# Supplementary for Detecting Localized Deepfake Manipulations Using Action Unit-Guided Video Representations

Tharun Anand Siva Sankar Pravin Nair
Indian Institute of Technology Madras
ed20b068@smail.iitm.ac.in, ch20b103@smail.iitm.ac.in, pravinnair@ee.iitm.ac.in

In this supplementary document, we provide additional details to demonstrate the effectiveness of the proposed method. We start by providing detailed training procedure for both the pretext tasks and the deepfake detection framework. Further, we evaluate each pretext task individually through reconstruction performance and visual comparisons. We also include additional metrics, such as Average Recall (AR) and mean F1 score (mF1), along with the primary metrics of AUC and Average Precision (AP) shown in the manuscript. This offers a more detailed comparison of our method against existing state-of-the-art detection methods across different deepfake generation methods.

In addition, we elaborate on the construction of the latest deepfake videos with local manipulations, including descriptions of the methods and parameters used to generate fake videos with localized subtle edits. As a part of the supplementary material, we include samples of real and fake videos and a table that includes the probability comparison between our method and existing deepfake detection techniques.

## 1. Comprehensive Training Details

The two pretext tasks are trained using the CelebV-HQ dataset [24], which contains approximately $35,000$ real facial videos. The first pretext model for the reconstruction of masked frames minimizes a variant of $\ell_1$ reconstruction loss (Huber loss) between ground truth frames and frames reconstructed from masked inputs, following a VideoMAE-like approach [20]. The second pretext model for Facial Action Unit (AU) detections is trained using Huber loss between predicted AU maps and ground truth maps, to predict 16 AUs for each frame. To generate the ground-truth attention map for every action unit, we define landmarks corresponding to the different AUs following the conventional approach in [9, 16]. Elliptical regions are fitted to these landmarks as initial AU regions, which are then smoothed using a Gaussian filter of radius 3. This process yields 16 distinct AU maps, each corresponding to a specific localized action, for a single frame.

| | DVA | STIT | DFE | TF | FZ | T2L | V2P |
|---|---|---|---|---|---|---|---|
| MAE: Random Masking | 2.71e-8 | 2.6e-8 | 2.64e-8 | 2.48e-8 | 2.35e-8 | 2.40e-8 | 2.38e-8 |
| MAE: AU Detection | 1.23e-8 | 1.17e-8 | 1.19e-8 | 1.05e-8 | 9.8e-9 | 1.02e-8 | 1.01e-8 |

Table 1. **Reconstruction Error for Pretext Tasks:** The pretext models for random masking and AU detection are evaluated to demonstrate their standalone effectiveness. MAE between ground-truth and reconstructions is tabulated across diverse datasets (target data is normalized between 0 and 1). The negligible MAE values, ranging from $10^{-9}$ to $10^{-8}$, highlight the effectiveness of the learned representations in both the pretext tasks.

We trained both the pretext models using the Adam optimizer with a batch size of 8 for 600 epochs. Gradient accumulation was applied every 20 steps. We used the pretrained checkpoints from VideoMAE [20] to initialize our weights for both the pretext tasks.

During fine-tuning for deepfake detection, the fused encoder shown in Fig. 4 in the manuscript, is trained with a classifier on the FF++ dataset [15], consisting of 700 real and $3,600$ fake videos generated via four manipulation methods [3, 8, 18, 19]. Focal Loss [14] is used to address class imbalance. For finetuning, we used a batch size of 8 for 100 epochs. A learning rate of $1e$-5 with an exponential decay of $1e$-3 is used for both the pretext tasks and the finetuning stage.

## 2. Evaluation on Pretext tasks

We evaluate the performance of pretext models independently to demonstrate the effectiveness of the representations learned by the respective encoders, VFE (Video Frame Encoder) and AUE (Action Unit Encoder). For the first self-supervised task - reconstruction of face-centered frames from masked input frames - we compute the Mean Absolute Error (MAE) between the output reconstructed frames and the ground truth. MAE is first computed across all 16 input frames for each video to obtain a video-level MAE. This score is then averaged over all videos across diverse methods, and presented in the first row of Table 1.

