Guarding Digital Identity: Attention-Guided Fusion for Detecting Forged ID Documents (Student Abstract)

Gargi Surendra Yeole, Poulomi Bhattacharya, Akshay Agarwal

Trustworthy BiometraVision Lab, IISER Bhopal, India
{surendra21, poulomi24, akagarwal}@iiserb.ac.in

Abstract

Government verification systems are increasingly relying on internet-based platforms, where users authenticate their identities by uploading images captured with ordinary mobile devices. However, the rapid advancements in generative algorithms have enabled the creation of highly realistic forged ID cards that can easily bypass such verification pipelines. These forgeries are not restricted to a single modality; they may target facial imagery, textual content, or both, posing significant challenges to existing detection approaches. We present a framework that analyzes visual features for ID forgery detection by integrating feature fusion with attention mechanisms, leveraging both convolutional neural network (CNN) architectures, such as ResNet-50 and EfficientNet, and transformer-based models, including ViT-16 and Swin Transformer. This study emphasises the significance of feature fusion and attention-driven representation learning in developing robust and trustworthy ID forgery detection systems for real-world deployment.

Introduction

The rapid advancement of generative models and image editing tools has made it increasingly easy for malicious actors to forge identity documents, posing serious threats to security-critical applications such as Know Your Customer (KYC), border control, and remote identity verification. Manual inspection is often unreliable and inefficient, while conventional automated methods struggle to capture the diverse visual cues present in modern, high-quality forgeries. These challenges have driven research toward deep learning–based solutions that integrate complementary visual representations and aim to develop robust, generalizable systems for detecting document forgeries.

Recent research in document forgery detection has progressed along several directions. Early works used camera model fingerprints and device-specific artifacts to expose manipulations (Cozzolino and Verdoliva 2019), while later studies introduced hybrid CNN–Transformer models that jointly capture local textures and global semantics (George and Marcel 2025). Benchmark datasets, such as FantasyID, have enabled standardised evaluation under realistic digital and printed attacks (Korshunov et al. 2025). Transformer-based fusion frameworks, which leverage multimodal cues such as RGB and sensor noise, have demonstrated improved robustness (Guillaro et al. 2023). Building on these advances, we propose a framework that integrates feature fusion with a self-attention mechanism using complementary CNN (ResNet-50, EfficientNet-B4) and Transformer (ViT-16, Swin-T) backbones for robust ID forgery detection.

Proposed Attention-Guided Network

In this research, we develop a robust detection framework by combining feature fusion with a self-attention mechanism as depicted in Figure 1. Specifically, we employ a diverse set of pre-trained feature extractors, including convolutional neural networks (CNNs) such as ResNet-50 and EfficientNet-B4, as well as transformer-based models like Swin Transformer (Swin-T) and ViT-16. Feature embeddings from different backbones are concatenated in pairs to leverage their complementary strengths, and a one-head self-attention layer is then incorporated. This module projects the fused input into query-key-value spaces, computes scaled dot-product attention, and refines the attended features through a linear transformation to yield context-aware representations that highlight the most discriminative cues. To thoroughly assess the contribution of each component, we also conduct ablation studies: first, by evaluating each backbone independently without fusion, and second, by excluding the self-attention module.

Experimental Setup

We evaluate our approach on the FantasyID dataset (Korshunov et al. 2025), which contains 786 bona fide and 1,572 manipulated IDs in multiple non-English languages, captured across diverse mobile devices. Manipulations include face swaps and text edits, with two attack types. Images are resized to 224×224, ImageNet-normalized, and split into training, validation, and testing sets as 1,519, 380, and 459, with class imbalance handled via a weighted sampler. Training uses RandAugment, flips, rotations, color jitter, and random crops. Fused backbone features pass through an attention layer and a lightweight classifier (Linear–ReLU–Dropout 0.5–Linear). Models are trained in PyTorch with AdamW (lr 3×10^{-5}, weight decay 0.01),

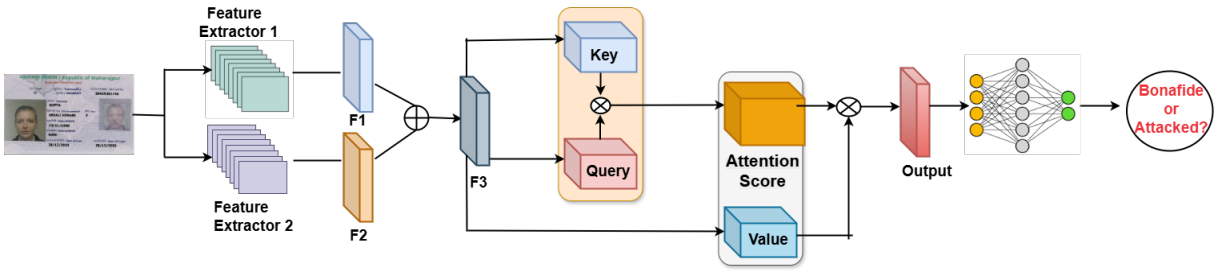


Figure 1: Attention-guided fusion framework, in which feature extractor 1 and feature extractor 2 correspond to different CNN- and Transformer-based models.

