Detecting Deepfakes in Video Conferencing: Domain-Specific Challenges and Model Adaptations

Brendan P. Murphy

bigsur@stanford.edu

Introduction

This research addresses the emerging threat of real-time deepfake attacks in video conferencing platforms. Using a modified AltFreezing architecture, I developed a system that processes 32-frame video clips to detect facial manipulation artifacts specific to video conferencing scenarios. The model achieves 85% AUC on Zoom video detection, demonstrating effective domain adaptation for video conferencing environments.

Datasets

FF++ DLC

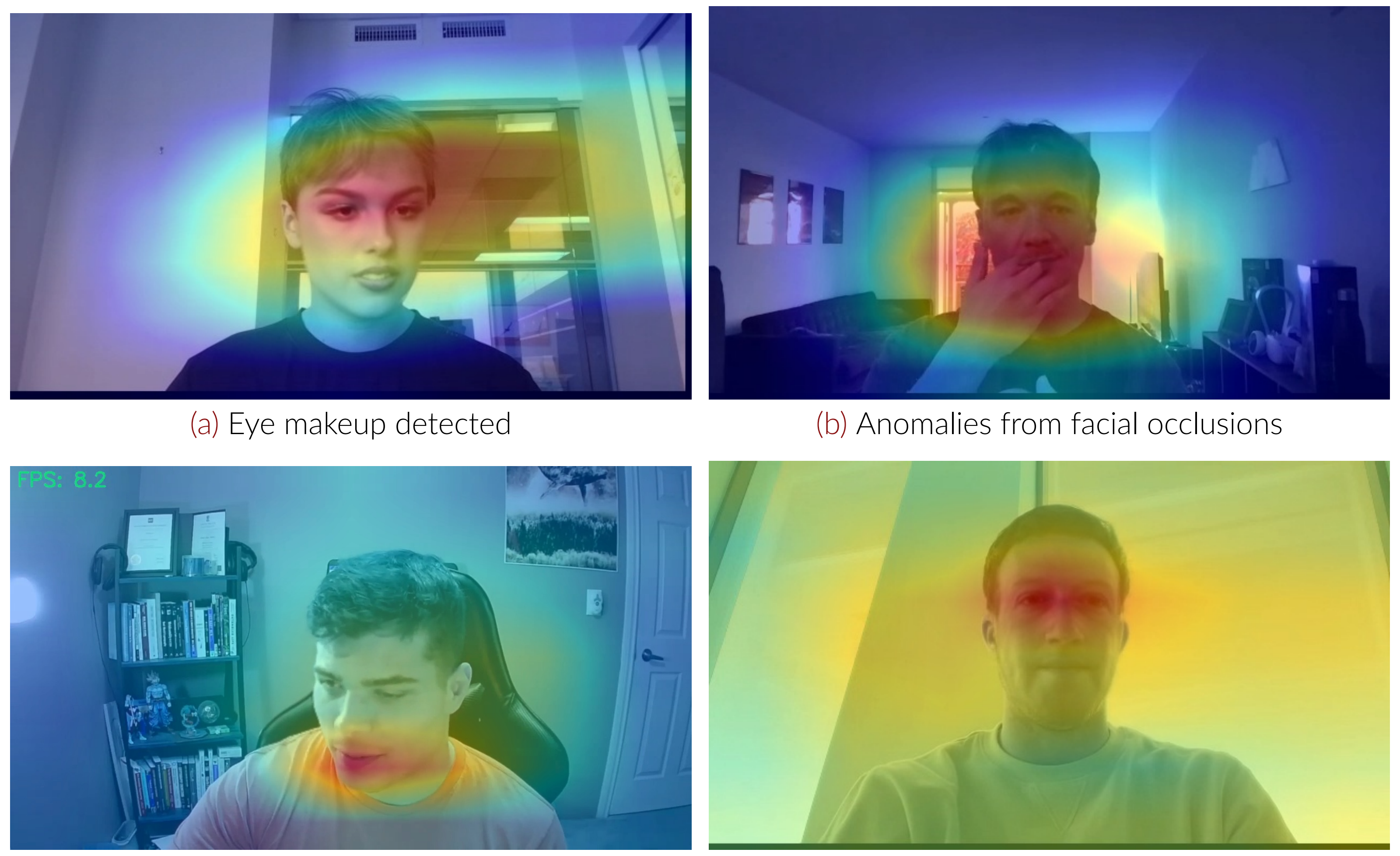
- Constructed by modifying the FaceForensics++ dataset using DLC face-swapping
- Contains 120 videos (100 training, 20 validation) split evenly between real and fake
- Segmented into 4,500 clips of 32 frames at 224x224 resolution
- Preserves consistent lighting and framing while introducing identity manipulation artifacts