For the second self-supervised pretext task - reconstruction of AU maps for each video - we compute the MAE

| Detection Methods | DVA | | | | STIT | | | | DFE | | | | Tokenflow | | | | VideoP2P | | | | TextLive | | | | Fatezero | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | AP | AR | mf | AUC | AP | AR | mf | AUC | AP | AR | mf | AUC | AP | AR | mf | AUC | AP | AR | mf | AUC | AP | AR | mf | AUC | AP | AR | mf |
| FTCN | 27.1 | 30.1 | 25.4 | 27.2 | 34.8 | 37.0 | 32.1 | 34.5 | 33.5 | 35.8 | 30.9 | 33.3 | 29.5 | 32.1 | 27.0 | 29.5 | 31.0 | 33.7 | 29.2 | 31.5 | 30.2 | 32.9 | 28.6 | 30.8 | 28.7 | 31.3 | 26.9 | 29.1 |
| RealForensics | 37.5 | 40.3 | 36.2 | 38.2 | 46.9 | 49.8 | 44.5 | 47.1 | 45.6 | 48.6 | 43.2 | 45.8 | 41.2 | 44.2 | 39.0 | 41.6 | 42.8 | 45.8 | 40.4 | 43.0 | 42.0 | 45.1 | 39.8 | 42.3 | 40.5 | 43.4 | 38.5 | 40.9 |
| Lip Forensics | 33.8 | 37.1 | 32.4 | 34.7 | 42.0 | 45.8 | 40.3 | 43.0 | 40.9 | 44.6 | 39.0 | 41.7 | 36.5 | 40.3 | 34.7 | 37.4 | 38.0 | 41.9 | 36.0 | 38.8 | 37.3 | 41.2 | 35.4 | 38.2 | 35.8 | 39.5 | 33.9 | 36.6 |
| EfficientNet+ViT | 36.2 | 38.4 | 34.9 | 36.6 | 44.7 | 47.1 | 42.8 | 44.9 | 43.5 | 45.9 | 41.3 | 43.5 | 39.0 | 41.5 | 37.2 | 39.3 | 40.8 | 43.2 | 38.6 | 40.8 | 40.1 | 42.5 | 38.0 | 40.2 | 38.6 | 40.8 | 36.5 | 38.6 |
| Face X-Ray | 33.4 | 36.1 | 31.5 | 33.8 | 41.1 | 43.6 | 38.8 | 41.2 | 40.3 | 42.9 | 38.2 | 40.5 | 36.0 | 39.1 | 34.2 | 36.5 | 37.5 | 40.7 | 35.4 | 37.8 | 36.8 | 40.0 | 34.8 | 37.3 | 35.2 | 38.2 | 33.5 | 35.8 |
| LAA Net | 61.5 | 58.0 | 55.2 | 56.6 | 72.5 | 69.3 | 66.4 | 67.8 | 71.2 | 68.1 | 64.8 | 66.3 | 66.4 | 62.9 | 60.3 | 61.6 | 68.0 | 64.5 | 62.0 | 63.3 | 67.1 | 63.7 | 60.9 | 62.2 | 65.7 | 61.8 | 59.3 | 60.6 |
| SBI | 65.2 | 62.8 | 59.4 | 61.0 | 75.5 | 73.2 | 70.1 | 71.6 | 73.3 | 71.5 | 68.5 | 69.9 | 69.0 | 66.4 | 63.5 | 64.9 | 70.8 | 68.1 | 65.2 | 66.6 | 70.2 | 67.5 | 64.7 | 66.1 | 68.6 | 65.9 | 63.1 | 64.5 |
| **Ours** | **87.2** | **85.8** | **82.5** | **84.1** | **92.5** | **90.7** | **88.1** | **89.4** | **93.1** | **91.6** | **89.5** | **90.5** | **91.7** | **89.4** | **87.0** | **88.2** | **90.3** | **90.2** | **86.9** | **88.0** | **89.1** | **87.9** | **85.5** | **86.6** | **88.5** | **86.0** | **84.2** | **85.1** |

Table 2. **Cross-Dataset Quantitative Comparison:** AUC, AP, AR, and mF1 scores evaluated across the latest deepfake generation methods. The results highlight the superior detection performance of our method, significantly surpassing existing state-of-the-art approaches in identifying fine-grained localized edits.

