Supplementary Material: Learning Natural Consistency Representation for Face Forgery Video Detection

Daichi Zhang[1,2,3], Zihao Xiao[4], Shikun Li[1,2], Fanzhao Lin[1,2], Jianmin Li[3], and Shiming Ge[1,2,*]

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] Department of Computer Science and Technology, Institute for AI, BNRist, Tsinghua University [4] RealAI
{zhangdaichi20,lishikun19,linfanzhao21}@mails.ucas.ac.cn, lijianmin@mail.tsinghua.edu.cn, zihao.xiao@realai.ai, geshiming@iie.ac.cn

In this supplementary material, we provide more details of our intuition, related datasets, baselines, additional experimental results, and further ethical discussion on how our method will contribute to the forgery video detection field, described in detail as follows.

1 Intuition

Our method's intuition is based on the observation that fake videos struggle to preserve the natural consistency as real face videos (defined as the semantic-level spatiotemporal coherence in natural face videos, and the opposite as inconsistency, such as facial movements and expression changes) even under unknown generation methods and different corruptions, since current generation methods take less account of the spatiotemporal coherence and this high-level inconsistency is naturally robust against corruptions compared to low-level features. We aim to explore this representation by introducing natural consistency learning on only real face videos with two specifically designed self-supervised tasks. We provide samples from FF++ in Fig.1 of the manuscript to support our intuition, and further provide more samples that can't preserve the natural consistency from other four investigated datasets in Fig.1.

Based on the observation above, we further design two self-supervised tasks. Our major intuition is fake videos struggle to maintain natural spatiotemporal consistency in real videos and masking/shuffling real videos could expose the unnatural change to benefit natural consistency representation learning. As shown in Fig.2(a), we calculate the L_1 distance of each frame from real/fake and real(masked)/real(shuffled) clips to the first frame of each video in FF++, and find real clips exhibit less local "spikes" between frames than other three, which suggests real clips are more naturally consistent and supports our intuition. Note that we don't entirely separate the spatial and temporal consistency since they may be correlated for both real and fake videos. The representation optimized by our tasks contains both spatial and temporal clues extracted by

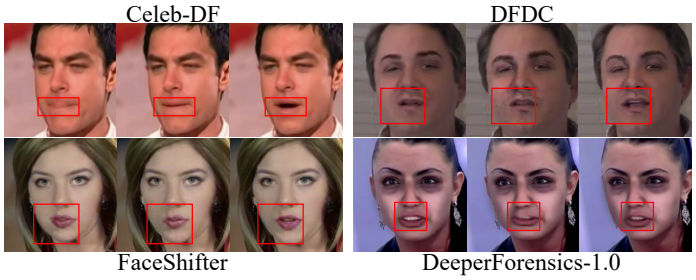

Fig. 1: Forgery videos from other four investigated complex datasets which still cannot preserve the natural consistency.

CNNs and Transformer. The two designed visual-only pretraining tasks (SPM and TCM) aim to learn natural consistency on unlabeled real videos in latent space. SPM leverages masking to mine the context of spatiotemporal clues in latent space, rather than spatial pixel space in MAE. TCM directly regularizes the distance between inconsistent shuffled clips and natural clips in latent space, which is different from [20] that exposes inconsistency by blending the shuffled clip with natural clip to generate augmented fake faces in pixel space. There are also some previous observations [17, 25, 35] supporting that shuffling could expose spatiotemporal inconsistency.

We employ each designed task for pretraining respectively then use t-SNE to visualize the representations of four different clips above in Fig. 2 (b), which suggests both pretraining tasks could learn natural spatiotemporal consistency in real videos to provide strong discriminability.

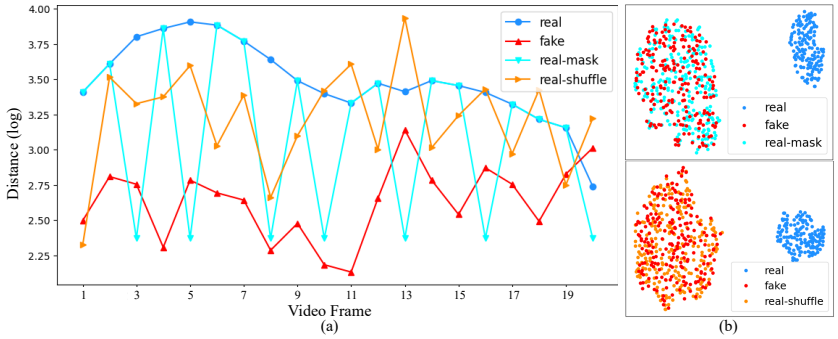

Fig. 2: Intuition(a) and ablation(b) for our pretraining tasks.

