the final predictions.

More specifically, given an input video $\mathbf{V}_t \in \mathbb{R}^{h \times w \times c \times T_s}$ processed as a 2D landmark sequence denoted by $\mathbf{X}_t = \Psi_1(\mathbf{V}_t) \in \mathbb{R}^{2 \times n \times T_s}$, we select K specific regions. We denote the position of the set of the n_k landmarks belonging to the region of index $k \in \{1, ..., K\}$ as $\mathbf{X}_t^k =\in \mathbb{R}^{2 \times n_k \times T_s}$. For each $k \in \{1, ..., K\}$, an autoencoder that aims at learning the distribution of authentic region-specific landmark trajectories is considered. To this end, given an encoder $\text{Enc}_k(.)$ and a decoder $\text{Dec}_k(.)$, for all windowed sequences $\mathbf{X}_t^k$, our model is trained as,

$$\begin{cases} \mathbf{z} = \text{Enc}_k(\mathbf{X}_t^k), \\ \hat{\mathbf{X}}_t^k = \text{Dec}_k, \end{cases} \tag{2}$$

with $\mathbf{z} \in \mathbb{R}^{T_s \times d}$ being the d-dimensional latent representation. The learning is optimized using the mean squared distance formulated as,

$$\mathcal{L}_r = \frac{1}{N_b} \sum_{t=0}^{N_b} ||\hat{\mathbf{X}}_t^k - \mathbf{X}_t^k||_2^2, \tag{3}$$

with N_b being the total batch samples and $|| \cdot ||_2$ denoting the L2-norm. Similarly to [14, 15], a statistical model termed Peak Over Threshold (PoT) [16] is used to automatically select an adequate threshold λ based on the training sequences. Such an approach identifies a suitable value at risk by fitting the distribution of the training data with a Generalized Pareto Distribution. Hence, given λ and a window $\mathbf{X}_t^k$, the predic-

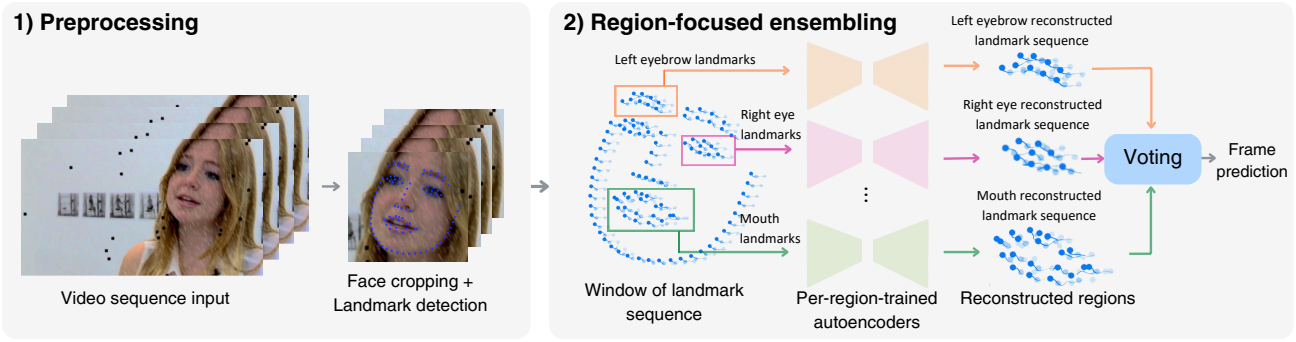

Fig. 2. Overview of the proposed facial region-focused ensembling.

tion l_t^k associated with the frame t is expressed as,

$$l_t^k = \mathbb{1}\left(\frac{1}{2 \times n_k \times T_s} \sum_{i}^{2n_k} \sum_{j}^{T_s} ||\hat{\mathbf{X}}_{t,i,j}^k - \mathbf{X}_{t,i,j}^k||_2^2 > \lambda\right). \quad (4)$$

with $\mathbb{1}(.)$ being an indicator function. Finally, given the K predictions l_t^k from different autoencoders, the final prediction is built via the soft majority voting rule.

4. EXPERIMENTS

4.1. Experimental settings

4.1.1. Dataset

We used the ForgeryNet dataset [6], a comprehensive benchmark for temporal forgery localization. It is formed by 2,896,062 images and 221,247 partially manipulated videos. Nevertheless, we select[1] only the data that contain one person per video. In total, we consider 9866 real videos from the official training set and 1,516 videos from the validation set for testing (as annotations for the test set are not yet available). Note that our test set comprises six different forgery methods.

