The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2/W9-2025
ISPRS Intl. Workshop "Photogrammetric and computer vision techniques for environmental and infraStructure monitoring, Biometrics and Biomedicine"
PSBB25 , 9–11 June 2025, Moscow, Russia

VCF: A Real-World Video Conference Deepfake Benchmark for Face-Swap Detection and Robustness Evaluation

Maksim Krasilnikov[1,2], Mikhail Nikitin[2], Anton Konushin[3,1]

[1] Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, 119991 Moscow, Russia - (anton.konushin, maksim.krasilnikov)@graphics.cs.msu.ru
[2] Video Analysis Technologies LLC, 119634 Moscow, Russia
[3] AIRI, Moscow, Russia – konushin@airi.net

Keywords: Dataset, Benchmark, Deepfake detection, Face-swapping detection

Abstract

The rapid advancement of deepfake generation techniques poses significant security and privacy risks, particularly in video conferencing scenarios where variable resolutions, compression artifacts, and environmental factors complicate detection. Existing benchmarks often fail to address these context-specific challenges, limiting their applicability to real-world communication platforms. To bridge this gap, we introduce VCF (**V**ideo **C**onference Deep**F**akes) dataset, the first, to the best of our knowledge, specialized benchmark designed for evaluating deepfake detection in video conferencing contexts. VCF leverages the VCD dataset as target videos and the LaPa dataset as a set of source faces, algorithmically ranking sources by similarity in gender, ethnicity, age, and facial hair to select optimal matches for enhanced deepfake visual plausibility. The dataset incorporates multi-resolution videos, H.264 compression artifacts from different compression rates, and diverse backgrounds to simulate conditions specific to video conferences. Comprehensive evaluations of 14 detection methods reveal significant performance degradation under video quality variations. Our results emphasize the critical need for robust detection frameworks resilient to resolution shifts, compression artifacts, and diverse generation pipelines. VCF provides a standardized, scenario-specific benchmark to drive advancements in securing digital communication platforms against evolving deepfake threats.

1. Introduction

Deepfake technology has rapidly evolved in both sophistication and accessibility, creating potential security and privacy risks that threaten digital integrity. In particular, the rise of video conferencing platforms for professional, educational, and personal interactions has increased interest in methods for detecting manipulated videos under realistic and diverse conditions. However, existing benchmarks for deepfake detection largely focus on general video manipulation scenarios and frequently overlook the unique challenges arising from video conferencing setups. For instance, these setups include variable resolutions based on network conditions, diverse compression artifacts, lighting conditions, and potential virtual background modifications. Such factors necessitate the development of a specialized framework to comprehensively evaluate deepfake detection methods in a video conferencing context.

To address this need, we present **VCF** (**V**ideo **C**onference Deep**F**akes), the first, to the best of our knowledge, specialized benchmark designed to evaluate techniques for detecting deep forgeries in video conferencing context. VCF leverages two primary data sources: the VCD dataset, which contains authentic video conferencing recordings in high-quality 1080p, and the LaPa dataset, used for inserting new faces through face swapping. We selectively match target and source faces based on criteria such as gender, ethnicity, facial hair, and age — thereby minimizing the required transformations and improving the visual believability of the generated deepfakes. Furthermore, VCF deliberately includes variations in input video resolutions and compression levels to more accurately reflect the breadth of possible real-world conditions. This holistic approach to dataset construction ensures that researchers and practitioners can evaluate how detection algorithms perform when encountering typical video conferencing artifacts and alterations.

Initial tests with modern deepfake detection algorithms reveal that state-of-the-art methods still face notable challenges on the VCF benchmark. These challenges emphasize the dataset's capacity to expose performance gaps that might remain hidden in more generic deepfake detection datasets. In particular, we observe degradation in detection accuracy across different resolutions and compression levels, emphasizing the need for robust solutions against adversarial manipulations performed on low- and mid-quality video streams. Our findings further reinforce the importance of specialized benchmarks in driving algorithmic improvements designed for realistic user-centric scenarios.