2 More Explanation for Generalization

We provide an intuitive explanation for the improved generalization of our model here: given a natural real clip $v = \{f_1, f_2, ..., f_n\}$ with n frames and a fake clip $v^* = \{f_1, ..., f_m^*, ..., f_n\}$ with a manipulated frame f_m^*, supervised detectors may easily overfit to forgery clue in f_m^* or inconsistency between $f_{m\pm k}$ and f_m^*, which are two patterns related to specific forgery methods and datasets. Our method instead focuses on the natural consistency between $f_{m\pm k}$ and f_m to learn general representations that discriminate all inconsistencies, avoiding overfitting to the previous two patterns and leading to improved generalization.

3 More Details about Datasets

Section 4.1 of the manuscript presents a brief description of the datasets we used. For better comprehension, we provide more details of the real face video datasets, VoxCeleb2 [4], and forgery video datasets, including FaceForensics++ (FF++) [27], Celeb-DF-v2 (CDF) [23], DeepFake Detection Challenge (DFDC) [7], FaceShifter (FSh) [18], and DeeperForensics (DFo) [16], as presented in Tab 1 which includes the total number of real and fake videos, the manipulation types to generate fake videos and where the source videos from. Besides, the FaceForensics++ dataset includes three different quality subsets, raw/c23/c40 (from uncompressed to heavily compressed). The DeeperForensics dataset introduces various perturbations to the videos to simulate real-world scenarios. For our fine-tuning stage, we follow the official data split if provided (such as FaceForensics++ and Celeb-DF), otherwise we randomly split the training, validation and testing subsets by 6:2:2.

Table 1: Details of real face video and forgery video datasets, including the total number of real and fake videos, the manipulation methods used to generate fake videos and where the source videos from.

Dataset	Real Videos	Fake Videos	Manipulation Types	Source Videos
VoxCeleb2 [4]	150,480	-	-	Youtube
FaceForensics++ [27]	1,000	5,000	Deepfakes/FaceSwap/Face2Face/NeuralTextures	consented actors
Celeb-DF [23]	590	5,639	improved DeepFake synthesis algorithm	Youtube
Deepfake Detection Challenge [7]	23,564	104,500	DFAE/NTH/MM-NN swap/FSGAN/StyleGAN	consented individuals
FaceShifter [18]	-	1,000	FaceShifter	FaceForensics++
DeeperForensics [16]	50,000	10,000	DeepFake Variational Auto-Encoder	consented actors

4 More Details about Baselines

We collect and conduct experimental comparisons with other 20 representative and state-of-the-art methods mentioned above in the manuscript and this supplementary material, including Xception [27], CNN-aug [32], Patch-based [2], Two-branch [24], Face X-ray [19], CNN-GRU [28], Multi-task [26], DSP-FWA [22], S-MIL-T [21], TD-3DCNN [36], STIL [11], HCIL [12], SBI+EB4 [30], PCL+I2G [39],

LipForensics [14], RealForensics [13], FTCN [40], ISTVT [38], NoiseDF [33] and AltFreezing [34]. The details of selected baselines are presented in Tab. 2, including the detection type (image or video-level), backbone (CNN or Transformer), training data modality (image, video or audio), and training scheme (supervised or self-supervised). The selected baselines include representative CNN detector such as Xception [27], spatiotemporal model such as CNN-GRU [28] and TD-3DCNN [36], transformer-based detector such as FTCN [40], supervised detector such as AltFreezing [34], and self-supervised detector such as RealForensics [13]. Additionally, it can be observed that our NACO is the only self-supervised video-level detector requiring video-only modality without label.

Table 2: Details of the collected state-of-the-art methods and our proposed method in type (image detector or video detector), backbone (CNN-based or Transformer-based), training data modality (image, video or audio), and scheme (supervised or self-supervised).