4.1.2. Baselines

In addition to the proposed ensembling approach, we evaluate seven recent anomaly detection methods for time-series. As discussed in [17], these approaches adopt different learning paradigms. First, four reconstruction-based methods are considered, namely **TranAD** [15], **USAD** [18], **OmniAnomaly** [14] and **MAD-GAN**. TranAD and USAD are respectively based on transformer and AE architectures that are trained adversarially. OmniAnomaly uses a stochastic Recurrent Neural Network (RNN) and a planar normalizing flow to generate reconstruction probabilities. Finally, MAD-GAN is a GAN-inspired approach using an RNN as a base model for modeling spatio-temporal dependencies. Second, forecasting approaches are also tested for the use case of deepfake localization, including **CAE-M** [11], **DAGMM** [19] and **GDN** [20]. CAE-M feeds a reconstruction error and the learned feature representations to an auto-regressive network that predicts future feature values. DAGMM constrains the feature space to follow Gaussian mixture model distribution. Then, an RNN is employed for predicting a future data point. Last but not least, GDN models the relationships between data features as a graph coupled with an attention mechanism. For comparing with supervised deepfake localization methods [6, 8, 7, 5], only **MDS** [5], a multimodal technique with decoupled audio-video networks, is compatible with our setting. It maximizes the similarity of real audio and real visual features and minimizes it otherwise. The other baselines either require audiovisual input data or are not accessible.

[1]The processed sequences will be released.

4.1.3. Evaluation metrics

For evaluating the proposed ensembling strategy as well as state-of-the-art methods, we report the following metrics: the standard Precision, Recall, and F1-score metrics. Note that the results are reported with and without the Point Adjustment protocol, referred to as (**PA**) and (**non-PA**), respectively. The PA protocol proposed in [21] is commonly used for evaluating unsupervised anomaly detection in time-series [15, 18, 14]. It assumes that if a single point within an anomalous segment is detected, then the entire segment is correctly predicted as anomalous. Furthermore, we compute the t-Precision, the t-Recall, and the t-F1-scores [22], which are metrics tailored for time-series by taking into account factors like the location of detected anomalies and the cumulative overlap between predicted and ground-truth segments. Finally, similar to deepfake detection methods, we also report the Area Under the Curve (AUC) metric. In all our experiments, **Bold** and underline report the best and second best results, respectively.

4.1.4. Implementation details

For each video, we detect and crop the faces. Then, we extract from each frame a total of 98 landmarks using SPIGA [23]. The landmark values are normalized between 0 and 1. The average lengths of training and testing sequences are respectively equal to 160 and 119 frames. We use the same au-

	Method	AUC	Precision	Recall	F1-score
Sup.	MDS [5]	0.4943	0.4704	0.6931	0.5604
Unsup.	TranAD [15]	0.7177	0.8018	0.4878	0.6066
	USAD [18]	0.7779	0.8330	0.5000	0.6046
	DAGMM [19]	0.7573	0.8314	0.5644	0.6724
	GDN [20]	0.7837	**0.8397**	0.6187	0.7125
	MAD-GAN [25]	0.8497	0.7980	0.7859	0.7919
	OmniAnomaly [14]	0.7068	0.7998	0.4642	0.5874
	CAE-M [11]	0.9182	0.8385	0.9130	0.8742
	Ours	**0.9302**	0.8090	**0.9538**	**0.8754**

Table 1. Results in terms of standard performance metrics on ForgeryNet under the PA protocol.

	Method	AUC	Precision	Recall	F1-score
Sup.	MDS [5]	0.4663	**0.5104**	**0.3325**	**0.4027**
Unsup.	TranAD [15]	0.4967	0.2757	0.0459	0.0787
	USAD [18]	0.5080	0.3657	0.0647	0.1100
	DAGMM [19]	0.5015	0.3155	0.0527	0.0904
	GDN [20]	0.5153	0.4097	0.0820	0.1367
	MAD-GAN [25]	0.5425	0.4629	0.1715	0.2503
	OmniAnomaly [14]	0.4933	0.2417	0.0370	0.0641
	CAE-M [11]	0.5314	0.4720	0.1222	0.1941
	Ours	**0.5491**	0.4597	0.1916	0.2704

Table 2. Results in terms of standard performance metrics on ForgeryNet under the non-PA protocol.

toencoder architecture as proposed in CAE-M [11] from this repository[2]. The models are trained 5 epochs, one sequence at a time on an NVIDIA RTX A4000 GPU. We use the AdamW [24] optimizer with a learning rate of 10^{-3} and weight decay of 10^{-5}.