In summary, our main contributions are:

1. The introduction of VCF, a specialized deepfake benchmarking dataset made for video conferencing, incorporating realistic scene setups, a broad range of resolutions, and compression artifacts. The dataset is publicly available [1].
2. A detailed methodology for generating high-quality deepfake videos, including careful face matching strategies (based on gender, ethnicity, facial hair, and age) to ensure realistic face swaps, as well as multiple publicly available face-swapping methods to showcase variability in generated deepfakes.
3. Extensive evaluation of existing deepfake detection algorithms, highlighting their strengths and limitations when dealing with the unique challenges of video conferencing environments.

[1] https://github.com/mirmashel/vcf_dataset

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-169-2025 | © Author(s) 2025. CC BY 4.0 License.

2. Related Work

Recent achievements in the development of neural network and generative technologies have spurred a diverse range of studies in deepfake generation and deepfake detection.

Many works and open source methods for creating deepfakes began to appear. Face2Face (Thies et al., 2016) pioneered real-time facial reenactment by mapping expressions from a source to a target face with high fidelity. FaceShifter (Li et al., 2020a) improved identity preservation using a two-phase architecture, thus facilitating more realistic face-swapping results. SimSwap (Chen et al., 2020a) introduced an Injection Module which transfers the identity information of the source face into the target face at feature level, enhancing transfer of facial features. DiffFace (Kim et al., 2022) integrated diffusion-based generative models to capture fine-grained details in synthetic faces. HyperDreamBooth (Ruiz et al., 2024) focused on text-driven face generation via hypernetwork-based fine-tuning, allowing precise control over facial stylization. Deep-Live-Cam (Estanislao, 2024) an open-source repository addressed real-time manipulation in video conferencing, optimizing latency and visual consistency.

Due to the development of deepfake generation methods, there was a request for detection methods, which began to develop in an adversarial manner. A variety of detection methods have been proposed to counter evolving face manipulation techniques. Early approaches, such as Meso4 and MesoInception4 (Afchar et al., 2018), focus on mesoscopic properties of images to detect subtle artifacts in manipulated media. Capsule networks (Nguyen et al., 2019) utilize dynamic routing between capsules to model hierarchical spatial relationships. F3Net (Qian et al., 2020) employs frequency-aware decomposition and local frequency statistics to deeply mine the forgery patterns, while FFD (Dang et al., 2020) using the attention mechanism to improve the classification and localization of altered areas. SRM (Luo et al., 2021), originally designed for steganalysis, has been repurposed to detect noise patterns indicative of deepfakes. The SPSL framework (Liu et al., 2021) combines spatial and phase spectrum features via shallow learning to suppress high-level features and focus on the local region. Recce (Cao et al., 2022) utilizes reconstruction neural networks to model the distributions of only real faces, which enhances the learned representations to be aware of forgery patterns. SIA (Sun et al., 2022) focuses on self-information attention to enhance the feature representation. UCF (Yan et al., 2023a) separated general deepfake artifacts from domain-specific features, facilitating improved generalization.

In order to train and compare detection methods with each other, training samples and benchmarks naturally began to appear. FaceForensics++ (Rossler et al., 2019) compiled a large-scale corpus of manipulated videos, establishing a foundational benchmark for training and evaluating detection models. Celeb-DF (Li et al., 2020b) introduced higher-fidelity face swaps, highlighting the challenge of subtle artifacts and improved identity preservation. DFDC (Dolhansky et al., 2020) arose from an industry-led initiative, providing a large-scale challenge to benchmark detection methods in diverse real-world conditions. Deepfakes in the Wild (Pu et al., 2021) included videos collected from the internet and captured under less controlled environments, widely used for testing various detection systems. ForgeryNet (He et al., 2021) broadened the range of manipulations, encompassing multiple synthetic techniques across various facial regions. DF-Platter (Narayan et al., 2023) presented a diverse dataset with low- and high-resolution videos and distinct sets for single- and multi-subject deepfakes, with face images of Indian ethnicity.

Overall, these existing works highlight the rapid advancements and growing complexity of deepfake technology. While considerable effort has been directed toward both generation and detection, our work addresses a critical gap in the literature — the lack of a dedicated benchmark dataset designed for video conferencing scenarios.