Method	Type	Backbone	Modality	Scheme
Xception [27]	Image	CNN	Image	Supervised
CNN-aug [32]	Image	CNN	Image	Supervised
Patch-based [2]	Image	CNN	Image	Supervised
Two-branch [24]	Video	CNN+LSTM	Video	Supervised
Face X-ray [19]	Image	CNN	Image	Self-supervised
CNN-GRU [28]	Video	CNN+RNN	Video	Supervised
Multi-task [26]	Video	CNN	Video(labeled)	Semi-supervised
DSP-FWA [22]	Video	CNN	Image	Self-supervised
S-MIL-T [21]	Video	CNN	Video	Supervised
TD-3DCNN [36]	Video	CNN	Video	Supervised
STIL [11]	Video	CNN	Video	Supervised
HCIL [12]	Video	CNN	Video	Supervised
SBI+EB4 [30]	Image	CNN	Image	Self-supervised
PCL+I2G [39]	Image	CNN	Image	Self-supervised
LipForensics [14]	Video	CNN	Video(labeled)	Self-supervised
RealForensics [13]	Video	CNN	Video+Audio	Self-supervised
FTCN [40]	Video	CNN+Transformer	Video	Supervised
ISTVT [38]	Video	Transformer	Video	Supervised
NoiseDF [33]	Video	CNN	Video	Supervised
AltFreezing [34]	Video	CNN	Video	Supervised
NACO (ours)	Video	CNN+Transformer	Video	Self-supervised

5 Additional Experimental Results

5.1 Intra-Dataset Evaluation

Follwing [13, 14], we investigate intra-dataset performance on FF++ under different video compression levels. Specifically, we fine-tune separate model for each

video quality, including: (1) uncompressed (raw), (2)slightly compressed (c23), and heavily compressed (c40). The results are presented in Tab. 3. We observe that frame-level detectors which focus on low-level clues significantly suffer from video compression, such as [19, 27]. This empirically demonstrates our motivation that these low-level clues such as artifacts or boundaries could be corrupted under common perturbations, such as compression. Other video-level detectors, such as [13, 14, 40] and our NACO, which utilize high-level spatiotemporal clues to detect suffer less under video compression. And our method achieves the best intra-dataset AUC score on all three different compression levels, which also indicates the effectiveness and superiority of our intuition and our designed method.

Moreover, we also make intra-dataset comparisons with other recent state-of-the-art detectors on Celeb-DF-v2(CDF) [23] and DFDC [7] datasets, as shown in Tab. 4. We can find that our method consistently outperforms other detectors on these two more challenging datasets, also indicating the effectiveness and superiority of our method.

Table 3: Intra-dataset evaluation. Video-level comparisons on FF++ when fine-tuned and tested on uncompressed (raw), slightly compressed (c23), and heavily compressed (c40) videos. Other methods' results are from [13].

Method	FF++ (raw)		FF++ (c23)		FF++ (c40)	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Xception [27]	99.0	99.8	97.0	99.3	89.0	92.0
CNN-aug [32]	98.7	99.8	96.9	99.1	81.9	86.9
Patch-based [2]	99.3	**99.9**	92.6	97.2	79.1	78.3
Two-branch [24]	-	-	-	99.1	-	91.1
Face X-ray [19]	99.1	99.8	78.4	97.8	34.2	77.3
CNN-GRU [28]	98.6	**99.9**	97.0	99.3	90.1	92.2
LipForensics [14]	98.9	**99.9**	98.8	99.7	94.2	98.1
FTCN [40]	-	-	**99.1**	**99.8**	-	98.3
RealForensics [13]	99.3	**99.9**	**99.1**	**99.8**	96.3	99.5
NACO (ours)	**99.5**	**99.9**	98.9	**99.8**	**97.0**	**99.6**

5.2 Ablation Study on Mask Ratio of SPM

We conduct further ablations to investigate the impact of mask ratios on our proposed SPM. Specifically, we set the random mask ratio α to $1/4$, $2/4$, and $3/4$ then evaluate the robustness and generalization under both settings. The results are presented in Tab. 5. We observe that both robustness and generalization decline when setting the ratio too low or high, *i.e.*, averagely 1.05% and 11.85% AUC drop respectively. This suggests that keeping or ignoring too much spatial information in SPM constrains the detecting performance, particularly for generalization. Therefore, the mask ratio should be chosen carefully for better natural consistency representation learning.

Table 4: Intra-dataset evaluation. Video-level comparisons on Celeb-DF(CDF) and DFDC. Other methods' results are from their original papers.

