Beyond DeepFakes: Detecting Facial Retouching with CLIP Adaptation

Berkay Buğra Gök

CMPE537 - Computer Vision - Lale Akarun - 2025/2026 Fall

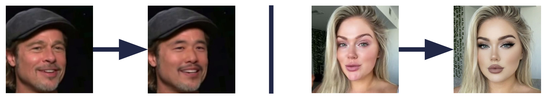
Abstract

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.

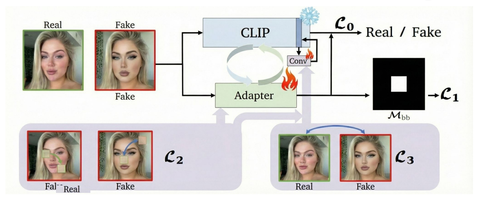
Beauty filters differ fundamentally from conventional face forgeries, as they apply constrained, image-space cosmetic edits such as skin smoothing and color adjustment—while preserving facial identity and expression dynamics. As a result, widely used forgery detection datasets, including FaceForensics++ (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.

Introduction

Face manipulation detection has primarily focused on identifying explicit forgeries such as identity replacement and expression reenactment. In contrast, beauty filters apply subtle, identity-preserving cosmetic edits that are increasingly prevalent in social media content and can interfere with security-sensitive face analysis systems. Existing large-scale datasets, including FaceForensics++ (FF++) c23, contain only forgery-based manipulations and do not explicitly model beauty filter artifacts, limiting direct supervision for this task.



In this work, we investigate whether a CLIP-based model trained on forgery data can generalize to the detection of beauty filters. To address spurious attention to background and image boundaries observed in prior approaches, we introduce a central patch selection strategy that restricts the model's focus to facial regions. This design emphasizes filter-relevant cues while suppressing non-facial artifacts, enabling more robust discrimination between filtered and unfiltered face images.



Methodology

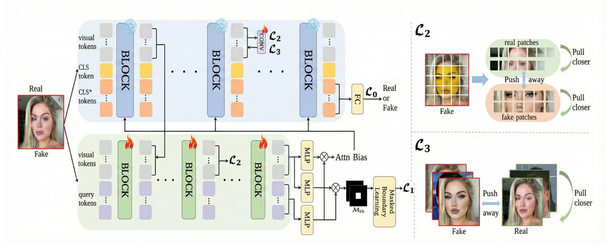
Our approach adapts the Forensics Adapter architecture to the task of face retouching detection, aiming to generalize from deepfake supervision on FaceForensics++ (FF++) c23 to unseen cosmetic filters using a CLIP-based vision–language model.

Architecture: CLIP with Parallel Adapter

We employ a frozen CLIP image encoder (ViT-L/14) to retain its strong semantic representations, complemented by a lightweight parallel Adapter (Tiny ViT) to capture fine-grained manipulation cues. The input image is tokenized into 16×16 patches and processed through a dual-stream design:

- CLIP stream: extracts global semantic features.
- Adapter stream: focuses on localized manipulation traces.

The two streams interact bi-directionally. The Adapter absorbs intermediate CLIP features (layers 1, 8, and 16) to guide local representation learning, and enhances CLIP by injecting an attention bias (Δ) into its self-attention layers, steering focus toward manipulated regions.



Training **Objectives**

The model is trained using a weighted combination of classification and contrastive objectives:

$$L = \lambda_0 L_0 + \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3$$

Classification Loss (L_0). Standard cross-entropy loss applied to the CLIP classification token to distinguish real and manipulated images.

Boundary Loss (L_1). Masked MSE loss supervising blending boundary prediction, restricted to the central 8 × 8 patch region to avoid border artifacts:

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

Patch-wise Contrastive Loss (L_2). Encourages separation between real and manipulated patches by maximizing intra-class similarity and minimizing inter-class similarity:

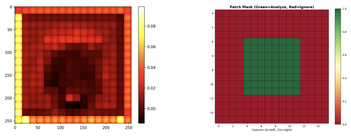
$$L_2 = -\log \frac{\exp(\delta(x_i, x_j)/\tau)}{\exp(\delta(x_i, x_j)/\tau) + \sum_{x_k \in X^*} \exp(\delta(x_i, x_k)/\tau)}$$

Sample-wise Contrastive Loss (L_3). Operates on global embeddings to cluster real samples and separate manipulated samples in feature space:

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

Center-Focused Boundary Learning (Our Modification)

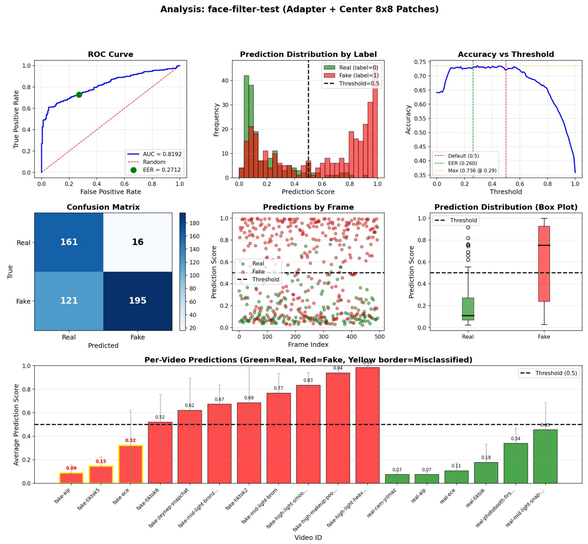
The original model applies a Masked Blending Boundary Loss across all patches. We observe that this encourages reliance on border artifacts (e.g., compression and watermarks) prevalent in FF++. To better align the model with beauty filter detection, we restrict boundary learning to the central 8×8 patch region. This region contains the majority of facial features affected by cosmetic filters.



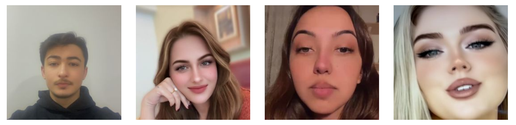
Results

Figure below summarizes the performance of the proposed center-focused adapter model on the face filter test set. The model achieves an **AUC of 0.82** with an **EER of 0.27**, indicating moderate separability between filtered and unfiltered samples. Using the default decision threshold of 0.5 yields an overall **accuracy of 72.2%**, with threshold analysis showing stable performance around this operating point.

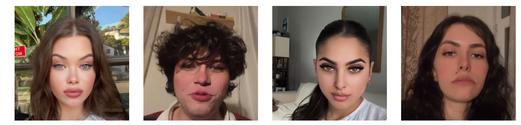
Prediction score distributions reveal a clear shift between real and filtered samples, though partial overlap remains, particularly for heavily compressed or low-quality videos. Frame-level predictions exhibit increased variance, while **video-level aggregation improves robustness**, correctly separating most filtered and unfiltered videos. Misclassifications predominantly occur in samples affected by strong compression or suboptimal lighting, consistent with observed sensitivity to video quality.



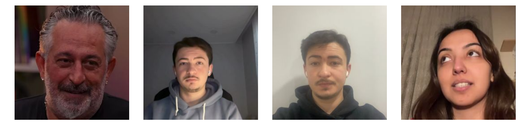
Some of the Correctly Identified Filtered Images



Some of the Incorrectly Identified Filtered Images



Some of the Correctly Identified Unfiltered Images



Conclusion

This work demonstrates that CLIP-based models trained on forgery datasets can generalize to face retouching detection when spatial attention is properly constrained. Restricting learning to central facial patches yields a consistent performance gain, improving accuracy from 69.1% to 72.2%, and confirms that controlling attention is critical for detecting subtle and also identity-preserving manipulations.

Our experiments further reveal that video compression and lighting conditions significantly impact performance, even at reduced spatial resolutions. Heavily compressed and low-light videos degrade detection accuracy, with online video sources performing worse than recordings captured under controlled conditions, indicating sensitivity to encoding artifacts. Finally, we identify Flickr-Faces-HQ-Retouching (FFHQR) as a more suitable dataset for future retraining, offering task-aligned supervision for robust beauty filter detection.

Acknowledgements

This work builds upon the **Forensics Adapter** framework proposed by Cui *et al.* (Ocean University of China; Southwest Jiaotong University), which serves as the primary architectural foundation of our approach. We also acknowledge prior studies on the impact of beauty filters on face analysis systems by Concas *et al.* (University of Cagliari), and the **RetouchingFFHQ** dataset introduced by Ying *et al.* (Fudan University and NVIDIA), which informed our understanding of fine-grained face retouching artifacts. Finally thanks to our instructor, **Lale Akarun**, for valuable guidance and feedback throughout this work.