3. VCF Dataset

The VCF dataset was synthesized using a systematic pipeline designed to emulate realistic deepfake generation processes within video conferencing environments. This involves selecting videos where the scene is typical of video conferencing scenarios, such as a person sitting in front of a monitor and engaging in various activities. The dataset incorporates real-time generation methods and employs a range of manipulations, including adjustments to the input resolution and compression of the final videos. The entire generation pipeline can be seen in Figure 1.

3.1 Source data

The VCF dataset is based on two other open source datasets:

1. **VCD (Video Conference Dataset)** (Naderi et al., 2024): The dataset consists of 160 videos simulating real video conferences, all captured in 1080p resolution. These videos are specifically designed for video conferencing settings, featuring various participants engaging in simple actions in front of a monitor. The scenes are shot under diverse lighting conditions and include different types of backgrounds, such as blurred, natural, virtual, and mobile. This dataset served as the target set for face swap operations, meaning these videos were used as the base for integrating different faces through generative techniques.

2. **LaPa (Landmark guided face Parsing dataset)** (Liu et al., 2020): The large-scale image dataset for face parsing comprises over 22,000 facial images, offering a wide array of variations in expression and pose. Images from this dataset were utilized as sources for deepfake creation, meaning these faces were used as the origins for face swap operations. The dataset's diversity aids in selecting the most suitable face for deepfake generation, enhancing the authenticity and effectiveness of the swaps.

Prior to deepfake generation, the original VCD videos were systematically downsampled to 540p, 360p, and 270p resolutions. This multi-resolution framework enables to study detector robustness to input quality degradation – a critical factor in real-world applications, where attackers may exploit low-resolution video streams to mask manipulation artifacts or accelerate the generation to real-time if they do not have large computing resources.

3.2 Face matching protocol

To maximize the plausibility of synthetic forgeries, source-target pairs were selected using a multivariate similarity metric that prioritizes alignment between LaPa (source) and VCD