Method	CDF		DFDC	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)
S-MIL-T [21]	98.8	-	85.1	-
TD-3DCNN [36]	81.1	88.8	82.6	79.0
STIL [11]	**99.8**	-	89.8	-
HCIL [12]	**99.8**	-	95.1	-
STDT [37]	91.7	97.2	97.4	99.1
PCL+I2G [39]	-	**99.9**	-	94.4
SBI+EB4 [30]	-	93.7	-	88.7
NACO (ours)	**99.8**	**99.9**	**99.4**	**99.9**

Table 5: Effect of mask ratios in SPM. Comparisons of different mask ratios α in SPM under both robustness and cross-dataset settings.

Ratio	FF++ (c40)		CDF	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)
1/4	89.8	94.2	74.3	78.9
2/4	91.2	95.4	81.2	89.5
3/4	87.8	94.5	74.5	76.4

5.3 Ablation Study on Window Size of TCM

We conduct further ablation study on the temporal window size of the TCM module, as presented in Tab. 6, which also indicates that too-short or too-long video length would both decrease the performance and the drop is more significant with a small window size. Thus, the window size should also be chosen carefully for better natural consistency representation learning.

Table 6: Effect of temporal window size in TCM. Comparisons of different temporal window in TCM under both robustness and cross-dataset settings.

Size	FF++ (c40)		CDF	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)
10	89.1	93.3	77.6	82.9
20	91.2	95.4	81.2	89.5
40	91.1	94.7	80.8	88.4

5.4 Ablation Study on Backbone

We conduct further studies to demonstrate the advantages of our designed architecture in learning natural consistency representation. Specifically, we evaluate

two traditional CNNs [15] and ViTs [8] for learning spatiotemporal representation, and a proposed 3D CNN detector [36] to directly learn video representations from pixel space. The results are presented in Tab. 7. We observe that ViT outperforms CNNs in both settings, indicating the effectiveness of our architecture by using Transformer to learn the long-range spatiotemporal representation. Moreover, the performance using ViT is also better than using 3DCNN for learning video representation directly from pixel space, demonstrating our hypersisi that pixel space lacks crucial information for natural consistency representation learning and also suggesting the superiority of our architecture by incorporating CNNs for local receptive spatial features and Transformer for long-range spatiotemporal representation learning.

Table 7: Ablation on different backbones. Comparisons of different architectures on learning natural consistency from real videos.

Architecture	FF++ (c40)		CDF	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)
ResNet18	79.2	87.5	66.2	56.1
ResNet50	83.2	90.4	67.1	60.8
3DCNN	80.1	88.4	68.8	65.2
ViT-Small	81.7	87.2	72.8	73.6
ViT-Base	91.2	95.4	81.2	89.5

5.5 Ablation on Different Pretraining Dataset and Method

We conduct further experiments by using a larger pretraining dataset, *i.e.*, both VoxCeleb2 and Kinetics-400, as shown in Tab. 8, which indicates that a larger pretraining dataset can indeed improve the detection performance. Thus, considering the additional computational cost, a balance between the pretraining data scale and the detection performance may be essential.

Table 8: Evaluation with larger pretraining dataset. We report the video-level AUC (%) under both robustness and cross-dataset settings.

Pretraining Dataset	FF++ (c40)		CDF	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Kinetics-400	86.3	92.9	70.8	69.6
VoxCeleb2	91.2	95.4	81.2	89.5
VoxCeleb2 + Kinetics-400	91.4	95.5	82.1	90.2

Furthermore, following RealForensics and [20], we report the results in Tab. 9, where the performance slightly declines. This suggests the performance may be

related to both dataset size and data quality: small-scale LRW and FF++(real) may contain more useful information about natural face movement, while large-scale VoxCeleb2 may have more redundant information about face identity, consistent with the results of RealForensics pretrained on VoxCeleb2. The results also suggest competitive performance can be obtained by tuning on a different and smaller dataset.

Table 9: AUC(%) when pretrained on LRW and FF++(real).