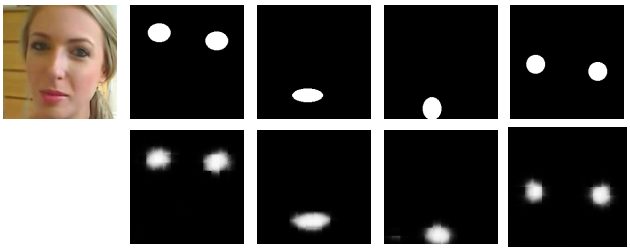


Figure 1. **AU Detection Maps Comparison:** Comparison of four AU maps for a sample test image, with ground-truth maps (top row) and reconstructed maps (bottom row). The accurate reconstruction across action units highlights the effectiveness of the pretext task in preserving spatio-temporal localization.

between the 16 reconstructed AU maps and the corresponding ground truth maps for every frame. For diverse methods, this per-frame MAE is initially averaged across all 16 frames for each video. These video-level MAE values are then averaged across the all the videos corresponding to a particular deepfake generation method to obtain the final reconstruction error, as reported in the second row of Table 1. The low MAE values for both the pretext tasks demonstrates effectiveness of their respective learned representations. In Fig. 1, a qualitative comparison is shown, where, for a single frame, we display selected ground-truth AU maps alongside their reconstructed counterparts as output by the model. These visualizations highlight the model's capability in accurately capturing fine-grained facial details.

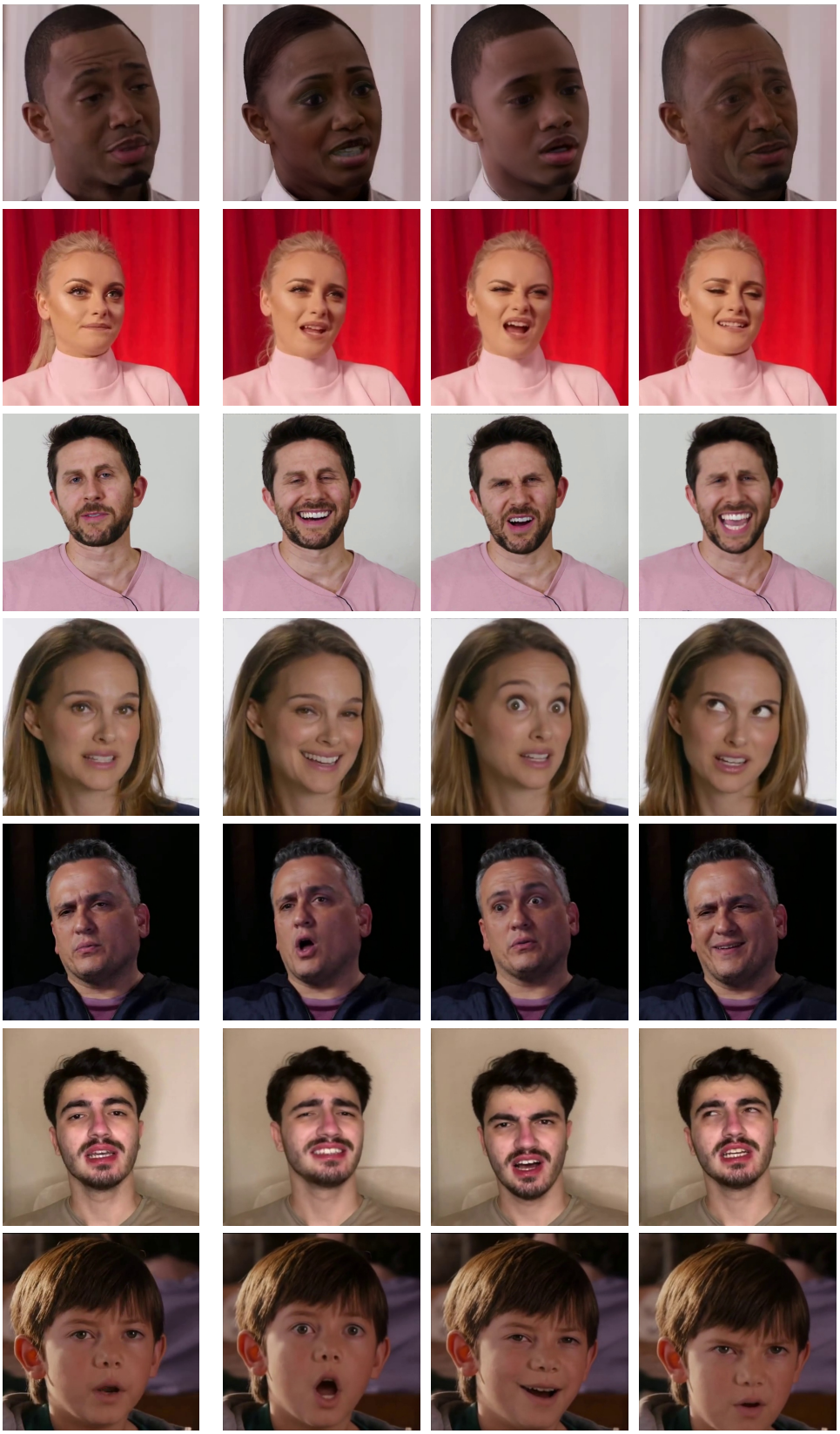
## 3. Latest Locally edited Deepfake Videos