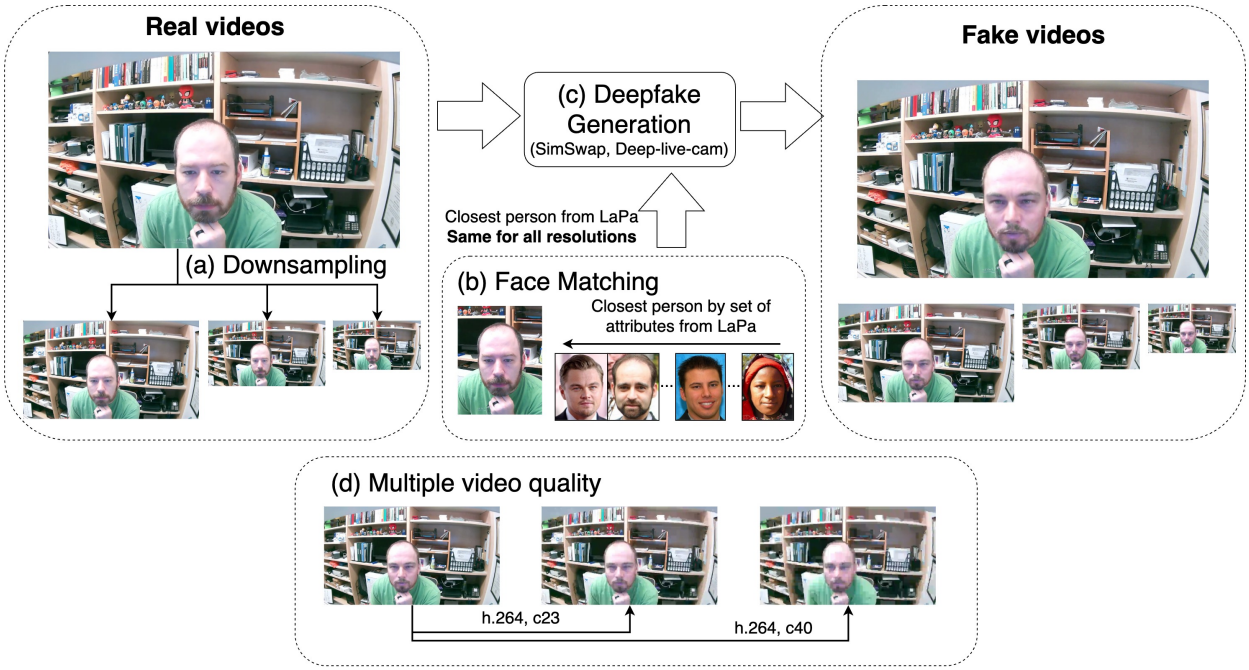

Figure 1. (a) Downsampling videos from 1080p resolution to lower resolutions 540p, 360p and 270p. (b) Match faces from LaPa dataset to every video from VCD dataset by gender, ethnicity, age, and facial hair. (c) Deepfake video generation with Deep-live-cam and SimSwap. (d) Compress every real and fake video with H.264 compression protocol with quantization parameters set to 23 and 40.

(target) subjects based on gender, ethnicity, age, and facial hair characteristics. This strategy simplifies the operation of the deepfake generator, which means that the quality of the output deepfake has also improved, because the generator needs to make fewer transformations. This approach reflects the strategic behavior of malicious actors seeking to minimize visual discrepancies during face-swapping in order to improve the quality of result deepfake.

3.3 Generation methods

Deepfake synthesis utilized two open-source frameworks representative of contemporary face-swapping techniques:

1. **SimSwap** (Chen et al., 2020b) is a framework designed to transfer the identity of an arbitrary source face into an arbitrary target face while preserving the attributes of the target face. It was used in two configurations, processing inputs and outputs at 224x224 and 512x512 resolutions. Accordingly, the first option generated lower-quality deepfakes, while the second required more computing power to ensure real-time generation.

2. **Deep-Live-Cam** (Estanislao, 2024) is a very popular framework for generating forgery videos in real time. It was employed in both standard and ”enhanced” modes, simulating scenarios ranging from casual manipulation attempts to high-effort adversarial attacks.

3.4 Multiple video quality

Following synthesis, the dataset underwent post-processing to emulate real-world video conferencing conditions. All videos were compressed using the H.264 codec – a standard for platforms such as Zoom and Microsoft Teams – with quantization parameters (QP) set to 23 (high quality) and 40 (low quality), simulating network-induced compression artifacts. Additionally, background variations in the original VCD dataset (blurred, virtual, or natural backgrounds) were preserved to evaluate detector resistance to common feature of video conferencing software that allows to hide what lies behind user.

3.5 Dataset Characteristics

The final VCF repository comprises **1,920 authentic video clips**, derived from 160 original recordings across four resolutions (1080p, 540p, 360p, 270p) and three compression rates (raw, 23, 40), alongside **7,680 deepfake variants** generated through combinations of resolution, generator configuration, and compression parameters. This structured diversity enables comprehensive evaluations of detection frameworks across three axes: (1) resolution adaptability, assessing performance degradation as input quality decreases; (2) compression robustness, measuring resilience to H.264 artifacts; and (3) background invariance, quantifying susceptibility to false positives induced by environmental distractions and conference software features.

4. Deepfake Detection Evaluation

We follow the pre-processing, training pipeline, available model weights and use the codebases of DeepfakeBench (Yan et al., 2023b). For the evaluation metric, we report the widely-used video-level Area Under the Curve (AUC) to compare approaches to each other. To provide a comprehensive results for comparison, we evaluated 14 competitive detectors. All detectors are trained on FF++ (c23) (Rossler et al., 2019) and tested in cross-domain manner on Celeb-DF (Li et al., 2020b) and different parts of ours VCF dataset. The results of our experiments can be seen in Table 1.

