

Towards a Universal Synthetic Video Detector: From Face or Background Manipulations to Fully AI-Generated Content

Supplementary Material

This supplementary document provides additional details and experimental analyses to complement the main paper. It includes extended discussions, ablations, and robustness studies to further validate the effectiveness of UNITE in detecting spatio-temporal inconsistencies in generated videos.

We begin by presenting Dataset Details (Sec. S-1), offering insights into the datasets used in training and evaluation, highlighting their diversity and complexity. A dedicated theoretical comparison of AD-Loss vs. Contrastive Loss (Sec. S-2) is provided, showcasing the fundamental differences of the AD-Loss implementation as compared to the contrastive loss.

To establish the versatility of UNITE, we change the foundation model backbone in Sec. S-3 and explore its robustness to the choice of synthetic training data (Sec. S-4). We show the t-SNE visualizations to juxtapose the separability of the features learned with and without AD-loss in Sec. S-5.

Additional experiments delve into: UNITE’s performance when trained on DeMamba [5] (Sec. S-6) to perform fair in-domain evaluation against state-of-the-art detectors which were highlighted in cross-domain settings in the main paper; evaluations on the DF40 [32] dataset (Sec. S-7), a recent benchmark for DeepFake detection; results obtained by UNITE in a 4-class fine-grained classification (Sec. S-8); and performance analyses with different compressions on the FF++ [22] dataset (Sec. S-9), shedding light on the resilience of UNITE under varying input conditions.

Hyperparameter and architectural choices are rigorously analyzed through ablations. The sensitivity to AD-Loss hyperparameters (Sec. S-10) investigates the effect of different hyperparameter configurations on model performance. We further analyze the importance of AD-loss feature center’s update in Sec. S-11. In Sec. S-12, we evaluate the impact of padding methods on spatio-temporal consistency, while Sec. S-13 assesses the influence of model design modifications.

Finally, the document concludes with a comprehensive ablation analysis (Sec. S-14), highlighting insights from the above experiments to underscore the critical components that contribute to the exceptional performance of UNITE. This supplementary material aims to provide a deeper understanding of the design and effectiveness of UNITE across diverse settings and challenges.

S-1. Dataset Details

The details of all the datasets used to train/evaluate our UNITE model are as follows:

- **FaceForensics++ (FF++)** [22]: This is the most widely used DeepFake dataset for training detectors, containing fake videos generated by 4 methods: DeepFakes [6], Face2Face [25], FaceSwap [7], and NeuralTextures [26]. FF++ contains 1000 real videos and corresponding four manipulated versions, equaling 4000 fake videos. These videos are provided at three different compression levels to assess robustness under varying quality (experiments with this are in Sec. S-9).
- **CelebDF** [15]: This dataset consists of DeepFake videos featuring celebrity faces. Videos are crafted to minimize noticeable artifacts, improving realism compared to earlier datasets while consisting variations in lighting, background, and facial movements. CelebDF includes 590 real videos and 5639 DeepFake videos.
- **DeeperForensics** [11]: This is one of the largest datasets consisting of 60000 videos constituted by a total of 17.6 million frames. Manipulations are applied to real-world scenarios, addressing challenges like occlusions, lighting changes, and compression.
- **DeepFakeTIMIT** [14]: DeepFakeTIMIT is one of the earliest DeepFake datasets, focusing on low-resource scenarios. The real videos are derived from the VidTIMIT [23] dataset, which contains real videos of 43 individuals speaking scripted sentences. The 640 fake videos are generated by swapping faces using two types of GAN-based approaches.
- **HifiFace** [28]: It is a high-fidelity face-swapping dataset that prioritizes maintaining identity features and perceptual quality. HifiFace consists of 1000 fake videos that maintain sharpness and detail, ideal for high-resolution detection tasks.
- **UADFV** [33]: The University of Albany DeepFake Video Dataset or UADFV is one of the first publicly available (small-scale) datasets for DeepFake detection. It consists of 49 real videos and their corresponding 49 fake videos.
- **AVID** [35]: AVID [35] is a recently proposed video inpainting model. We collected 24 videos from the publicly available supplementary website of the paper¹. The videos do not contain human faces, with the majority featuring no human subjects at all. Instead, the videos focus on some form of background manipulations- from chang-

¹<https://zhang-zx.github.io/AVID/supp/index.html>

Table S-1. Results (accuracy) obtained by UNITE when trained with DINOv2 features of FF++ and GTA-V instead of SigLIP-So400m.

Training Loss	FF++ [22]	CelebDF [15]	AVID [35]	GTA-V [10]	DeMamba [5]
CE Loss only	93.15%	85.40%	33.33%	100.00%	52.46%
CE+AD Loss	98.96% ($\uparrow 5.81\%$)	95.11% ($\uparrow 9.71\%$)	100.00% ($\uparrow 66.67\%$)	100.00% ($\uparrow 0.00\%$)	83.39% ($\uparrow 30.93\%$)

Table S-2. Performance of UNITE when trained with FF++ [22] and DeMamba [5]. For comparison purposes, we have also mentioned the values obtained when UNITE was trained with FF++[22] and GTA-V [10] which were reported in the main paper. The results obtained here show that the model is invariant to the type of synthetic data used during training. As long as there is diversity in the training data, the cross-dataset performances are similar and not data-dependent.