We leveraged seven state-of-the-art methods to test proposed deepfake detection method : Diffusion Video Autoencoders (DVA) [7], Stitch It In Time (STIT) [21], Disentangled Face Editing (DFE) [23], Tokenflow [5], VideoP2P [10], FateZero [13], Text2Live [1]. For all the methods, we utilized their official source code and generated 50 videos each. These methods enabled localized edits targeting eyes, mouth, expressions, age, and gender transformations. For DVA, we used 1000 sampling steps, a learning rate of 0.002 (for finetuning), and an editing scale of 0.5. For StyleGAN2 [6] based editing methods STIT and DFE, we followed the common pipeline for editing, which involves video inversion to latent space, finetuning the generator for a specific video, and editing the latent vector. For STIT, we used 50 steps for finetuning the generator, along with an editing range of $+6$ to $-6$. Similarly, for DFE, we used 50 steps for finetuning and edtiting range between $-10$ to $+10$. TokenFlow, Video-P2P, and FateZero utilize pre-trained diffusion models during inference, standardized with 50 DDIM inversion steps and a classifier-free guidance scale of 7.5 for text fidelity. Video-P2P further employs a cross-attention replacement ratio of 0.4 to enhance temporal consistency. Text2LIVE, in contrast is trained for each video using a video-specific generator for 1,000 optimization steps.

## 4. Additional Experimental results

In this section, we evaluate our method's generalization capability in a cross-dataset setting. As shown in Table 3, our method consistently achieves high performance on standard datasets, exceeding 90% AUC and matching the performance of recent deepfake detection models, LAANet [12], SBI [17], AltFreezing [22] and CADMM [4], as shown in Table 3. In the case of latest locally manipulated video, existing SOTA methods experience a significant drop in performance. The current SOTA methods exhibit AUCs as low as 30-75%, as shown in Table 2, whereas our method demonstrates robust generalization, achieving an AUC as high as 93%. A similar trend is noticeable in the case of all other metrics as well. Notably, our approach exhibits a superior average recall, across all the compared videos, indicating high accuracy in detecting fake videos (considered as positives), with significantly fewer false negatives and a manageable number of false positives, ensuring efficient and reliable detection even for localized manipulations by recent deepfake methods.