4.2. Results

4.2.1. Performance using standard metrics

Table 1 and Table 2 report the obtained results in terms of Precision, Recall, and F1-score with and without the PA protocol, respectively. In the former, it can be noted that the proposed ensemble generally outperforms other approaches including the supervised baseline. This confirms the adequacy of following a region-based strategy for spatially modelling deepfake artifacts. It can also be seen that except CAE and MAD-GAN, most approaches unsupervised achieve comparable results. Hence, it remains unclear whether reconstruction-based on forecasting methods are more suitable for the complex scenario of deepfake localization. This might also suggest that both reconstruction and forecasting approaches are able to capture discrepancies. In the latter case, when the predictions are not adjusted, it can be observed that all approaches suffer from an expected significant performance drop. Nevertheless, our ensemble still surpasses unsupervised state-of-the-art methods, including CAE-M with an increase of 1.77% and 7.63% in terms of AUC and F1-score, against 1.2% and 0.12% under the PA protocol, respectively. In comparison with MDS, although they reach better precision, recall and F1-score under the non-PA protocol, we achieve a higher AUC of 93.02% and 54.91% with the PA and non-PA protocols respectively. This suggests that with our approach the forged and real frames are more separable than with the supervised baseline. Additionally, contrary to MDS, our method is trained using a single modality and does not require annotated deepfake data. Notably, MDS presents overfitting signs since it achieves significantly higher AUC under the in-dataset setting reported in [5], with more than 90% against 46.63% under the cross-dataset setting (see Table 2).

[2]https://github.com/imperial-qore/TranAD/

	Method	t-Precision	t-Recall	t-F1-score
Sup.	MDS [5]	0.3039	**0.3348**	0.2376
Unsup.	TranAD [15]	0.3424	0.0552	0.0815
	USAD [18]	0.4101	0.0763	0.1110
	DAGMM [19]	0.3730	0.0639	0.0937
	GDN [20]	0.4226	0.0933	0.1229
	MAD-GAN [25]	**0.5066**	0.2068	0.2239
	OmniAnomaly [14]	0.3130	0.0434	0.0657
	CAE-M [11]	0.4683	0.1546	0.2041
	Ours	0.4354	0.2362	**0.2706**

Table 3. Results in terms of range-based metrics (t-Precision, t-Recall and t-F1-score) proposed in [22] on ForgeryNet under the non-PA protocol.

4.2.2. Performance using range-based metrics

Table 4 and Table 3 report the obtained results on ForgeryNet in terms of range-based metrics including the t-Precision, the t-Recall, and the t-F1-score proposed in [22], under the PA and the non-PA protocols, respectively. It can be observed from Table 3 that the obtained results with range-based metrics are consistent with the standard metrics results shown in Table 2 and Table 1. In fact, the proposed ensemble achieves the highest t-F1-score, followed by MAD-GAN and CAE-M. This demonstrates the robustness of our strategy as compared to unsupervised state-of-the-art techniques, suggesting that it can detect consecutive anomalies rather than random point anomalies. However, in Table 4, we observe that our method is no longer the best performing. This can be explained by the fact that the adjustment harms our performance, by boosting the t-Recall at the expense of the t-Precision. Notably, PA does not always yield a reliable comparison. As shown in [26], it can boost the performance a random detector making it comparable to a well-trained one. Furthermore, the compatibility of this protocol with range-based metrics is debatable. In fact, by treating an anomalous segment and a single point equally, the temporal information that range-based metrics aim to capture, based on the predicted anomaly location and cumulative overlap, is dissipated. Hence, we report only non-PA results in the following experiments.