Train		Test						
FF++ & GTA-V	FF++ & DeMamba	Dataset	Accuracy	AUC	Precision@0.5	Recall@0.5	Precs@Rec=0.8	Rec@Precs=0.8
<i>Face Manipulated Data</i>								
✓		FF++	99.96%	99.89%	100.00%	99.84%	100.00%	99.96%
✓		CelebDF	95.11%	94.36%	96.82%	68.75%	96.53%	68.75%
✓		DeeperForensics	99.62%	100.00%	100.00%	99.62%	100.00%	99.63%
✓		DeepFakeTIMIT	91.90%	91.33%	90.45%	88.39%	100.00%	91.95%
✓		HifiFace	75.62%	81.24%	79.55%	71.71%	75.62%	72.47%
✓		UADFV	97.01%	94.95%	96.89%	100.00%	94.12%	100.00%
	✓	FF++	99.92%(-0.04)	99.78%(-0.11)	99.92%(-0.08)	100.00%(+0.16)	99.78%(-0.22)	100.00%(+0.04)
	✓	CelebDF	95.66%(+0.55)	95.66%(+1.30)	95.66%(-1.16)	100.00%(+31.25)	95.66%(-0.87)	100.00%(+31.25)
	✓	DeeperForensics	100.00%(+0.38)	100.00%(+0.00)	100.00%(+0.38)	100.00%(+0.00)	100.00%(+0.00)	100.00%(+0.37)
	✓	DeepFakeTIMIT	92.89%(+0.99)	92.89%(+1.56)	92.89%(+2.44)	100.00%(+11.61)	92.89%(-7.11)	90.56%(-1.39)
	✓	HifiFace	100.00%(+24.38)	100.00%(+18.76)	100.00%(+20.45)	100.00%(+28.29)	100.00%(+24.38)	100.00%(+27.53)
	✓	UADFV	94.32%(-2.69)	89.69%(-5.26)	89.69%(-7.20)	100.00%(+0.00)	88.34%(-5.78)	96.35%(-3.65)
<i>Background Manipulated Data</i>								
✓		AVID	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	✓	AVID	100.00%(+0.00)	100.00%(+0.00)	100.00%(+0.00)	100.00%(+0.00)	100.00%(+0.00)	100.00%(+0.00)
<i>Fully Synthetic Data</i>								
✓		GTA-V	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
✓		DeMamba	87.12%	93.75%	92.76%	89.60%	89.81%	92.12%
	✓	GTA-V	100.00%(+0.00)	100.00%(+0.00)	100.00%(+0.00)	100.00%(+0.00)	100.00%(+0.00)	100.00%(+0.00)
	✓	DeMamba	100.00%(+12.88)	100.00%(+6.25)	100.00%(+7.24)	100.00%(+10.40)	100.00%(+10.19)	100.00%(+7.88)

ing the background scene, to uncropping.

- **GTA-V [10]:** The SAIL-VOS-3D [10] dataset features videos from the GTA-V game (fully synthetic), most of which are human-centric videos. Although the dataset is not AI-generated, the videos from this data is a good representation of fully synthesized content. The training set of this data consists of 161 videos and the validation set (which we use for evaluation/test) consists of 41 videos.
- **DeMamba [5]:** This recently proposed dataset is generated by a number of text-to-video (T2V) and image-to-video (I2V) models, with its training and validation sets consists of a disjoint set of generators. In the training set there are 1.2 million real videos and 1.08 million fake videos. The validation set consists of 10,000 real and 9588 fake videos. In the main paper, we primarily use the validation set for cross-dataset analyses. However in Sec. S-4 and S-6, we utilize a subset of the training data to train our UNITE model for ablation experiments. Due to some publicly available training data files being corrupted in the current version, we could only use the subset of data that was intact and accessible for training.

S-2. AD-Loss vs. Contrastive Loss

At first glance, the Attention-Diversity (AD) loss (Sec. 3.4 of the main paper) and the contrastive loss [13] may appear similar. However, there are key differences between the two, despite both being fundamentally inspired by Fisher Discriminant Analysis [9] for deep networks.

- Contrastive loss is generally used to maximize the similarity between related pairs (positives) and minimize it between unrelated pairs (negatives) in a learned embedding space, to ensure that similar data points are closer and dissimilar ones are further apart. AD-loss, on the other hand, focuses on encouraging spatial diversity of attention across multiple heads within a transformer network. Instead of comparing sample pairs, it ensures that different attention heads look at different regions of a video frame, promoting a more holistic and diverse coverage of features in a single example.
- In contrastive loss, sample pairs or triplets, consisting of an anchor (random sample) a positive (sample of the same class as the anchor) and a negative (sample of a different class from the anchor) are compared, and their distances in the latent space are optimized through either supervised labels or unsupervised similarity constraints. AD-loss operates *within a single instance* and works by diversify-

Table S-3. **SOTA Comparison on DeMamba synthetic data:** Comparison of the results obtained by UNITE and state-of-the-art methods when trained on the DeMamba [5] *train* set and tested on the DeMamba *val* set. We also mention the results obtained when UNITE was trained on FF++ [22] and GTA-V [10], which was reported in the main paper. **Bold** shows the current best results and the previous best and second-best results are highlighted in **red** and **blue** respectively. The performance gain is highlighted in **green**. (*P*: Precision@0.5 and *R*: Recall@0.5)

