Beyond DeepFakes: Detecting Facial Retouching with CLIP Adaptation

Berkay Buğra Gök

***Abstract*—We propose a CLIP-based approach for classifying filtered and unfiltered face images. While significant progress has been made in detecting face forgeries such as DeepFakes, Face2Face, FaceSwap, and NeuralTextures, comparatively little attention has been given to face retouching and beauty filters. This omission is increasingly important as cosmetic filters become more popular on social media each day and pose risks for security-sensitive systems vulnerable to deepfake and morphing attacks. Beauty filters differ fundamentally from conventional face forgeries, as they apply constrained, image-space cosmetic edits such as skin smoothing and color adjustment—while preserving facial identity and expression dynamics. As a result, widely used forgery detection datasets, including FaceForensics++ (FF++) c23, do not explicitly represent such manipulations. Our method builds on a slightly modified model trained exclusively on FF++ c23, which contains only face forgery techniques. Despite the absence of direct supervision for beauty filters, our results indicate that CLIP-based representations can generalize beyond the training distribution and effectively discriminate between filtered and unfiltered facial images. These findings highlight the potential of vision–language models for more generalizable face manipulation detection.**

***Index Terms*—facial retouching detection, image manipulation, deepfake detection, CLIP, vision–language models, image forensics, digital media forensics**

I. Introduction

The rapid evolution of generative models has prioritized the detection of *explicit forgeries*, such as identity swaps and reenactments. While recent detectors achieve high performance by modeling spatial and cross-modal artifacts [2]–[4], they largely overlook *facial retouching* (beauty filtering). Unlike deepfakes, retouching introduces subtle, identity-preserving edits that, despite their benign usage, can significantly compromise biometric security systems [5].

A primary challenge is the lack of appropriate supervision. Benchmarks like FaceForensics++ (FF++) [1], [7] focus on neural rendering rather than the texture smoothing and geometric warping typical of cosmetic filters. Consequently, models trained on explicit forgeries often fail to generalize to the distinct domain of retouched faces.

In this work, we investigate this gap using a CLIP-based forgery detector [4]. We observe that standard adaptations often attend to spurious background cues rather than facial manipulation. To address this, we introduce a *central patch selection* strategy that restricts the model's attention to the facial region. This forces the learner to focus on relevant cues—such as skin texture and shading consistency—while suppressing background artifacts. Our results show that this simple spatial constraint enables a model trained solely on FF++ to effectively discriminate beauty filters, demonstrating that large vision–language models possess transferable forensic knowledge.

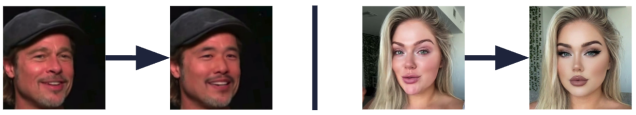

Figure 1: Deepfake vs. Retouching

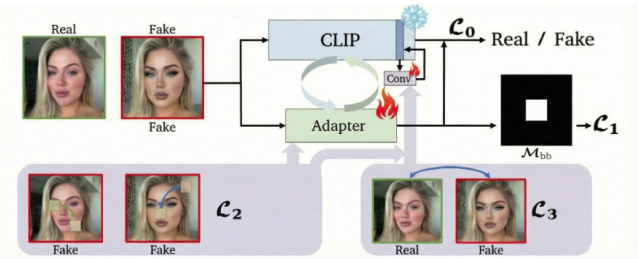

Figure 2: High Level View of the Parallel Model

II. Related Work

A. CLIP-based Deepfake and Synthetic Media Detection

Recent work has leveraged CLIP for synthetic media detection, though its standard features lack explicit sensitivity to forensic traces. To bridge this gap, specialized adaptation mechanisms have been developed. Liu *et al.* proposed the Forgery-aware Adaptive Transformer (FAT) [3], which integrates spatial and frequency-domain adapters. By aligning visual features with forgery-related text prompts, FAT captures low-level artifacts alongside semantic content, significantly improving robustness against various generative models. However, this model uses CLIP as a mere feature extractor, meaning that the knowledge carried by the CLIP model in earlier layers is lost towards the end layers. To overcome this limitation, alternative methods that better integrate with the CLIP are proposed.

In the specific domain of face manipulation, Cui *et al.* introduced the Forensics Adapter [4]. This approach injects a lightweight module that interacts with CLIP's visual tokens to model blending boundaries and manipulation traces. By avoiding full backbone fine-tuning, it preserves CLIP's generalizability while enabling strong performance on deepfakes.