Network-1 ↓ Network-2 →	Accuracy	Network fusion							
		ResNet50		ViT-b-16		Swin-T		EfficientNet-B4	
		w/o Attn.	w/ Attn.	w/o Attn.	w/ Attn.	w/o Attn.	w/ Attn.	w/o Attn.	w/ Attn.
ResNet50	0.66	–	–	0.69	0.81	0.96	0.94	0.86	0.87
ViT-B-16	**0.72**	0.69	0.81	–	–	**0.96**	**0.95**	0.76	0.73
Swin Trans.	0.67	0.96	0.94	**0.96**	**0.95**	–	–	**0.96**	**0.95**
EfficientNet-B4	0.62	0.86	0.87	0.76	0.73	**0.96**	**0.95**	–	–

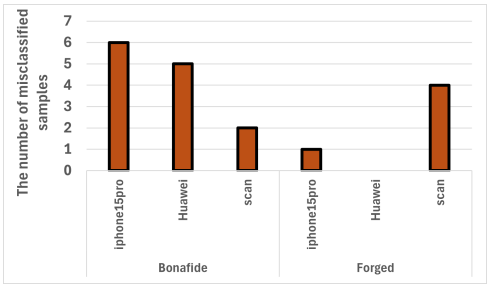
Table 1: Performances of single-extractor and dual-extractor models with and without attention. The Accuracy column (second column) shows the performance of each backbone when used alone. The remaining columns report results when the backbone in the row is fused with the backbone in the first column; each pair is shown with and without the attention module (w/o Attn. vs. w/ Attn.). – represents that the fusion of the same network has not been performed to avoid replication. It shows that the proposed fusion drastically improved the ID forgery detection compared to the best value of $\mathbf{0.72}$ obtained with ViT-B16 alone.

CosineAnnealingLR, label-smoothed cross-entropy ($\epsilon = 0.1$), and gradient clipping (max-norm 1.0).

ID Forgery Detection Results and Analysis

The performance of the proposed attention-guided fusion framework is reported in Table 1. We find that individual CNN or Transformer backbones perform relatively weaker compared to their fused counterparts. For instance, ResNet-50 alone achieves 66% accuracy, while ViT yields 69%; when combined with attention, performance improves to 81%. A similar pattern is observed with Swin-T, which increases from 67% accuracy on its own to 96% when fused with ResNet-50. Likewise, EfficientNet-B4 alone achieves 62%, but its fusion with Swin-T reaches 95%. These results highlight that CNNs, which excel at capturing fine-grained local textures, and Transformers, which model global semantic dependencies, provide complementary information. Their fusion enables richer feature representations, while the attention module further emphasizes the most discriminative cues, resulting in substantial improvements in accuracy.

We further analyze Figure 2(a), which shows that across all fusion settings, false negatives (manipulated IDs mislabeled as bona fide) outnumber false positives, indicating that detecting subtle forgeries is harder than rejecting genuine documents. Figure 2(b) provides examples of bona fide mislabeled as attacked and attacked mislabeled as bona fide.



(a)

(b)

Figure 2: (a) Device-specific misclassification analysis for the EfficientNet-B4 + Swin Transformer model, showing false negatives (attacks classified as bona fide) and false positives (bona fide classified as attacks) across three device types. (b) Example visualizations of misclassified samples, left: attacked samples predicted as bona fide; right: bona fide samples predicted as attacks.

Conclusions and Future Work

Our attention-guided fusion framework effectively combines CNNs and Transformers to detect manipulated ID documents, achieving strong generalization across devices and manipulation techniques. This highlights the promise of attention-driven fusion for robust identity verification. In future work, we plan to explore lightweight architectures for real-time deployment and evaluate scalability on larger, more diverse ID datasets.

References

Cozzolino, D.; and Verdoliva, L. 2019. Noiseprint: A CNN-based camera model fingerprint. *IEEE TIFS*, 15: 144–159.

George, A.; and Marcel, S. 2025. EdgeDoc: Hybrid CNN-Transformer Model for Accurate Forgery Detection and Localization in ID Documents. *arXiv preprint arXiv:2508.16284*.

Guillaro, F.; Cozzolino, D.; Sud, A.; Dufour, N.; and Verdoliva, L. 2023. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *IEEE/CVF CVPR*, 20606–20615.

Korshunov, P.; Mohammadi, A.; Vidit, V.; Ecabert, C.; and Marcel, S. 2025. FantasyID: A dataset for detecting digital manipulations of ID-documents. *arXiv preprint arXiv:2507.20808*.