Among the tested approaches, X-CLIP achieves the highest score (0.862) on Celeb-DF, followed by methods such as SRM (0.840), UCF (0.838), and Capsule (0.834). This result indicates that X-CLIP is particularly effective on more established

Method	Celeb-DF	VCF							
		Full	270p, c40	540p, c23	1080p, raw	DLC	DLC-E	SimS-224	SimS-512
Meso4 (Afchar et al., 2018)	0.529	0.469	0.459	0.469	0.465	0.493	0.437	0.446	0.497
Capsule (Nguyen et al., 2019)	0.834	0.513	0.529	0.516	0.528	0.611	0.415	0.519	0.503
Meso4Inc (Afchar et al., 2018)	0.700	0.514	0.548	0.496	0.511	0.579	0.398	0.561	0.515
SIA (Sun et al., 2022)	0.816	0.524	0.529	0.530	0.544	0.583	0.423	0.554	0.539
F3Net (Qian et al., 2020)	0.787	0.531	0.526	0.539	0.534	0.594	0.415	0.580	0.534
FFD (Dang et al., 2020)	0.743	0.545	0.513	0.552	0.604	0.599	0.435	0.611	0.532
Effnetb4 (Tan and Le, 2019)	0.808	0.553	0.514	0.564	0.558	0.625	0.483	0.564	0.539
SRM (Luo et al., 2021)	0.840	0.554	0.523	0.570	0.592	0.600	0.452	0.600	0.561
CORE (Ni et al., 2022)	0.810	0.563	0.562	0.564	0.586	0.639	0.404	0.617	0.591
Recce (Cao et al., 2022)	0.824	0.563	0.538	0.584	0.597	0.617	0.485	0.592	0.559
SPSL (Liu et al., 2021)	0.799	0.580	0.518	0.599	0.638	0.623	0.406	0.671	0.619
Xception (Rossler et al., 2019)	0.816	0.591	0.544	0.612	0.636	0.642	0.499	0.632	0.590
UCF (Yan et al., 2023a)	0.838	0.599	**0.564**	0.615	0.661	0.633	0.517	0.667	0.580
X-CLIP (Ma et al., 2022)	**0.862**	**0.667**	0.560	**0.714**	**0.804**	**0.698**	**0.616**	**0.736**	**0.619**

Table 1. Evaluation results using different quality and resolution parts and different generation methods from VCF. The metric is the video-level AUC. **Full** - the whole VCF dataset; **270p, c40** - the worst in quality videos from VCF, low-resolution and c40 compression; **540p, c23** - the mid-quality videos from VCF; **1080p, raw** - the best quality videos from VCF, high resolution and no compression; **DLC** - videos generated with Deep-Live-Cam; **DLC-E** - videos generated with Deep-Live-Cam in Enhanced mode; **SimS-224** - videos generated with SimSwap with 224 resolution generator; **SimS-512** - videos generated with SimSwap with 512 resolution generator.

Dataset	Real Videos	Fake Videos	Total Videos	Total Subjects	Generation Techniques	Multi Res./Qual.	VideoConf Setup
FF++ (Rossler et al., 2019)	1000	4000	5000	N/A	4	✓	×
Celeb-DF (Li et al., 2020b)	590	5639	6229	59	1	×	×
DFDC (Dolhansky et al., 2020)	23654	104500	128154	960	8	×	×
DeepfakeTIMIT (Korshunov and Marcel, 2018)	640	320	960	32	2	✓	×
Deepfakes in the Wild (Pu et al., 2021)	1896	1869	3738	N/A	N/A	×	×
ForgeryNet (He et al., 2021)	99630	121617	221247	5400+	36	×	×
DF-Platter (Narayan et al., 2023)	764	132496	133269	454	3	✓	×
VCF (ours)	1920	7680	9600	160	4	✓	✓

Table 2. Comparison with existing open-source deepfake datasets.

deepfake benchmarks, consistently surpassing many baseline and recently proposed deepfake detection models.

When evaluating the same set of methods on our new VCF benchmark — which includes multiple video resolutions and compression levels - overall scores drop relative to Celeb-DF, emphasizing the difficulty of this new dataset. In addition, you can see how randomly the quality of the methods is distributed over the years. Nonetheless, X-CLIP remains among the top performers in many configurations. X-CLIP reaches 0.667 on full VCF dataset, which is the best among the reported methods. Therefore, unless otherwise stated, results for X-CLIP will be presented.

Examining the VCF subsets reveals that many top methods on Celeb-DF do not maintain their performance as the video quality degrades. X-CLIP quality metrics drops from 0.804 to 0.560 when when video quality decreases ("1080p, raw" compared to "270p, c40"). The picture is approximately the same for other methods. This drop highlights a critical need to develop detection schemes that are robust to a wide range of video qualities - an essential real-world scenario in video conferencing. Additionally, UCF slightly outperforms X-CLIP (0.564 vs 0.560), suggesting certain methods are more robust at lower resolutions or higher compression levels, although the difference remains modest.