Method	Metrics	Sora [3]	Morph Studio [1]	Runway ML (Gen2) [21]	HotShot [18]	Lavie [29]	Show-1 [34]	Moon Valley [17]	Crafter [4]	Model Scope [27]	Wild Scrape [5]	Avg
TALL [31]	P	71.15%	96.89%	98.51%	79.38%	84.59%	79.38%	98.79%	99.02%	92.70%	76.47%	87.91%
	R	91.07%	98.28%	97.83%	83.00%	76.57%	79.57%	99.52%	98.93%	94.14%	66.31%	88.52%
F3Net [20]	P	68.27%	99.89%	99.67%	89.35%	57.00%	36.57%	99.52%	99.71%	93.80%	88.41%	88.73%
	R	83.93%	99.71%	98.62%	77.57%	85.24%	63.17%	99.58%	99.89%	89.43%	76.78%	81.88%
NPR [24]	P	91.07%	99.57%	99.49%	24.29%	89.64%	57.71%	97.12%	99.86%	94.29%	87.80%	82.45%
	R	91.07%	99.57%	99.49%	24.29%	89.64%	57.71%	97.12%	99.86%	94.29%	87.80%	84.08%
STIL [8]	P	57.21%	99.08%	99.32%	86.19%	82.24%	70.43%	99.25%	98.96%	97.18%	81.32%	87.12%
	R	67.86%	96.00%	98.41%	96.14%	77.14%	80.43%	97.44%	96.93%	96.29%	68.36%	82.22%
MINTIME-CLIP-B [5]	P	83.21%	99.99%	99.67%	50.84%	99.20%	99.27%	99.76%	99.99%	91.83%	91.77%	91.55%
	R	89.29%	100.00%	98.99%	26.43%	96.79%	98.14%	99.84%	100.00%	84.29%	82.38%	87.62%
FTCN-CLIP-B [5]	P	91.79%	99.99%	99.79%	45.94%	99.76%	97.80%	99.99%	99.99%	94.69%	92.32%	92.21%
	R	87.50%	100.00%	98.91%	17.71%	97.71%	91.86%	100.00%	100.00%	85.29%	82.83%	86.18%
CLIP-B-PT [5]	P	67.80%	43.56%	70.88%	29.97%	52.97%	35.36%	55.52%	66.03%	44.23%	42.99%	44.83%
	R	85.71%	82.43%	90.36%	71.00%	79.29%	75.43%	89.62%	86.29%	82.14%	75.16%	81.74%
DeMamba-CLIP-PT [5]	P	25.87%	95.14%	96.23%	73.43%	83.31%	75.49%	90.17%	95.06%	95.06%	69.95%	79.97%
	R	58.93%	96.43%	93.12%	68.00%	69.36%	69.00%	89.14%	91.86%	96.14%	56.59%	78.86%
XCLIP-B-PT [5]	P	16.39%	72.16%	87.77%	39.86%	65.57%	54.26%	75.23%	84.80%	61.60%	55.28%	61.29%
	R	81.34%	82.15%	83.35%	80.98%	81.82%	81.55%	82.14%	82.98%	81.93%	81.10%	81.93%
DeMamba-XCLIP-PT [5]	P	18.26%	93.50%	94.72%	69.94%	78.08%	71.50%	83.95%	92.23%	93.54%	68.10%	76.38%
	R	66.07%	95.86%	94.64%	77.86%	75.36%	80.29%	90.89%	92.50%	96.00%	66.41%	83.59%
XCLIP-B-FT [5]	P	64.42%	99.73%	96.78%	70.98%	90.35%	77.28%	97.34%	99.84%	82.01%	88.97%	86.77%
	R	82.14%	99.57%	93.62%	61.29%	79.36%	69.71%	97.92%	99.79%	77.14%	83.59%	84.41%
UNITE (Trained on FF+GTA-V)	P	88.57%	100.00%	100.00%	90.16%	89.91%	98.34%	99.52%	100.00%	98.96%	92.56%	92.76%(+0.55)
	R	92.11%	100.00%	94.62%	96.93%	98.12%	99.86%	98.69%	100.00%	96.29%	89.89%	89.60%(+1.08)
UNITE (Trained on DeMamba)	P	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%(+7.79)	100.00%(+11.48)
	R	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%(+11.48)

ing the spatial focus of attention heads. By minimizing overlap among different attention heads’ focus, it ensures better feature extraction across both the foreground and background, addressing challenges posed by partial and fully synthetic manipulations.

- Contrastive loss is typically used in tasks like metric learning [30], clustering [16], and self-supervised pre-training [12], aiming to create distinct clusters of data representations. AD-loss is tailored specifically for Deep-Fake detection, enhancing the transformer model’s ability to capture subtle, diverse cues across frames—be it for faces, backgrounds, or synthetic content—thereby improving detection accuracy without limiting focus to just a few regions.

S-3. Ablation on Foundation Model

We conducted an ablation study by replacing the SigLIP foundation model with DINOv2-ViT-L/14 [19] in our UNITE framework. This substitution was primarily aimed at demonstrating that the choice of foundation model does not significantly impact the overall performance improvements observed in our framework. Since the foundation model is used to extract image-level features, we hypothesized that its specific architecture would not be a critical factor in our system’s performance gains.

The results from this study, shown in Table S-1, support our hypothesis, showing that the substantial improvements in accuracy across various datasets are due to our novel video-transformer architecture combined with the AD-loss,

rather than the specific choice of foundation model. This suggests that our framework’s performance enhancements are robust and can be achieved with different foundation models, as long as they provide suitable image-level features. Therefore, the key innovation and contribution of our work lie in the design of the video-transformer and the use of AD-loss, which can be effectively combined with a variety of foundation models to achieve superior performance.