To visually illustrate our model's superior performance, we display frames of videos with various localized edits in Fig. 2, along with probability scores for detection. All real videos utilized in this experiment are from the publicly available dataset [11]. Our method consistently achieves confidence scores exceeding 90% in detecting localized edits within fake videos, as compared to the existing state-of-



***Gender, Young, Old***

| Method | Fake Prob. |
|---|---|
| AltFreeze | 20.33 |
| CADMM | 36.0 |
| SBI | 45.80 |
| LAANet | 50.13 |
| Ours | **91.25** |

***Sad, Anger, Disgust***

| Method | Fake Prob. |
|---|---|
| AltFreeze | 34.47 |
| CADMM | 40.30 |
| SBI | 68.10 |
| LAANet | 59.17 |
| Ours | **90.36** |

***Smile, Anger, Disgust***

| Method | Fake Prob. |
|---|---|
| AltFreeze | 31.27 |
| CADMM | 35.0 |
| SBI | 73.31 |
| LAANet | 63.69 |
| Ours | **93.98** |

***Smile, EyeRaise, EyeGaze***

| Method | Fake Prob. |
|---|---|
| AltFreeze | 32.13 |
| CADMM | 36.0 |
| SBI | 51.19 |
| LAANet | 53.74 |
| Ours | **87.89** |

***Shock, EyeRaise, Smile***

| Method | Fake Prob. |
|---|---|
| AltFreeze | 33.0 |
| CADMM | 33.69 |
| SBI | 66.25 |
| LAANet | 59.07 |
| Ours | **89.13** |

***Disgust, Anger, EyeGaze***

| Method | Fake Prob. |
|---|---|
| AltFreeze | 31.03 |
| CADMM | 37.33 |
| SBI | 74.0 |
| LAANet | 53.5 |
| Ours | **93.62** |

***Shock, Smile, Anger***

| Method | Fake Prob. |
|---|---|
| AltFreeze | 35.59 |
| CADMM | 37.48 |
| SBI | 70.35 |
| LAANet | 54.6 |
| Ours | **93.15** |

**Original** **Fake Manipulations** ***Edit types*** **& Probability score**

Figure 2. **Visual Detection Comparison for Locally Manipulated Videos:** A real video (left) undergoes three types of localized manipulations, generating fake videos that are visually indistinguishable from the original. The reported confidence scores, averaged across the three manipulations, highlight our method's superior ability to detect subtle edits compared to the best methods.

| Method | CDF2 | | | | DFD | | | | DFW | | | | DFDC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | AP | AR | mf | AUC | AP | AR | mf | AUC | AP | AR | mf | AUC | AP | AR | mf |
| LAA Net [12] | **95.4** | **97.64** | 87.71 | **92.41** | **99.5** | **99.8** | 95.47 | **97.59** | 87.6 | 85.08 | 69.66 | 78.56 | 86.94 | **97.7** | 73.37 | 83.81 |
| SBI [17] | 93.18 | 85.16 | 82.68 | 83.90 | 97.56 | 92.79 | 89.49 | 91.11 | 84.83 | **88.37** | 81.64 | 84.60 | 86.16 | 93.24 | 71.58 | 80.99 |
| AltFreezing [22] | 89.5 | 88.46 | 85.50 | 86.24 | 98.50 | 97.86 | 97.0 | 97.41 | 72.6 | 70.86 | 68.5 | 69.66 | **94.0** | 92.57 | **91.1** | **91.80** |
| CADMM [4] | 93.0 | 91.12 | 77.00 | 83.46 | 99.03 | 99.59 | 82.17 | 90.04 | 75.0 | 72.80 | 71.26 | 72.14 | 88.3 | 86.7 | 85.62 | 86.1 |
| EfficientNet+ViT [2] | 79.0 | 75.61 | 74.5 | 75.05 | 87.0 | 88.09 | 85.8 | 86.93 | 72.0 | 68.74 | 67.0 | 67.85 | 91.0 | 85.12 | 83.7 | 84.39 |
| Ours | 93.84 | 95.27 | **92.66** | 92.17 | 97.15 | 95.28 | **98.6** | 97.23 | **91.0** | 88.25 | **88.63** | **87.40** | 93.0 | 91.93 | 90.38 | 91.26 |