Pretrain Dataset	CDF	DFDC	FSh	DFo	*Avg*
LRW	87.2	75.7	99.5	99.4	90.5
FF++(Real)	86.1	74.6	98.7	98.2	89.4
VoxCeleb2	89.5	76.7	99.4	99.5	91.2

Furthermore, to demonstrate the advantages of our pretraining method on face forgery detection task, we conduct further studies by using different video datasets (not face but action videos): Something-Something-V2(SS-V2) [10] and Kinetics-400 [1] and different method: VideoMAE [31] for pretraining in Tab.10. We observe other video datasets generally improve performance than no-pretraining, and other pretraining methods also lead to acceptable results, but the performance on face videos is better, and none of them match the robustness and generalization achieved by our method, indicating its superiority. It is also suggested that our pretrained models are beneficial for face forgery video detection task, which can leverage off-the-shelf resources.

Table 10: Comparisons with other pretraining methods and datasets when fine-tuned on FF++ (c23).

Method	Pretrain Dataset	FF++ (c40)		CDF	
		ACC	AUC	ACC	AUC
VideoMAE	SS-V2	80.2	88.5	66.8	65.6
	Kinetics-400	81.6	90.2	67.3	66.7
	VoxCeleb2	84.8	94.0	71.6	74.9
NACO (ours)	None	69.0	72.6	66.2	67.9
	SS-V2	84.8	91.1	69.1	69.3
	Kinetics-400	86.3	92.9	70.8	69.6
	VoxCeleb2	91.2	95.4	81.2	89.5

5.6 Forgery Localization

In Fig. 4 of our manuscript, we present Grad-CAM [29] visualization results on five consecutive frames from each subset of FF++ (c23) to investigate whether

our method can effectively localize the forgery traces in fake videos. This section presents more Grad-CAM examples on each subset of FF++ (c23), including real, Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). The additional results are illustrated in Fig. 3. We observe that our model can still effectively respond to the forgery traces and localize the inconsistencies in fake videos, such as the mouth and nose regions in F2F and FS, while the heatmap is average over the entire face area for real videos since there is no inconsistency. The results provide more support that our method can effectively localize the forgery traces in fake videos.

Fig. 3: Forgery localization. Additional Grad-CAM [29] results on consecutive frames of FF++ (c23), including real, Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). We find that our proposed method can effectively respond to the inconsistencies in fake videos and localize the forgery areas.

5.7 Further Discussions with SOTA Detectors

In addition to Tab. 1 of our manuscript, we make further comparisons with the recent SOTA RealForensics when using the same auxiliary dataset VoxCeleb2 as shown in Tab. 11. We observe that our method achieves better performance with an average 1.2% AUC improvement than RealForensics when using the same auxiliary dataset for pretraining, indicating the superiority of our method. Besides, the RealForensics requires audio for pre-training and introduces multi-task learning during fine-tuning, but our method only requires video modality and doesn't require any other auxiliary tasks during fine-tuning, which is more scalable and flexible.

The results in Tab. 1,2&4 of our manuscript indeed show that our method achieves improvement compared to recent SOTA RealForensics and AltFreezing

on generalization and robustness. Moreover, our method has significantly fewer tuned parameters but a better performance compared to the two competitors as presented in Tab. 3 of the manuscript, further demonstrating its advantages.

Additionally, since the emergence of large multi-modal models, how to integrate the audio appropriately can be a future direction of our work. We don't integrate audio since we aim to explore visual-only video representation, which is unexplored currently and more general than requiring audio. We can integrate audio following similar intuition to learn the audio consistency representation (such as also by masking and shuffling) or learn the cross-modal inconsistency between video and audio, by adding an audio branch to learn the audio representation. We consider this as one of our future directions.

Table 11: Further comparison with RealForensics when using the same auxiliary dataset. Video-level AUC (%) is reported. * means the method is pretrained on VoxCeleb2.

Method	CDF	DFDC	FSh	DFo	*Avg*
RealForensics* [13]	82.9	**78.9**	99.3	98.8	90.0
NACO (ours)	**89.5**	76.7	**99.4**	**99.5**	**91.2**

5.8 Discussions with Other Self-Supervised Detectors

We make further cross-dataset comparisons with more recent self-supervised detectors (LPS [9] and OST [3]) as presented in Tab.12. We find our method can achieve competitive and impressive performance on generalization. Further, LPS, OST, and RealForensics all focus on self-supervised learning with following key differences: LPS learns the spatiotemporal representation by predicting future state, OST generates forgery samples during testing, RealForensics is pretrained to learn audio-visual correspondence and fine-tuned by multi-task learning, while our method is pretrained and fine-tuned on visual-only modality by exploring the natural consistency representation in real face videos with only optimizing a MLP header during fine-tuning.