S-4. Robustness to the Choice of Synthetic Training Data

To demonstrate that the performance of UNITE is not specifically reliant on the GTA-V [10] dataset, we evaluate its capability to generalize when trained with diverse synthetic datasets. Specifically, we trained UNITE on the FF++ [22] and DeMamba [5] *train* splits and evaluated its performance across all datasets.

The results, presented in Table S-2, show that UNITE achieves comparable performance across datasets, with a notable improvement on the DeMamba synthetic dataset [5] since this evaluation becomes in-domain in nature and GTA-V [10] now becomes out-of-domain evaluation. On the other datasets (excluding HifiFace, where the results increased by a significant margin), the performance remains consistent between the model trained using GTA-V [10] data and the one trained using DeMamba [5], illustrating that UNITE is robust to the choice of synthetic data used during training and can effectively utilize any synthetic dataset to enhance training diversity.

S-5. t-SNE Visualizations

To further understand the impact of the AD-loss on the feature space of our UNITE framework, we conducted a t-SNE analysis (Fig. S-1) on the features extracted when trained on FF++ and GTA-V datasets. This analysis was performed with and without the AD-loss to visually assess the separability between real (red) and fake (blue) classes.

The results from the t-SNE plots reveal a significant improvement in class separability when the AD-loss is incorporated into the training process. The features learned with AD-loss exhibit a clearer distinction between real and fake samples, indicating that the AD-loss helps to create a more discriminative feature space. This enhanced separability is particularly notable in cross-dataset settings, where the model is trained on one dataset but evaluated on another. The improved class separation in these scenarios suggests that the AD-loss not only enhances the model’s performance on the training dataset but also improves its generalizability across different datasets.

This observation supports our claim that the AD-loss plays a crucial role in enhancing the robustness and effectiveness of our UNITE framework by promoting a more structured and discriminative feature representation. The visual evidence from the t-SNE analysis complements the quantitative performance improvements observed in our experiments, providing a deeper insight into how the AD-loss contributes to the overall success of our approach.

S-6. UNITE trained on DeMamba

We trained UNITE on the DeMamba [5] *train* set to ensure a fair comparison with state-of-the-art (SOTA) detectors (refer to Table 3 in the main paper), which were also trained on the same dataset. In the main paper, UNITE was instead trained on FF++ [22] and GTA-V [10], focusing on cross-dataset analysis. Despite this cross-domain setting, UNITE outperformed SOTA detectors in terms of average performance. To further demonstrate the versatility and effectiveness of the UNITE model, we additionally conducted this in-domain evaluation as an ablation experiment.

The results of this comparison are presented in Table S-3. Notably, UNITE achieved perfect performance (100%) on the DeMamba *val* split, outperforming the previous best method by $\sim 8\%$ in precision and $\sim 11\%$ in recall, demonstrating its ability to effectively capture the spatio-temporal inconsistencies inherent in videos generated by T2V and I2V models.

S-7. Experiments on DF40 Dataset

DF40 [32] is a recently proposed diverse DeepFake dataset which was generated with 40 DeepFake generation techniques including face-swapping and face-synthesis. The currently available version of the dataset, however, has a

Table S-4. Results obtained by our UNITE model (trained on FF++ [22] and GTA-V [10] on the DF40 [32] dataset.

Metric	Value
Accuracy	99.97%
AUC	99.82%
Precision@0.5	99.85%
Recall@0.5	99.97%
Precs@Rec=0.8	100.00%
Rec@Precs=0.8	99.97%

major problem- although it is a video DeepFake dataset, several of the generation techniques have single unrelated images in the dataset. So for our evaluation, we have removed those sets, and only used the video data available, making the “usable” set as 23 DeepFake generators, out of the 40 used in the DF40 [32]. The results obtained by UNITE on this usable set is reported in Table S-4, which indicates that UNITE (which was trained with FF++ [22] and GTA-V [10]) achieved a near-perfect performance even in this cross-dataset setting.

S-8. Finer-grained Classification

To further evaluate the robustness and fine-grained detection capabilities of our UNITE framework, we conducted additional experiments using a four-class classification setup. The classes included real, face-swap, face-reenactment, and fully synthetic videos. We maintained the same training protocol, using FF++ and GTA-V datasets for training.

The results of these fine-grained experiments are presented in Table S-5. Our UNITE framework demonstrates robust performance across different datasets and classes. On the FF++ dataset (in-domain evaluation), it achieves high accuracy for all classes, resulting in an average accuracy of 97.15%. Similarly, on the DF40 dataset (cross-dataset evaluation), the model performs well with an average accuracy of 86.97%.

Overall, the fine-grained experiments demonstrate the ability of our UNITE framework to distinguish between different types of manipulated and real videos effectively. The results underscore the robustness of our approach in handling various manipulation techniques and datasets, reinforcing its potential for real-world applications in video authenticity detection.