Table 3. **Cross-Dataset Quantitative Comparison:** Evaluation of AUC, AP, AR, and mF1 scores across standard deepfake datasets, focused on face swapping and reenactment. The proposed method is competitive with existing SOTA approaches across all metrics. Notably, our method achieves superior AR values, indicating high sensitivity in detecting fake videos (positives).

**Fake Score:** 56.6 **Fake Score:** 60.2 **Fake Score:** 67.3

Figure 3. **Limitations:** Detection of fake videos (bottom row) generated from real videos (top row) with localized edits. A noticeable drop in confidence scores (20-35%) is observed in case of occlusions or side-facing poses, since the proposed representations do not capture action unit dynamics effectively.

the-art detection methods. This observation holds consistently across a diverse range of localized edits, including expressions such as smiles, shock, disgust, sadness, anger, and modifications like eyebrow raises, eye gaze adjustments, and gender or age transformations. The supplementary material also includes the real and fake videos corresponding to the frames depicted in Fig.5 in the manuscript and Fig. 2.

Next, in Fig. 3, we present examples where our method exhibits a noticeable drop in confidence scores for detecting fake videos generated through localized edits applied to three real videos in [11]. Most of these cases occur when the subject is facing sideways or when occlusions hinder the learned representations to accurately capture facial dynamics through action units. The corresponding videos are included in the supplementary material.

## References


[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. *European Conference on Computer Vision*, abs/2204.02491, 2022. 2

[2] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. *Proc. International Conference on Image Analysis and Processing*, pages 219–229, 2022. 4

[3] Deepfakes Community. Deepfakes github repository. *GitHub Repository*, 2024. Accessed: 2024-11-09. 1

[4] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3994–4004, 2023. 2, 4

[5] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *International Conference on Learning Representations*, abs/2307.10373, 2023. 2

[6] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 2

[7] Gyeongman Kim, Hajin Shim, Hyunsu Kim, Yunjey Choi, Junho Kim, and Eunho Yang. Diffusion video autoencoders: Toward temporally consistent face video editing via disentangled video encoding. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6091–6100, 2023. 2

[8] Marek Kowalski. Faceswap. *GitHub Repository*, 2024. Accessed: 2024-11-09. 1

[9] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. EAC-Net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2583–2596, 2018. 1

[10] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8599–8608, 2023. 2

[11] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020. 2, 4

[12] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. LAA-Net: Localized artifact attention network for quality-agnostic and generalizable deepfake

detection. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17395–17405, 2024. 2, 4

[13] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15886–15896, 2023. 2

[14] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2980–2988, 2017. 1

[15] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *Proc. IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019. 1

[16] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. *Proceedings. European Conference on Computer Vision*, pages 705–720, 2018. 1

[17] Kaede Shiohara, Xingchao Yang, and Takafumi Taketomi. BlendFace: Re-designing identity encoders for face-swapping. *Proc. IEEE/CVF International Conference on Computer Vision*, pages 7634–7644, 2023. 2, 4

[18] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time face capture and reenactment of rgb videos. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 1

[19] 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

[20] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Proc. Advances in Neural Information Processing Systems*, 35:10078–10093, 2022. 1

[21] Rotem Tzaban, Ron Mokady, Rinon Gal, Amit Bermano, and Daniel Cohen-Or. Stitch It In Time: Gan-based facial editing of real videos. *Proc. SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2

[22] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. AltFreezing for more general video face forgery detection. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4129–4138, 2023. 2, 4

[23] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. *Proc. IEEE/CVF International Conference on Computer Vision*, pages 13789–13798, 2021. 2

[24] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. *Proc. European Conference on Computer Vision*, pages 650–667, 2022. 1