Table 12: Comparisons on cross-dataset generalization with other self-supervised detectors. Other methods' results are from their original papers.

Method	DFDC	Celeb-DF
LPS	73.2	87.7
OST	83.3	74.8
NACO (ours)	76.7	89.5

Table 13: Confidence interval (±%) for intra-dataset(on FaceForensics++, left), cross-dataset(middle) and cross-manipulation(right) at 0.05 significance level.

	raw	c23	c40	FSh	DFo	DF	FS	F2F	NT
AUC	0.086	0.058	0.132	0.143	0.175	0.083	0.136	0.134	0.162

5.9 Confidence Interval

The experimental results in our manuscript are the mean value of 3 independent tests and we further repeat each test for 7 times, and the confidence intervals (by t-Test [5], df=9) at 0.05 significance level of intra-dataset, cross-dataset and cross-manipulation results are shown in Tab.13, which prove our detection is stable and precise.

5.10 Significance Level

To demonstrate the significance of our method against other competitors, we employ Wilcoxon signed-ranks test [6] to show whether our NACO has a significant performance against other methods when AUC is above 95%. Note that the performance of these methods used for testing is from intra-, cross- and robust cases where our AUC is above 95%. As shown in Tab.14, our method outperforms other baselines on AUC at 0.05 significance level.

Table 14: Summary of the Wilcoxon signed-ranks test for our NACO against other approaches when our AUC is above 95% at 0.05 significance level. The p-values are shown in the brackets.

Xception	CNN-aug	Patch-based	X-Ray	CNN-GRU	LipForensics	FTCN
win[0.009]	**win**[0.001]	**win**[0.001]	**win**[0.001]	**win**[0.004]	**win**[0.043]	tie[0.439]

5.11 Failure Cases

We consider the samples with more complex real-world scenarios and provide several failure cases with their Grad-CAM from DeeperForensics-1.0, as presented in Fig.4. We observe that complicated conditions in real videos (such as illumination) may mislead the detection, and the high-realistic forgery frame with minor movements in fake videos could also increase the detection difficulty. We can intuitively explain the results: the illumination increases the difficulty of natural consistency learning and the fake videos with minor movements contain less consistency clues compared to obvious movements. How to handle complicated conditions in real face videos and minor movements in forgery videos could be our future direction.

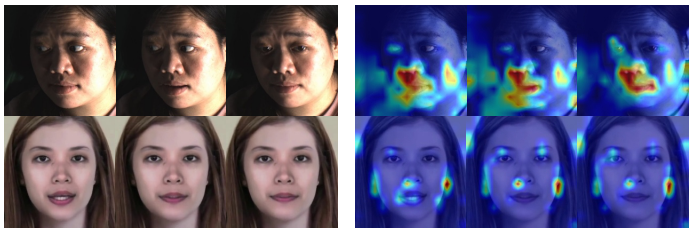

Fig. 4: Failure cases from DeeperForensics-1.0 when fine-tuned on FF++ (c23). The first row is real video but misclassified as fake, while the second row is fake but misclassified as real.

6 Ethical Discussions

The competition between generation and detection is always in progress. New generators could promote the development of detectors then new detectors will force the generators to synthesize more realistic results. Our method could detect the forgery videos that exist spatiotemporal inconsistency, which exists in most current forgery videos. If the forgery videos in the future can entirely eliminate this inconsistency to resist our detector, which is difficult we assume, all detectors based on this may fail. But our method can still provide foundation and insight for the future detectors, especially for how to detect by learning representation on real data to distinguish fake ones.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: CVPR. pp. 4724–4733 (2017)
2. Chai, L., Bau, D., Lim, S.N., Isola, P.: What makes fake images detectable? understanding properties that generalize. In: ECCV. pp. 103–120 (2020)
3. Chen, L., Zhang, Y., Song, Y., Wang, J., Liu, L.: Ost: Improving generalization of deepfake detection via one-shot test-time training. Advances in Neural Information Processing Systems **35**, 24597–24610 (2022)
4. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: Interspeech. pp. 1086–1090 (2018)
5. Dekking, F.M., Kraaikamp, C., Lopuhaä, H.P., Meester, L.E.: A Modern Introduction to Probability and Statistics: Understanding why and how, vol. 488. Springer (2005)
6. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine learning research **7**, 1–30 (2006)
7. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C.: The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854 (2019)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)