S-9. Experiments with Different Compression Levels

The FF++ [22] dataset provides three levels of compression—“raw”, “c23”, and “c40”—that reflect varying levels of video quality, with “raw” being uncompressed (high quality or HQ), c23 being a medium-compression level

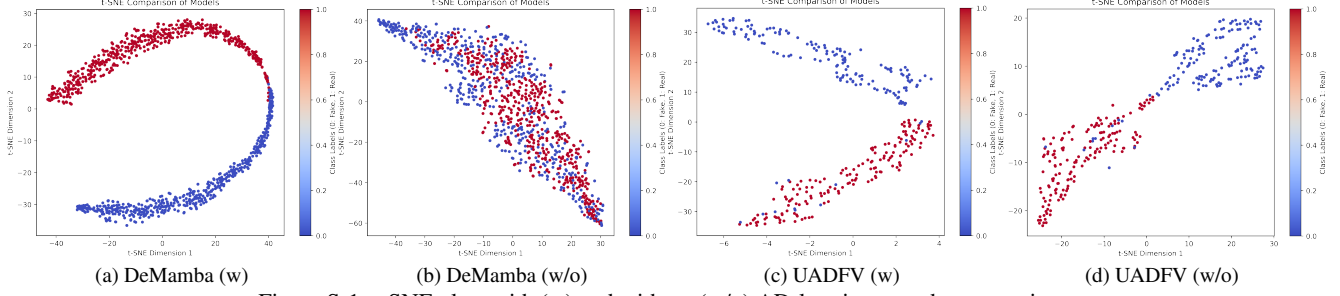


Figure S-1. t-SNE plots with (w) and without (w/o) AD-loss in cross-dataset settings.

Table S-5. **4-class Fine-Grained Results:** We divide the existing DeepFake datasets into four categories- (1) face-swap, (2) face-reenactment (3) fully synthetic and (4) real, to perform a 4-class fine-grained classification using UNITE.

Dataset	Face-swap	Face-Reenact	Fully Synthetic	Real	Average
FF++ [22]	98.21%	92.75%	-	97.64%	97.15%
DF40 [32]	85.79%	87.94%	-	88.65%	86.97%
GTA-V [10]	-	-	100.00%	-	100.00%
DeMamba [5]	-	-	59.39%	63.12%	61.35%

(medium quality or MQ) and c40 representing high compression (low quality or LQ). In our experiments, we trained UNITE on the c23 split to strike a balance between video quality and real-world applicability. To assess the robustness of the trained model across different compression levels, we conducted an ablation study by evaluating its performance on the HQ and LQ splits, the results of which are shown in Table S-6. This experiment is crucial as it demonstrates the model’s ability to generalize across varying video qualities, which is representative of diverse real-world scenarios for detection of in-the-wild DeepFakes. The results of this evaluation reveal the resilience of UNITE to compression artifacts, ensuring its applicability to practical deployments where video quality can vary significantly.

S-10. Sensitivity to AD-loss hyperparameters

To analyze the sensitivity of UNITE to the choice of the δ_{within} and $\delta_{between}$ hyperparameters (refer to Sec. 3.4 of the main paper), we conducted an ablation study varying their values, the results of which are shown in Fig. S-2. Our results reveal that the performance of UNITE remains relatively consistent across a wide range of δ_{within} and $\delta_{between}$ values. Notably, in the case of δ_{within} ablation in Fig. S-2(a), this robustness is most pronounced when the signs of the first and second components are opposite, suggesting that the loss function benefits when the initial feature centers are away from each other. This behavior underscores the stability of UNITE’s optimization process under varied hyperparameter settings, reducing the need for extensive tuning, even under cross-dataset settings.

S-11. Ablation on Feature-Center Update in AD-Loss

To investigate the importance of updating feature centers in the AD-loss during training, we conducted an ablation study where we disabled the feature centers’ update. This experiment was performed on the FF++ and GTA-V datasets, which were used for training our UNITE framework. The results are presented in Table S-7.

The findings from this study clearly indicate that updating the feature centers is a crucial component of the AD-loss. Without this update, the model’s performance significantly deteriorates, highlighting the importance of this mechanism in refining the feature space. The dynamic update of feature centers plays a vital role in allowing the model to better distinguish between real and fake videos. This process helps to adapt the model to the evolving characteristics of the data during training, leading to improved generalization and robustness.

Overall, this ablation study highlights the importance of the feature centers’ update in the AD-loss, reinforcing the effectiveness of our proposed method in enhancing the discriminative power of the feature representations.

S-12. Ablation on Padding Choice

As mentioned in Sec. 3.1 of the main paper, in our experiments, UNITE employs padding using the last frame of each video segment, a strategy that ensures the temporal consistency of the input sequence. To evaluate the impact of this padding choice, we conducted an ablation study comparing two padding strategies: (1) padding with zeros, and (2) padding with the last frame of the video segment.

Table S-6. **Ablation on different compression factors:** Results obtained by the UNITE model (trained on “c23” compression) on different compression factors of the FF++ [22] dataset. The difference in performances across the compression factors are minimal indicating robustness of the UNITE model.

Compression	Accuracy	AUC	Precision@0.5	Recall@0.5	Precs@Rec=0.8	Rec@Precs=0.8
raw	99.38%	99.57%	99.60%	99.64%	99.60%	99.64%
c23	99.96%	99.89%	100.00%	99.84%	100.00%	99.96%
c40	95.69%	94.57%	95.60%	96.33%	93.29%	94.57%

Table S-7. **Ablation on Feature Centers Update:** We compare the results (accuracy) obtained by UNITE with and without the feature centers’ update step (Eq. 3 of the main paper) in AD-loss.