Zoom DLC

- Curated specifically for video conferencing scenarios
- Contains 114 videos (48 training, 66 evaluation) with real and fake samples
- Segmented into 4,300 clips incorporating compression, lighting, and resolution variations
- Manipulated using DLC to simulate real-world impersonation attempts

Celeb-DF

- Used as external benchmark for out-of-distribution evaluation
- Original dataset contains 590 real and 5,639 synthetic videos, used 100 videos (50 real, 50 fake)
- Maintains high-quality deepfake content without video conferencing artifacts



(a) Eye makeup detected

(b) Anomalies from facial occlusions

(c) Mouth artifacts detected

(d) Artifacts near eyes detected

Figure 1. Examples of artifact detection on Zoom videos. Artifacts include eye makeup effects, hand occlusions, and synthetic patterns around mouth and eyes. These visualizations show the model's sensitivity to subtle manipulations.

Methods

Feature Extraction

- **Spatial Features**: Per-frame analysis captures facial inconsistencies and compression artifacts across individual frames
- **Temporal Features**: Cross-frame analysis identifies motion discontinuities and lighting variations between sequential frames
- **3D ResNet-50**: Backbone network processes both spatial and temporal dimensions simultaneously for unified feature representation

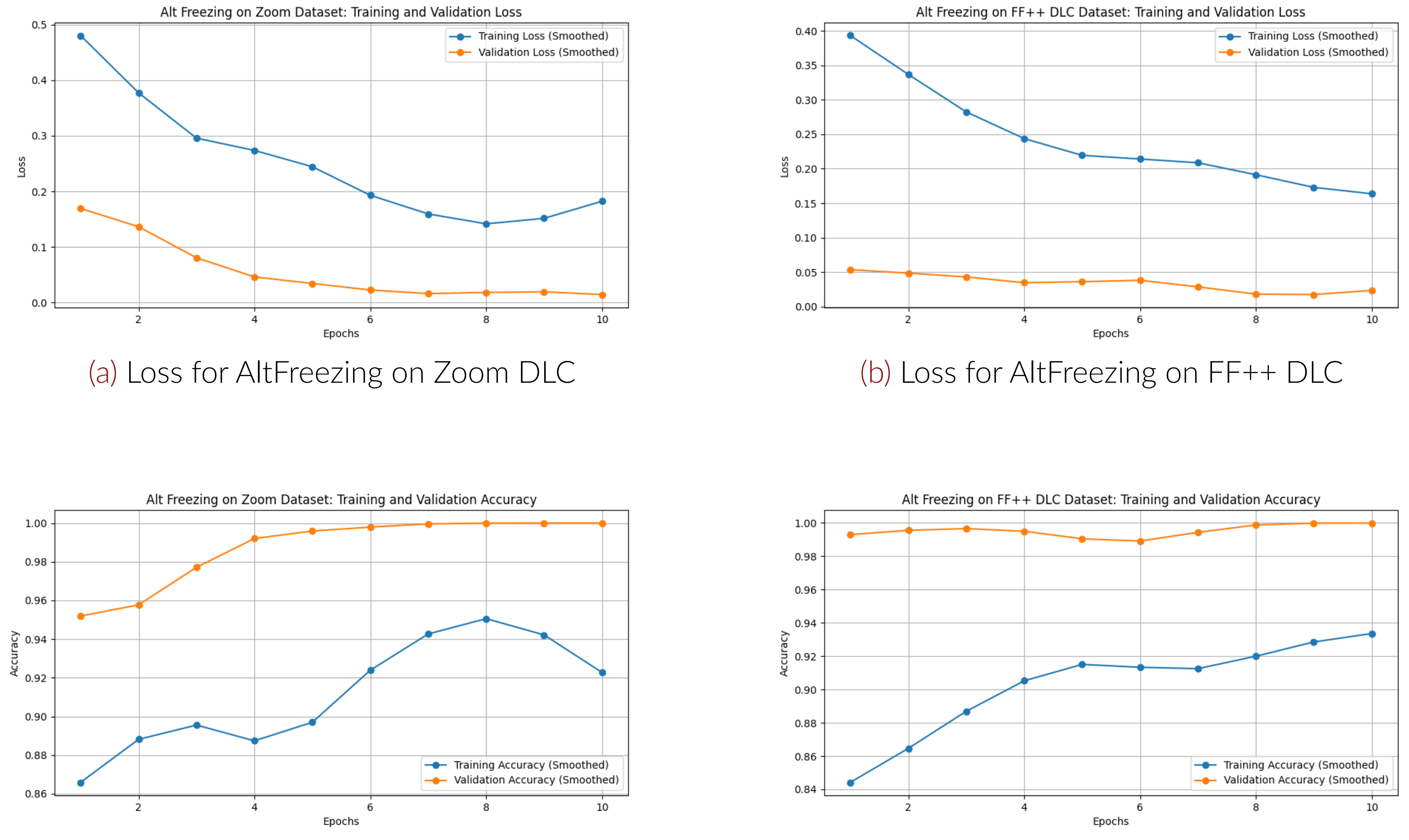
AltFreezing Model Architecture

- **Backbone Network**: Pre-trained 3D ResNet-50 $\phi(X)$ extracts spatio-temporal features
- **Classification Head**: Module $h(\hat{u})$ processes features for final prediction
- **Output Layer**: Sigmoid activation σ produces binary classification: $f_\theta(X) = \sigma(h(\phi(X)))$

Fine-tuning Strategies

- **Gradual Unfreezing**: Incrementally unfreezes backbone layers to balance utilizing pre-trained features while adapting to domain-specific characteristics
- **Alternate Freezing**: Switches between updating spatial and temporal parameters to enforce learning both feature types
- **Implementation**: Uses Adam optimizer with learning rate 5e-6 and weight decay 0.005

Training Analysis



(a) Loss for AltFreezing on Zoom DLC

(b) Loss for AltFreezing on FF++ DLC

(c) Accuracy for AltFreezing on Zoom DLC

(d) Accuracy for AltFreezing on FF++ DLC

Figure 2. The validation curves highlight key differences between datasets. FF++ DLC shows high initial accuracy and minimal improvement suggesting pretraining on similar data limited further learning, while Zoom data shows noticeable improvement, demonstrating effective domain specific fine-tuning. These results underscore the importance of dataset diversity in training. *Note that training accuracies appear higher than final evaluation metrics due to the use of sliding window predictions on 32-frame clips during training versus video-level mean predictions during real-world evaluation.