However, these methods focus primarily on explicit forgeries such as identity swapping and reenactment. Their effectiveness on subtle, identity-preserving manipulations like facial retouching remains largely unexplored, a gap this work aims to address.

B. Face Retouching and Beauty Filter Analysis

Compared to face forgery detection, facial retouching has received significantly less attention in the forensic literature. Most existing benchmarks and detectors are designed to identify identity-altering manipulations, while beauty filters introduce subtle, identity-preserving changes that primarily affect skin texture, shading, and local facial geometry. These edits are visually natural and therefore more difficult to distinguish from genuine imagery.

RetouchingFFHQ [6] represents one of the first large-scale efforts to explicitly model this problem. The dataset provides fine-grained annotations of multiple retouching operations applied to high-quality face images and frames retouching analysis as a regression task, where the goal is to estimate the strength of different cosmetic edits rather than to perform binary classification. This formulation enables detailed modeling of retouching severity but requires specialized supervision and filter-specific labels.

In contrast, our work studies whether detectors trained only on forgery-based supervision can generalize to the detection of retouched faces without access to retouching labels or filter strength annotations. This setting is more aligned with real-world forensic deployment, where the type and degree of cosmetic editing are typically unknown.

III. Methodology

A. Overview

We adapt the Forensics Adapter architecture to the task of facial retouching detection, aiming to generalize from deepfake supervision on FaceForensics++ (FF++) c23 to unseen cosmetic filters. The model is built on a CLIP vision–language backbone and is trained only with forgery labels, without access to any retouching annotations. The goal is to evaluate whether forensic representations learned from explicit face forgeries can transfer to identity-preserving beauty edits.

B. CLIP with Parallel Adapter Architecture

We use a frozen CLIP image encoder (ViT-L/14) to preserve its strong semantic representations and pair it with a lightweight parallel Adapter based on a Tiny Vision Transformer. The input image is tokenized into 16×16 patches and processed by two interacting streams:

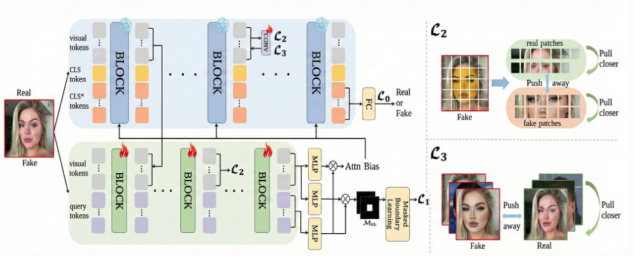

Figure 3: CLIP with parallel Forensics Adapter. The Adapter absorbs intermediate CLIP features and injects attention bias Δ to emphasize manipulation cues.

- **CLIP stream**, which extracts global semantic features.
- **Adapter stream**, which focuses on localized manipulation traces.

The two streams exchange information in a bi-directional manner. The Adapter receives intermediate CLIP features from layers 1, 8, and 16, allowing it to align its local representations with CLIP's semantic structure. In turn, the Adapter injects an attention bias Δ into CLIP's self-attention layers, steering the model toward manipulation-relevant regions.

C. Central Patch Selection

Beauty filters primarily modify facial appearance, while background and image borders often contain spurious cues correlated with dataset-specific preprocessing. To suppress such artifacts, we restrict boundary supervision to a central facial region. Specifically, only the central 8×8 patch grid is used to compute boundary-related losses, ensuring that learning focuses on the face rather than on background or frame edges.

Figure 4 visualizes the selected central region.

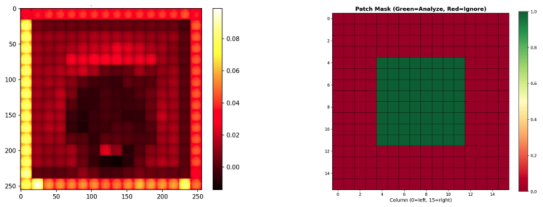

Figure 4: Central patch selection. Only the central 8×8 region is used for boundary supervision.

D. Training Objectives

The model is trained using a weighted combination of classification and contrastive losses.

1) Classification Loss: A standard cross-entropy loss L_0 is applied to the CLIP classification token to distinguish real and manipulated images.

2) Boundary Loss: The Adapter predicts blending boundaries from CLIP features. Let V_{bb} denote Adapter features and Q the corresponding CLIP queries. The boundary map is computed as

$$M_{bb} = \mathrm{Conv}(V_{bb} Q^{\top}). \tag{1}$$

A masked mean squared error loss is then applied:

$$L_1 = \mathrm{MSE}(M_{bb} \odot B, M'_{bb}), \tag{2}$$

where B denotes the central 8×8 mask and M'_{bb} is the ground-truth boundary map.