	Method	t-Precision	t-Recall	t-F1-score
Sup.	MDS [5]	0.2321	0.6521	0.3034
Unsup.	TranAD [15]	0.3357	0.4537	0.3653
	USAD [18]	0.3926	0.5582	0.4281
	DAGMM [19]	0.3564	0.5303	0.3959
	GDN [20]	0.3847	0.5852	0.4226
	MAD-GAN [25]	**0.4214**	0.7591	**0.4842**
	OmniAnomaly [14]	0.3039	0.4361	0.3364
	CAE-M [11]	0.3465	0.8635	0.4635
	Ours	0.2487	**0.9369**	0.3695

Table 4. Results using range-based metrics proposed in [22] on ForgeryNet under the PA protocol.

Facial regions	#Landmarks	AUC	Precision	Recall	F1-score
Pupils (P)	2	0.5186	0.4109	0.0916	0.1498
Left Eye (LE)	8	0.5182	0.4064	0.0923	0.1505
Right Eye (RE)	8	0.5060	0.3409	0.0602	0.1023
Eyes (E)	16	0.5216	0.4265	0.0976	0.1588
Left Brow (LB)	9	0.5257	0.4451	0.1065	0.1718
Right Brow (RB)	9	0.5115	0.3733	0.0756	0.1258
Brows (B)	18	**0.5346**	**0.4796**	**0.1257**	**0.1992**
Nose (N)	9	0.5104	0.3603	0.0786	0.1290
Mouth (M)	20	0.5092	0.3587	0.0710	0.1185
Jawline (J)	33	0.5296	0.4672	0.1121	0.1808

Table 5. Results using individual facial regions on ForgeryNet in terms of standard performance metrics under the non-PA protocol.

4.2.3. Selection of facial regions

Since we propose a facial region-focused ensemble, we report in Table 5 the performance of our AE trained on different facial regions. The best performance is achieved using the jawline and eyebrows models. This can be explained by the fact that during the blending stage, deepfake generation methods fail to perfectly align the foreground and background faces, resulting in noisy landmarks within those facial areas. It is also interesting to observe the mismatch between the left and right facial areas. Specifically, a difference in terms F1-score of 4.70% and 4.77% can be observed between the right and the left eyebrows, and the left and right eye respectively.

4.2.4. Role of the ensembling

Table 6 gives the obtained results by considering different combinations of the three most relevant regions. Mainly, we compare the simple concatenation of the region-based geometric features against the proposed ensemble strategy. It can be noted that the ensembling consistently enhances the performance as compared to the direct concatenation of region-based features within a single model. This might be explained by the fact that implicitly learning region-based features with a single model is challenging. As discussed in [27], capturing local artifacts using a CNNs is not straightforward as successive convolution layers tend to eliminate low-level features.

Brows	Eyes	Jawline	Ensembled	AUC	F1-score
	✓	✓	✓	**0.5388**	**0.2446**
	✓	✓		0.5309	0.1869
✓		✓	✓	**0.5491**	**0.2704**
✓		✓		0.5173	0.1497
✓	✓		✓	**0.5392**	**0.2368**
✓	✓			0.5301	0.1854
✓	✓	✓	✓	**0.5513**	**0.2913**
✓	✓	✓		0.5374	0.2088

Table 6. Feature combination versus ensembling strategy of the three most relevant facial regions under the non-PA protocol. Experiments are performed on ForgeryNet.

	Method	#Parameters
Sup.	MDS [5]	122777092
Unsup.	TranAD [15]	3197004
	USAD [18]	50109
	DAGMM [19]	50016
	GDN [20]	20074
	MAD-GAN [25]	48613
	OmniAnomaly [14]	38872
	CAE-M [11]	7229
	Ours	7229

Table 7. Number of model parameters

4.2.5. Model size

Finally, Table 7 reports the number of parameters of each method. It can be seen that our method as well as CAE-M have a significantly lower number of parameters in comparison to state-of-the-art techniques, including the supervised baseline MDS (with 7229 against 122777092 parameters).

5. CONCLUSION

In this paper, temporal deepfake localization has been formulated as an unsupervised time-series anomaly detection problem. To assess the suitability of the proposed formulation, state-of-the-art methods in the general field of time-series anomaly detection have been benchmarked under the complex scenario of deepfake localization. Instead of using raw videos, a geometric representation is used, namely, the trajectories of facial landmarks, enabling the use of relatively lightweight architectures. Furthermore, to better model localized artifacts, a facial region-based ensembling strategy has been introduced. The obtained results have not only demonstrated the relevance of the proposed formulation, but have also shown the superiority of the introduced ensembling method as compared to state-of-the-art techniques. However, our approach might be sensitive to compression and noisy landmark extraction. In future works, strategies for including textural information will be explored to overcome this issue. As such, this paper opens the ground for future investigations in the field of unsupervised deepfake localization.