Evaluation

- **Visual Analysis**: Grad-CAM heatmaps provide qualitative evaluation of model attention and shows regions identified as manipulated, validating detection of specific artifacts.
- **AUC (Area Under Curve)**: Primary metric measuring detection performance across different classification thresholds, providing a threshold-independent assessment of model's ability.
- **AUPRC (Area Under Precision-Recall Curve)**: Evaluates model performance on imbalanced datasets by showing the trade-off between precision and recall, particularly important for real-world detection scenarios.
- **F1 Score and Accuracy**: Secondary metrics that measure classification performance and precision-recall balance, though less emphasized for cross-dataset comparisons due to their threshold dependency (evaluated at 0.1). In deployment, thresholds would be tuned based on specific application requirements.

Results

Model/Method	Train Data	Eval Data	AUC	AUPRC	F1	Accuracy
logistic regression	Zoom	Zoom	0.04	-	0.18	0.10
base model	N/A	Zoom	0.81	0.80	0.67	0.68
base model	N/A	Celeb DF	0.93	0.92	0.72	0.62
alternate freezing	FF++(dlc)	Zoom	0.83	0.79	0.70	0.62
alternate freezing	Zoom	Zoom	**0.85**	0.76	0.70	0.62
gradual unfreezing	FF++(dlc)	Zoom	0.83	0.79	**0.76**	**0.72**
gradual unfreezing	Zoom	Zoom	0.82	**0.81**	0.65	0.68
alternate freezing	Zoom	Celeb DF	0.52	0.48	0.62	0.50

Table 1. Performance comparison of models across training and evaluation configurations on Zoom and Celeb-DF datasets. Metrics include AUC, AUPRC, F1 Score, and Accuracy, with bold values indicating the best performance of the optimized models, highlighting trade-offs between domain specific optimization and generalization.

Discussion

The experiments revealed a clear trade off between domain specific performance and generalization. While alternate freezing achieved the highest AUC on Zoom data, it showed limited generalization to out-of-distribution samples. Gradual unfreezing demonstrated better generalization but lower domain-specific performance. Real-world challenges included monitor glare false positives and difficulty with poorly lit subjects, portrait mode videos with borders, and subjects far away from the camera. Future development will focus on significantly expanding the Zoom dataset with more diverse scenarios.

References

[1] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection, 2023.