3) Patch-wise Contrastive Loss: To encourage separation between real and manipulated patches, we apply a contrastive loss:

$$L_2 = -\log \frac{\exp(\delta(x_i, x_j)/\tau)}{\exp(\delta(x_i, x_j)/\tau) + \sum_{x_k \in X^*} \exp(\delta(x_i, x_k)/\tau)}, \tag{3}$$

where $\delta(\cdot,\cdot)$ denotes cosine similarity, τ is a temperature parameter, and X^* denotes negative patch samples.

4) Sample-wise Contrastive Loss: At the image level, a contrastive loss is applied to global embeddings:

$$L_3 = -\log \frac{\exp(\delta(X_i, X_j)/\tau)}{\exp(\delta(X_i, X_j)/\tau) + \sum_{X_k \in D^*} \exp(\delta(X_i, X_k)/\tau)}, \tag{4}$$

where D^* denotes negative image samples.

The total loss is a weighted sum:

$$L = \lambda_0 L_0 + \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3. \tag{5}$$

IV. Results

We evaluate the proposed center-focused Adapter model on the face filter test set, which contains both filtered (retouched) and unfiltered facial videos unseen during training. All models are trained exclusively on FaceForensics++ (c23) using forgery supervision only, and no retouching annotations are used at any stage. The face filter test set is a small set with nearly 20 videos, up to 32 frames per video.

A. Overall Performance

Figure 5 summarizes the quantitative performance of the model. The receiver operating characteristic (ROC) curve yields an area under the curve (AUC) of approximately 0.81, indicating moderate separability between filtered and unfiltered samples. The equal error rate (EER) is observed at 0.29, reflecting the inherent difficulty of detecting subtle, identity-preserving facial retouching compared to explicit face forgeries.

Using the default decision threshold of 0.5, the model achieves an overall accuracy of 72.2%. Threshold sweeping shows that performance remains stable in a broad region around this operating point, with a maximum accuracy of approximately 73.1% achieved at a slightly lower threshold. This suggests that the classifier is not overly sensitive to threshold selection and maintains consistent behavior across operating conditions.

B. Prediction Score Distributions

The prediction score distributions for real and filtered samples are shown in Figure 5. Filtered samples tend to produce higher confidence scores, while real samples are concentrated at lower values. Nevertheless, a noticeable overlap remains between the two distributions, particularly in the mid-confidence

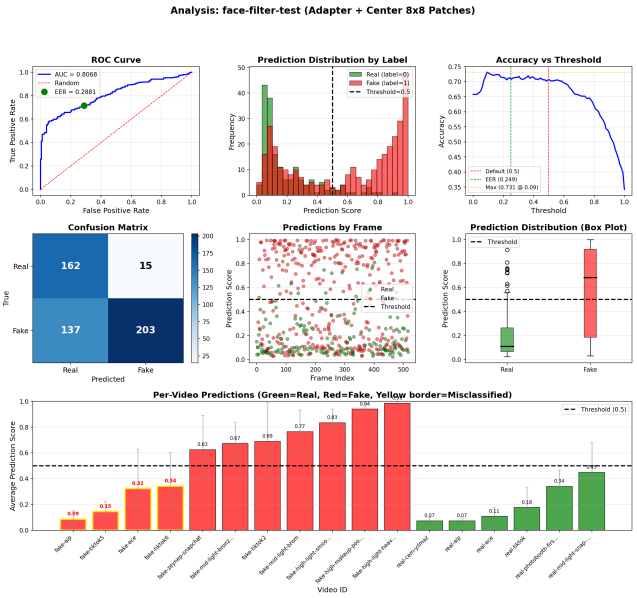

Figure 5: Model performance results on custom test set

range. This overlap is consistent with the visual subtlety of many beauty filters and indicates that some retouched samples closely resemble unfiltered faces at the pixel level.

Frame-level predictions exhibit substantial variance, as shown by the per-frame scatter plots. Individual frames affected by compression artifacts, motion blur, or lighting changes can lead to uncertain predictions. However, aggregating predictions at the video level significantly improves robustness. As illustrated in the per-video analysis, most filtered and unfiltered videos are correctly separated based on their average prediction scores, reducing the impact of noisy frames.