Method robustness varies by generation type. While X-CLIP consistently occupies a leading position across all four deepfake-generation strategies, its relative advantage over other methods fluctuates considerably. SPSL, for example, emerges as a close competitor on SimSwap data. The shift from Deep-Live-Cam to Deep-Live-Cam in enhanced mode (0.698 vs 0.616), as well as from SimSwap-224 to SimSwap-512 (0.736 vs 0.619), reveals that image generator strength and enhancement steps have a significant effect on detector performance. Most methods struggle as the generation quality increases or when additional synthetic "enhance" operations are applied.

For almost all experiments, the video ROC-AUC metric does not exceed 0.7, which is an unacceptable quality of work for real-world scenarios. These results highlight how current detection methods can fail to generalize across different deepfake pipelines and quality conditions. To ensure reliable detection in real-world video conferencing scenarios, future research should focus on resilience to different resolutions, compression levels, generation models and enhancement algorithms.

5. Comparison to existing deepfake datasets

Table 2 provides a systematic comparison of VCF with widely used deepfake datasets. While existing benchmarks such as DFDC (Dolhansky et al., 2020), ForgeryNet (He et al., 2021) and DF-Platter (Narayan et al., 2023) excel in scale (e.g., DFDC contains 104,500 fake videos), diversity of generation techniques (e.g., ForgeryNet covers 36 manipulation methods) and multi-face manipulations (e.g. this is the feature of DF-Platter), they lack explicit focus on video conferencing contexts. For instance, none incorporate specific environmental setups, platform-standard compression, variable resolutions, and virtual backgrounds — critical factors in real-world communication scenarios. Similarly, datasets like Celeb-DF (Li et al., 2020b) and FF++ (Rossler et al., 2019), though foundational, prioritize facial manipulation fidelity over environmental realism, limiting their utility for evaluating detectors in video conferencing applications.

VCF addresses these gaps by combining scenario-specific realism with technical diversity. With 1,920 authentic and 7,680 synthetic videos, VCF surpasses specialized datasets like DeepfakeTIMIT (Korshunov and Marcel, 2018) in scale while introducing four resolution tiers (1080p, 540p, 360p and 270p) and

three compression levels (raw, c23 and c40) — features absent in most benchmarks. This multi-resolution architecture enables precise analysis of detector robustness to quality degradation, a vulnerability highlighted by our experiments (e.g., X-CLIP's AUC drops from 0.804 on ”1080p, raw” videos to 0.5602 on ”270p, c40” which can be seen in table 1). Furthermore, VCF's curated face-swapping protocol, which pairs LaPa and VCD subjects based on demographic and phenotypic traits, ensures forgeries mimic adversarial attacks aimed at minimizing visual discrepancies. VCF preserves virtual and blurred backgrounds from the original VCD recordings, challenging detectors to distinguish forgery artifacts from environmental noise. This focus on contextual factors makes VCF suited for evaluating detectors in settings where attackers exploit low-quality streams or software features (e.g., background replacement) to evade detection.

In summary, VCF advances the field by prioritizing practical relevance over sheer scale. Its design highlights shortcomings in current detectors when faced with video conferencing-specific perturbations, providing a critical tool for developing robust, context-aware solutions. Future datasets may benefit from integrating VCF's scenario-driven approach while expanding the diversity of generation methods and demographic representation to further close the gap between lab benchmarks and real-world deployment.

6. Conclusion

As deepfake technology evolves, the need for detection frameworks tailored to real-world communication platforms grows increasingly urgent. VCF responds to this demand by providing a standardized, scenario-specific benchmark that challenges detectors with the technical and contextual complexities of video conferencing environments. By facilitating systematic analysis of resolution adaptability, compression robustness, and background invariance, the dataset empowers researchers to develop resilient, context-aware solutions. It provides an opportunity to test deepfake detection methods specifically for the use case of identifying deepfakes in video conferencing scenarios.