Feature centers	FF++ [22]	CelebDF [15]	AVID [35]	GTA-V [10]	DeMamba [5]
Without update	95.43%	80.07%	54.17%	100.00%	57.83%
With update	99.96%	95.11%	100.00%	100.00%	87.12%
	(↑4.53%)	(↑15.04%)	(↑45.83%)	(↑0.00%)	(↑29.29%)

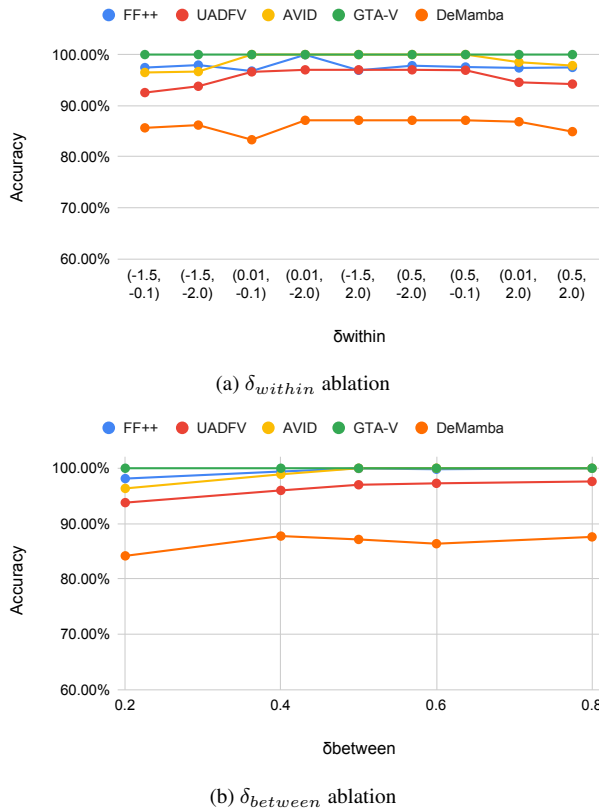


Figure S-2. **Ablation on AD-Loss Hyperparameters:** Performance comparison of UNITE across varying values of the (a) δ_{within} and (b) $\delta_{between}$ hyperparameters. The results indicate that the model’s learning is relatively robust to changes in the hyperparameters. Specifically in (a) the results are consistent when the signs of the first and second parameters of δ_{within} are opposite.

We performed this ablation in Table S-8, where UNITE was

trained *only* on FF++ [22] and in Table S-9, where UNITE was trained on both FF++ [22] and GTA-V [10].

The results revealed a significant performance gain when padding with the last frame compared to zero padding. This improvement is attributed to the preservation of semantic and temporal context provided by the last frame, which better aligns with the continuity of video data, as opposed to the abrupt discontinuity introduced by zero padding.

Moreover, this performance gain is even more pronounced when UNITE was trained on the combined datasets of FF++ [22] and GTA-V [10] in Table S-9. The diversity and domain disparity between these datasets amplify the importance of maintaining contextual integrity in input sequences. Padding with the last frame mitigates potential domain shifts introduced by abrupt padding artifacts, thereby enabling UNITE to generalize more effectively across heterogeneous datasets.

These findings underscore the critical role of padding strategies in video-based learning tasks, particularly when dealing with diverse dataset training scenarios with cross-dataset evaluations.

S-13. Ablation on Architecture Choice

To validate our choice of a transformer-based architecture, we compared UNITE with a simpler model using the same SigLIP-So400m [2] features (on FF++[22] and GTA-V [10] training set) but replaced the transformer with average pooling and classification layers using CE Loss (denoted as *AvgPool* model), the results of which are shown in Fig. S-3.

While this *AvgPool* baseline could perform basic feature aggregation, it lacked the ability to capture complex spatiotemporal dependencies. In contrast, UNITE’s multi-head self-attention mechanism effectively identified nuanced inconsistencies across frames, proving essential for robust synthetic video detection. This highlights the neces-

Table S-8. **Padding Ablation:** Results obtained by the UNITE model when trained on FF++ [22] *only*, using zero-padding vs. padding with the last available frame. **Green** shows the performance improvement.

Padding		Test						
Zeros	Last Frame	Dataset	Accuracy	AUC	Precision@0.5	Recall@0.5	Precs@Rec=0.8	Rec@Precs=0.8
<i>Face Manipulated Data</i>								
✓		FF++	99.55%	99.68%	99.70%	99.75%	99.70%	99.75%
✓		CelebDF	68.20%	94.44%	96.72%	65.08%	71.67%	65.08%
✓		DeeperForensics	83.49%	100.00%	100.00%	83.49%	100.00%	83.49%
✓		DeepfakeTIMIT	83.27%	80.84%	89.39%	80.90%	84.32%	82.90%
✓		HifiFace	64.63%	59.61%	65.98%	64.63%	54.54%	61.58%
✓		UADFV	93.48%	91.85%	93.71%	93.12%	93.71%	93.12%
	✓	FF++	99.53% (-0.02)	99.77% (+0.09)	99.94% (+0.24)	99.49% (-0.26)	99.94% (+0.24)	99.94% (+0.19)
	✓	CelebDF	72.61% (+4.41)	94.05% (-0.39)	96.45% (-0.27)	61.22% (-3.86)	80.45% (+8.78)	61.22% (-3.86)
	✓	DeeperForensics	91.35% (+7.86)	100.00% (+0.00)	100.00% (+0.00)	91.35% (+7.86)	100.00% (+0.00)	91.35% (+7.86)
	✓	DeepfakeTIMIT	86.90% (+3.63)	86.46% (+5.62)	83.61% (-5.78)	83.97% (+3.07)	88.90% (+4.58)	81.33% (-1.57)
	✓	HifiFace	63.63% (-1.00)	62.47% (+2.86)	67.12% (+1.14)	63.63% (-1.00)	59.30% (+4.76)	63.63% (+2.05)
	✓	UADFV	94.12% (+0.64)	94.38% (+2.53)	95.68% (+1.97)	97.11% (+3.99)	93.79% (+0.08)	94.38% (+1.26)
<i>Background Manipulated Data</i>								
✓		AVID	37.50%	33.33%	33.33%	37.50%	0.00%	33.33%
	✓	AVID	41.67% (+4.17)	33.33% (+0.00)	33.33% (+0.00)	41.67% (+4.17)	41.67% (+41.67)	33.33% (+0.00)
<i>Fully Synthetic Data</i>								
✓		GTA-V	55.23%	55.23%	60.19%	55.23%	59.47%	56.23%
✓		DeMamba	60.56%	54.54%	60.10%	29.15%	57.69%	50.12%
	✓	GTA-V	60.16% (+4.93)	61.52% (+6.29)	60.16% (-0.03)	58.73% (+3.50)	63.29% (+3.82)	58.73% (+2.50)
	✓	DeMamba	61.47% (+0.91)	57.38% (+2.84)	67.73% (+7.63)	33.01% (+3.86)	62.15% (+4.46)	54.16% (+4.04)

