# Guarding Digital Identity: Attention-Guided Fusion for Detecting Forged ID Documents

**Gargi Surendra Yeole, Poulomi Bhattacharya, Akshay Agarwal**

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

## Abstract

Government verification systems increasingly rely 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 convolution neural network (CNN) architectures, ResNet-50, EfficientNet, and transformer-based models, ViT-16, and Swin Transformer. This study highlights the importance of feature fusion and attention-driven representation learning in building robust and trustworthy ID forgery detection systems for real-world deployment.

## Introduction

Advancements in generative models and image editing tools have made it increasingly simple for malicious actors to forge identity documents. Such forgeries pose a significant risk to security-sensitive applications such as Know Your Customer (KYC), border control, and remote identity verification. Manual inspection is both error-prone and inefficient, while conventional methods often fail to capture the diverse visual cues present in sophisticated forgeries. This has motivated the development of deep learning approaches that integrate complementary visual features and design robust systems capable of generalizing across varied forgery scenarios.

Recent research in document forgery detection has explored several directions. Early works focused on camera model fingerprints and device-level artifacts to detect inconsistencies in manipulated images (Cozzolino and Verdoliva 2019). More advanced methods employ hybrid CNN-Transformer models that capture both local texture details and global semantics, showing strong results in identity document forensics (George and Marcel 2025). Benchmark datasets, such as FantasyID, have enabled systematic evaluation under realistic digital and printed attack conditions (Korshunov et al. 2025). In addition, transformer-based fusion frameworks that leverage multi-modal cues (e.g., RGB and sensor noise) have demonstrated improved robustness against subtle manipulations (Guillaro et al. 2023). Beyond model architectures, researchers are also investigating approaches that ensure generalization to unseen forgery types and preserve user privacy during training and evaluation. Collectively, these works highlight an active and evolving research landscape aimed at building generalizable, secure, and privacy-aware ID forgery detection systems.



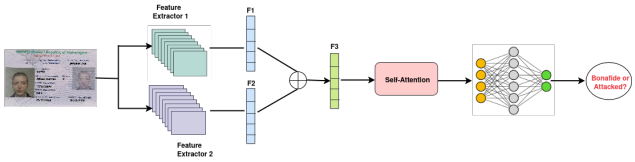


Figure 1: Attention-guided fusion framework, in which feature extractor 1 and feature extractor 2 correspond to different CNN- and Transformer-based models.

## 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 pretrained feature extractors, including convolutional neural networks (CNNs) such as ResNet-50 and EfficientNet-B4, as well as transformer-based models such as Swin Transformer (Swin-T) and ViT-16. Feature embeddings from different backbones are fused in pairs to leverage their complementary strengths, after which a self-attention layer is 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 to measure its impact on the overall performance.

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.

| Network-1 ↓<br>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** | – | – |

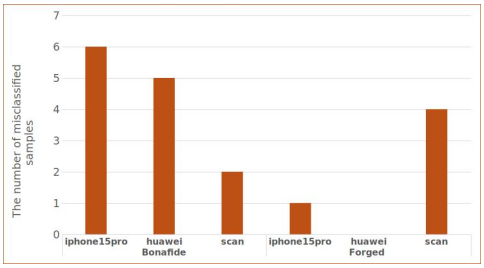


Figure 2: Device-Specific misclassification for the EfficientNet-B4 + Swin transformer model. The graphs show the counts of false negatives (attack as Bonafide) and false positives (Bonafide as Attack) across three device types. This pattern of device-specific vulnerability was consistent with other models evaluated.

## Experimental Setup

We conduct experiments on the FantasyID dataset (Korshunov et al. 2025), which contains ID documents in multiple non-English languages to capture cultural variations. The dataset includes 786 bona fide IDs captured on different mobile devices and 1,572 manipulated IDs created using face-swapping methods (InSwapper, Facedancer) and text inpainting techniques (TextDiffuser2), with two attack types combining face and text manipulations. All images are resized to $224 \times 224$ and normalized with ImageNet statistics. We use 1,519 images for training, 380 for validation, and 459 for testing, and address class imbalance with a weighted random sampler based on class frequencies. Training images are augmented with RandAugment, flips, rotations, color jitter, and random crops, while validation and test sets use only resizing and normalization. The fused backbone features followed by an attention layer are sent to a lightweight classification layer that consists of a linear layer, ReLU activation, dropout (0.5), and a second linear layer for binary classification. Training is performed in PyTorch with AdamW (learning rate $3 \times 10^{-5}$, weight decay 0.01), CosineAnnealingLR scheduling, label-smoothed cross-entropy loss ($\epsilon = 0.1$), 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, leading to substantial accuracy gains.

We further analyze misclassifications in Figure 2. Across all fusion settings, false negatives, manipulated IDs misclassified as bonafide, are more frequent than false positives. This suggests that while the fused models are highly effective overall, distinguishing subtle manipulations in forged IDs remains more challenging than avoiding genuine documents.

## 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*.