References

Afchar, D., Nozick, V., Yamagishi, J., Echizen, I., 2018. Mesonet: a compact facial video forgery detection network. *2018 IEEE international workshop on information forensics and security (WIFS)*, IEEE, 1–7.

Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., Yang, X., 2022. End-to-end reconstruction-classification learning for face forgery detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4113–4122.

Chen, R., Chen, X., Ni, B., Ge, Y., 2020a. Simswap: An efficient framework for high fidelity face swapping. *Proceedings of the 28th ACM international conference on multimedia*, 2003–2011.

Chen, R., Chen, X., Ni, B., Ge, Y., 2020b. Simswap: An efficient framework for high fidelity face swapping. *MM '20: The 28th ACM International Conference on Multimedia*.

Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A. K., 2020. On the detection of digital face manipulation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, 5781–5790.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C. C., 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Estanislao, K., 2024. Deep-live-cam. `https://github.com/hacksider/Deep-Live-Cam`.

He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., Liu, Z., 2021. Forgerynet: A versatile benchmark for comprehensive forgery analysis. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4360–4369.

Kim, K., Kim, Y., Cho, S., Seo, J., Nam, J., Lee, K., Kim, S., Lee, K., 2022. Diffface: Diffusion-based face swapping with facial guidance. arXiv 2022. *arXiv preprint arXiv:2212.13344*, 1(2), 3.

Korshunov, P., Marcel, S., 2018. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.

Li, L., Bao, J., Yang, H., Chen, D., Wen, F., 2020a. Advancing high fidelity identity swapping for forgery detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5074–5083.

Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S., 2020b. Celeb-df: A large-scale challenging dataset for deepfake forensics. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3207–3216.

Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., Yu, N., 2021. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 772–781.

Liu, Y., Shi, H., Shen, H., Si, Y., Wang, X., Mei, T., 2020. A new dataset and boundary-attention semantic segmentation for face parsing. *AAAI*, 11637–11644.

Luo, Y., Zhang, Y., Yan, J., Liu, W., 2021. Generalizing face forgery detection with high-frequency features. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16317–16326.

Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R., 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. *Proceedings of the 30th ACM International Conference on Multimedia*, 638–647.

Naderi, B., Cutler, R., Khongbantabam, N. S., Hosseinkashi, Y., Turbell, H., Sadovnikov, A., Zou, Q., 2024. Vcd: A video conferencing dataset for video compression. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 3970–3974.

Narayan, K., Agarwal, H., Thakral, K., Mittal, S., Vatsa, M., Singh, R., 2023. Df-platter: Multi-face heterogeneous deepfake dataset. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9739–9748.

Nguyen, H. H., Yamagishi, J., Echizen, I., 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2307–2311.

Ni, Y., Meng, D., Yu, C., Quan, C., Ren, D., Zhao, Y., 2022. Core: Consistent representation learning for face forgery detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12–21.

Pu, J., Mangaokar, N., Kelly, L., Bhattacharya, P., Sundaram, K., Javed, M., Wang, B., Viswanath, B., 2021. Deepfake videos in the wild: Analysis and detection. *Proceedings of The Web Conference 2021*.

Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J., 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. *European conference on computer vision*, Springer, 86–103.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M., 2019. Faceforensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.

Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K., 2024. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6527–6536.

Sun, K., Liu, H., Yao, T., Sun, X., Chen, S., Ding, S., Ji, R., 2022. An information theoretic approach for attention-driven face forgery detection. *European Conference on Computer Vision*, Springer, 111–127.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *International conference on machine learning*, PMLR, 6105–6114.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M., 2016. Face2face: Real-time face capture and reenactment of rgb videos. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2387–2395.

Yan, Z., Zhang, Y., Fan, Y., Wu, B., 2023a. UCF: Uncovering Common Features for Generalizable Deepfake Detection. *arXiv preprint arXiv:2304.13949*.

Yan, Z., Zhang, Y., Yuan, X., Lyu, S., Wu, B., 2023b. Deepfakebench: A comprehensive benchmark of deepfake detection. A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (eds), *Advances in Neural Information Processing Systems*, 36, Curran Associates, Inc., 4534–4565.