9. Ge, S., Lin, F., Li, C., Zhang, D., Wang, W., Zeng, D.: Deepfake video detection via predictive representation learning. ACM TOMM **18**(2s), 115:1–115:21 (2022)
10. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: ICCV. pp. 5842–5850 (2017)
11. Gu, Z., Chen, Y., Yao, T., Ding, S., Li, J., Huang, F., Ma, L.: Spatiotemporal inconsistency learning for deepfake video detection. In: ACM MM. pp. 3473–3481 (2021)
12. Gu, Z., Yao, T., Chen, Y., Ding, S., Ma, L.: Hierarchical contrastive inconsistency learning for deepfake video detection. In: ECCV. pp. 596–613 (2022)
13. Haliassos, A., Mira, R., Petridis, S., Pantic, M.: Leveraging real talking faces via self-supervision for robust forgery detection. In: CVPR. pp. 14930–14942 (2022)
14. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: CVPR. pp. 5039–5049 (2021)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
16. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In: CVPR. pp. 2886–2895 (2020)
17. Lee, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Unsupervised representation learning by sorting sequences. In: ICCV. pp. 667–676 (2017)
18. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Advancing high fidelity identity swapping for forgery detection. In: CVPR. pp. 5074–5083 (2020)
19. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: CVPR. pp. 5001–5010 (2020)
20. Li, M., Li, X., Yu, K., Deng, C., Huang, H., Mao, F., Xue, H., Li, M.: Spatio-temporal catcher: A self-supervised transformer for deepfake video detection. In: ACM MM. pp. 8707–8718 (2023)
21. Li, X., Lang, Y., Chen, Y., Mao, X., He, Y., Wang, S., Xue, H., Lu, Q.: Sharp multiple instance learning for deepfake video detection. In: ACM MM. pp. 1864–1872 (2020)
22. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. In: CVPRW. pp. 46–52 (2019)
23. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: A large-scale challenging dataset for deepfake forensics. In: CVPR. pp. 3204–3213 (2020)
24. Masi, I., Killekar, A., Mascarenhas, R.M., Gurudatt, S.P., AbdAlmageed, W.: Two-branch recurrent network for isolating deepfakes in videos. In: ECCV. pp. 667–684 (2020)
25. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: ECCV. pp. 527–544. Springer (2016)
26. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. In: BTAS. pp. 1–8 (2019)
27. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: ICCV. pp. 1–11 (2019)
28. Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., Natarajan, P.: Recurrent convolutional strategies for face manipulation detection in videos. In: CVPRW. pp. 80–87 (2019)

29. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017)
30. Shiohara, K., Yamasaki, T.: Detecting deepfakes with self-blended images. In: CVPR. pp. 18699–18708 (2022)
31. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. NeurIPS **35**, 10078–10093 (2022)
32. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-generated images are surprisingly easy to spot... for now. In: CVPR. pp. 8695–8704 (2020)
33. Wang, T., Chow, K.P.: Noise based deepfake detection via multi-head relative-interaction. In: AAAI. vol. 37, pp. 14548–14556 (2023)
34. Wang, Z., Bao, J., Zhou, W., Wang, W., Li, H.: Altfreezing for more general video face forgery detection. In: CVPR. pp. 4129–4138 (2023)
35. Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y.: Self-supervised spatiotemporal learning via video clip order prediction. In: CVPR. pp. 10334–10343 (2019)
36. Zhang, D., Li, C., Lin, F., Zeng, D., Ge, S.: Detecting Deepfake Videos with Temporal Dropout 3DCNN. In: IJCAI. pp. 1288–1294 (2021)
37. Zhang, D., Lin, F., Hua, Y., Wang, P., Zeng, D., Ge, S.: Deepfake video detection with spatiotemporal dropout transformer. In: ACM MM. pp. 5833–5841 (2022)
38. Zhao, C., Wang, C., Hu, G., Chen, H., Liu, C., Tang, J.: Istvt: interpretable spatial-temporal video transformer for deepfake detection. TIFS **18**, 1335–1348 (2023)
39. Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., Xia, W.: Learning self-consistency for deepfake detection. In: ICCV. pp. 15003–15013 (2021)
40. Zheng, Y., Bao, J., Chen, D., Zeng, M., Wen, F.: Exploring temporal coherence for more general video face forgery detection. In: ICCV. pp. 15024–15034 (2021)