Table S-9. **Padding Ablation:** Results obtained by the UNITE model when trained on FF++ [22] and GTA-V [10], using zero-padding vs. padding with the last available frame. **Green** shows the performance improvement.

Padding		Test						
Zeros	Last Frame	Dataset	Accuracy	AUC	Precision@0.5	Recall@0.5	Precs@Rec=0.8	Rec@Precs=0.8
<i>Face Manipulated Data</i>								
✓		FF++	99.52%	99.66%	99.69%	99.73%	99.69%	99.73%
✓		CelebDF	92.18%	94.17%	96.16%	60.13%	89.46%	60.13%
✓		DeeperForensics	96.45%	100.00%	100.00%	96.45%	93.47%	96.45%
✓		DeepfakeTIMIT	83.78%	85.63%	85.34%	51.01%	98.72%	88.52%
✓		HifiFace	70.66%	80.24%	75.41%	54.66%	69.48%	70.66%
✓		UADFV	94.41%	92.83%	94.38%	94.38%	94.38%	94.38%
	✓	FF++	99.96% (+0.44)	99.89% (+0.23)	100.00% (+0.31)	99.84% (+0.11)	100.00% (+0.31)	99.96% (+0.23)
	✓	CelebDF	95.11% (+2.93)	94.36% (+0.19)	96.82% (+0.66)	68.75% (+8.62)	96.53% (+7.07)	68.75% (+8.62)
	✓	DeeperForensics	99.62% (+3.17)	100.00% (+0.00)	100.00% (+0.00)	99.62% (+3.17)	100.00% (+6.53)	99.63% (+3.18)
	✓	DeepfakeTIMIT	91.90% (+8.12)	91.33% (+5.70)	90.45% (+5.11)	88.39% (+37.38)	100.00% (+1.28)	91.95% (+3.43)
	✓	HifiFace	75.62% (+4.96)	81.24% (+1.00)	79.55% (+4.14)	71.71% (+17.05)	75.62% (+6.14)	72.47% (+1.81)
	✓	UADFV	97.01% (+2.60)	94.95% (+2.12)	96.89% (+2.51)	100.00% (+5.62)	94.12% (-0.26)	100.00% (+5.62)
<i>Background Manipulated Data</i>								
✓		AVID	83.33%	100.00%	100.00%	83.33%	100.00%	83.33%
	✓	AVID	100.00% (+16.67)	100.00% (+0.00)	100.00% (+0.00)	100.00% (+16.67)	100.00% (+0.00)	100.00% (+16.67)
<i>Fully Synthetic Data</i>								
✓		GTA-V	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
✓		DeMamba	84.68%	89.61%	84.09%	87.92%	84.63%	90.89%
	✓	GTA-V	100.00% (+0.00)	100.00% (+0.00)	100.00% (+0.00)	100.00% (+0.00)	100.00% (+0.00)	100.00% (+0.00)
	✓	DeMamba	87.12% (+2.44)	93.75% (+4.14)	92.76% (+8.67)	89.60% (+1.68)	89.81% (+5.18)	92.12% (+1.23)

sity of the transformer approach for comprehensive detection performance.

S-14. Comprehensive Ablation Analysis

To evaluate the impact of various design choices on the performance of UNITE, we conduct an ablation study summarized in Table S-10. The baseline model uses average pooling over SigLIP-So400m features [2], which achieves reasonable performance but struggles to generalize, partic-

ularly on datasets with greater spatio-temporal complexity, such as AVID [35] and DeMamba [5].

The first modification replaces the simplistic *AvgPool* architecture with a transformer. While this improves performance across most datasets, owing to the transformer’s capability to capture temporal inconsistencies, the gains are limited for datasets with high diversity, such as AVID [35] and GTA-V [10] with nominally better than random performance on DeMamba [5], suggesting that architectural

Table S-10. **Evolution of detection performance under different ablation settings:** This table highlights the impact of various modifications to the UNITE training pipeline on the detection performance (accuracy) across multiple datasets. Starting from a base model using a simple average pooling model on SigLIP-So400m features (Sec. S-13), we show the effect of changing the architecture to a transformer, incorporating synthetic data into the training process, and adding the proposed AD-Loss. These settings progressively enhance performance, with the addition of AD-Loss achieving near-perfect or significantly improved results across all datasets.