C. Error Analysis

The confusion matrix reveals that misclassifications are asymmetric. False negatives (filtered videos classified as real) predominantly occur in heavily compressed videos or in scenarios with low or uneven lighting, where fine-grained manipulation traces are partially obscured. Conversely, false positives are less frequent and often correspond to challenging real videos exhibiting strong post-processing or capture artifacts.

D. Effect of Central Patch Masking

Compared to a baseline Adapter configuration without spatial restriction, incorporating central 8×8 patch masking yields a small but consistent performance improvement. By limiting boundary-related supervision to the central facial region, the model is less influenced by background textures and frame-level artifacts that are weakly correlated with retouching. This results in slightly improved AUC and more stable threshold behavior, particularly in video-level aggregation. While the gain is modest, it supports the hypothesis that spatially focused supervision better aligns the learned forensic features with facial manipulation cues.

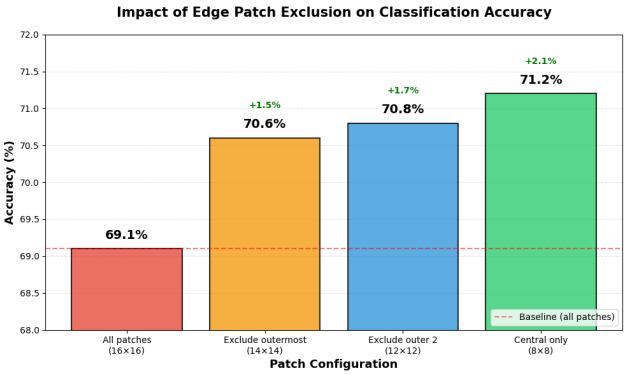

Figure 6: The effect of using different sizes of patch masks

Overall, these results demonstrate that forensic representations learned from explicit face forgeries can partially transfer to the detection of cosmetic facial filters, despite the absence of retouching supervision during training. The remaining performance gap highlights the need for more robust representations that are resilient to compression and capture-related degradation.

V. Conclusion

This work demonstrates that CLIP-based forensic models trained solely on explicit face forgery datasets can generalize to the task of facial retouching detection when spatial attention is properly constrained. By restricting boundary-related supervision to central facial patches, the proposed center-focused Adapter model achieves a consistent performance improvement, increasing classification accuracy from 69.1% to 72.2%. This finding confirms that controlling spatial attention is critical for detecting subtle, identity-preserving manipulations such as beauty filters, where background and border cues are largely uninformative.

Experimental results further indicate that video quality plays a significant role in detection performance. Compression artifacts and suboptimal lighting conditions substantially degrade model reliability, even when frames are downsampled to moderate resolutions. In particular, videos sourced from online platforms tend to perform worse than recordings captured under controlled conditions, highlighting the sensitivity of forensic features to encoding and post-processing effects.

While the proposed approach demonstrates promising cross-domain generalization without retouching supervision, the remaining performance gap suggests clear directions for future work. In particular, retraining or fine-tuning on task-aligned datasets such as Flickr-Faces-HQ-Retouching (FFHQR) may improve robustness to real-world retouching variations. Overall, this study provides evidence that forgery-trained forensic representations can serve as a viable foundation for facial retouching detection, provided that spatial supervision is carefully designed.

VI. Acknowledgements

Thanks Lale Akarun for their valuable guidance and supervision on determining the direction of the project.

References

[1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," *arXiv preprint arXiv:1901.08971*, 2019.

[2] R. Kundu, H. Xiong, V. Mohanty, A. Balachandran, and A. K. Roy-Chowdhury, "Towards a Universal Synthetic Video Detector: From Face or Background Manipulations to Fully AI-Generated Content," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[3] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, and Y. Zhao, "Forgery-aware Adaptive Transformer for Generalizable Synthetic Image Detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[4] X. Cui, Y. Li, A. Luo, J. Zhou, and J. Dong, "Forensics Adapter: Adapting CLIP for Generalizable Face Forgery Detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] S. Concas, S. M. La Cava, A. Panzino, G. Orrù, E. Masala, and G. L. Marcialis, "Deceptive Beauty: Evaluating the Impact of Beauty Filters on Deepfake and Morphing Attack Detection," in *Proc. IEEE International Joint Conference on Biometrics (IJCB)*, 2023.

[6] Q. Ying, J. Liu, S. Li, H. Xu, Z. Qian, and X. Zhang, "RetouchingFFHQ: A Large-scale Dataset for Fine-grained Face Retouching Detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models from Natural Language Supervision," arXiv preprint arXiv:2103.00020, 2021.