6. REFERENCES

[1] Jane Wakefield, "Deepfake presidents used in russia-ukraine war," Accessed: 2023-12-05.

[2] Ali Breland, "The bizarre and terrifying case of the "deepfake" video that helped bring an african nation to the brink," Accessed: 2023-12-05.

[3] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proc. of ICCV*, 2019, pp. 1–11.

[4] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," in *Proc. of CVPR*, 2022, pp. 18710–18719.

[5] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian, "Not made for each other-audio-visual dissonance-based deepfake detection and localization," in *Proc. of ACMMM*, 2020, pp. 439–447.

[6] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu, "Forgerynet: A versatile benchmark for comprehensive forgery analysis," in *Proc. of CVPR*, 2021, pp. 4360–4369.

[7] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat, "Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization," in *DICTA*, 2022, pp. 1–10.

[8] Rui Zhang, Hongxia Wang, Mingshan Du, Hanqing Liu, Yang Zhou, and Qiang Zeng, "Ummaformer: A universal multimodal-adaptive transformer framework for temporal forgery localization," in *Proc. of ACMMM*, 2023, pp. 8749–8759.

[9] Nesryne Mejri, Enjie Ghorbel, and Djamila Aouada, "Untag: Learning generic features for unsupervised type-agnostic deepfake detection," in *ICASSP*, 2023, pp. 1–5.

[10] Chao Feng, Ziyang Chen, and Andrew Owens, "Self-supervised video forensics by audio-visual anomaly detection," in *Proc. of CVPR*, June 2023, pp. 10491–10503.

[11] Yuxin Zhang, Yiqiang Chen, Jindong Wang, and Zhiwen Pan, "Unsupervised deep anomaly detection for multi-sensor time-series signals," *IEEE TKDE*, 2021.

[12] Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia, "Improving the efficiency and robustness of deepfakes detection through precise geometric features," in *Proc. of CVPR*, 2021, pp. 3609–3618.

[13] Zihan Liu, Hanyi Wang, and Shilin Wang, "Cross-domain local characteristic enhanced deepfake video detection," in *Proc. of ACCV*, 2022, pp. 3412–3429.

[14] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proc. of ACM SIGKDD*, 2019, pp. 2828–2837.

[15] Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings, "TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data," *Proc. of VLDB*, vol. 15, no. 6, pp. 1201–1214, 2022.

[16] Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouet, "Anomaly detection in streams with extreme value theory," in *Proc. of ACM SIGKDD*, 2017, pp. 1067–1075.

[17] Nesryne Mejri, Laura Lopez-Fuentes, Kankana Roy, Pavel Chernakov, Enjie Ghorbel, and Djamila Aouada, "Unsupervised anomaly detection in time-series: An extensive evaluation and analysis of state-of-the-art methods," *arXiv preprint arXiv:2212.03637*, 2022.

[18] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga, "Usad: Unsupervised anomaly detection on multivariate time series," in *Proc. of ACM SIGKDD*, 2020, pp. 3395–3404.

[19] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *ICLR*, 2018.

[20] Ailin Deng and Bryan Hooi, "Graph neural network-based anomaly detection in multivariate time series," in *Proc. of AAAI*, 2021, vol. 35, pp. 4027–4035.

[21] Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, et al., "Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications," in *Proc. of WWW*, 2018, pp. 187–196.

[22] Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich, "Precision and recall for time series," *Adv. Neural Inf. Process.*, vol. 31, 2018.

[23] Andrés Prados-Torreblanca, José M Buenaposada, and Luis Baumela, "Shape preserving facial landmarks with graph attention networks," in *BMVC*. 2022, BMVA Press.

[24] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[25] Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng, "Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks," in *Proc. of ICANN*. Springer, 2019, pp. 703–716.

[26] Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon, "Towards a rigorous evaluation of time-series anomaly detection," in *Proc. of AAAI*, 2022, vol. 36, pp. 7194–7201.

[27] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu, "Multi-attentional deepfake detection," in *Proc. of CVPR*, 2021, pp. 2185–2194.