Ablation Settings	FF++	CelebDF	UADFV	AVID	GTA-V	DeMamba
<i>AvgPool on SigLIP Features</i>	81.35%	69.11%	66.67%	0.00%	0.00%	45.24%
<i>Change Architecture to Transformer</i>	96.81%($\uparrow 15.46\%$)	70.03%($\uparrow 0.92\%$)	91.04%($\uparrow 24.37\%$)	0.00%($\uparrow 0.00\%$)	17.46%($\uparrow 17.46\%$)	55.48%($\uparrow 10.24\%$)
<i>Add Synthetic Data to Training</i>	98.69%($\uparrow 1.88\%$)	69.43%($\downarrow 0.60\%$)	94.03%($\uparrow 2.99\%$)	33.33%($\uparrow 33.33\%$)	100.00%($\uparrow 82.54\%$)	60.98%($\uparrow 5.50\%$)
<i>Add AD-Loss</i>	99.96%($\uparrow 1.27\%$)	95.11%($\uparrow 25.68\%$)	97.01%($\uparrow 2.98\%$)	100.00%($\uparrow 66.67\%$)	100.00%($\uparrow 0.00\%$)	87.12%($\uparrow 26.14\%$)

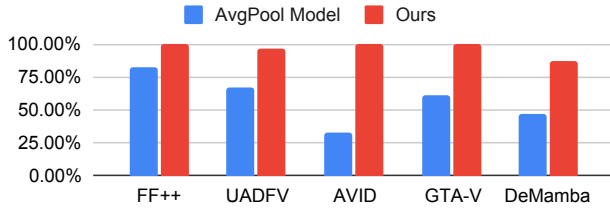


Figure S-3. **Architecture Evaluation:** Comparison of UNITE with a simplistic model with average pooling and classifier layers that works on the same SigLIP-So400m [2] features, to justify our choice of a complex transformer architecture.

changes alone are insufficient for robust generalization to all forms of DeepFakes.

Next, we introduce synthetic data (specifically GTA-V [10] into the training process, significantly enhancing performance, particularly for datasets like AVID [35] and GTA-V [10]. The addition of synthetic data improves the model’s ability to handle diverse scenarios, as it increases the training data diversity and exposes the model to a broader range of spatio-temporal inconsistencies.

Finally, adding the proposed Attention-Diversity (AD) loss further boosts the performance across all datasets. Notably, the model achieves near-perfect performance on most datasets, with significant improvements observed for challenging datasets such as DeMamba [5] and CelebDF [15]. This demonstrates the effectiveness of AD-Loss in leveraging the model’s attention mechanisms to better capture fine-grained inconsistencies.

Overall, the results highlight that the combined use of synthetic data and AD-Loss is critical for achieving state-of-the-art performance, demonstrating the robustness and adaptability of UNITE across diverse datasets.

References

- [1] Morph studio. <https://www.morphstudio.com/>, 2024. 3
- [2] Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting vit in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 7, 8
- [3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>, 3, 2024. 3
- [4] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 3
- [5] Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024. 1, 2, 3, 4, 5, 6, 7, 8
- [6] Deepfakes GitHub. <https://github.com/deepfakes/faceswap>, . 1
- [7] FaceSwap GitHub. <https://github.com/MarekKowalski/FaceSwap>, . 1
- [8] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3473–3481, 2021. 3
- [9] Harald Hanselmann, Shen Yan, and Hermann Ney. Deep fisher faces. In *BMVC*, 2017. 2
- [10] Yuan-Ting Hu, Jiahong Wang, Raymond A Yeh, and Alexander G Schwing. Sail-vos 3d: A synthetic dataset and baselines for object detection and 3d mesh reconstruction from video data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1418–1428, 2021. 2, 3, 4, 5, 6, 7, 8
- [11] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deepeforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2889–2898, 2020. 1

- [12] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *Advances in neural information processing systems*, 33: 16199–16210, 2020. 3
- [13] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 2
- [14] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 1
- [15] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020. 1, 2, 6, 8
- [16] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8547–8555, 2021. 3
- [17] moonvalley.ai. moonvalley.ai. <https://moonvalley.ai/>, 2022. 3
- [18] John Mullan, Duncan Crawbuck, and Aakash Sastry. Hotshot-XL, 2023. 3
- [19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3
- [20] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020. 3
- [21] Runway Research. Text driven video generation. <https://research.runwayml.com/gen2>, 2023. 3
- [22] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 1, 2, 3, 4, 5, 6, 7
- [23] Conrad Sanderson. The vidtimit database. 2002. 1
- [24] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 3
- [25] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 1
- [26] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 1
- [27] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscape text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 3
- [28] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hiface: 3d shape and semantic prior guided high fidelity face swapping. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021. 1
- [29] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yanan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 3
- [30] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009. 3
- [31] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668, 2023. 3
- [32] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Li Yuan, Chengjie Wang, Shouhong Ding, et al. Df40: Toward next-generation deepfake detection. *NeurIPS 2024*, 2024. 1, 4, 5
- [33] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019. 1
- [34] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pages 1–15, 2024. 3
- [35] Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yanan Zhao, Peter Vajda, Dimitris Metaxas, and Licheng Yu. Avid: Any-length video inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7162–7172, 2024. 1, 2, 6, 7, 8