DeepID Challenge: Incode Team Solution

Renat Khizbullin, Lazar Lazarevic, Anastasiia Evteeva, Alexey Tochin, and Aleksei Grishin
Incode Technologies Inc.
{renat.khizbullin, lazar.lazarevic, anastasiia.evteeva, alexey.tochin, aleksei.grishin}@incode.com

Abstract

Document forgery detection is increasingly critical due to the rise of deepfake and image manipulation technologies. We present Incode's solution for the ICCV 2025 DeepID challenge [1], which combines three complementary detectors: a general-purpose forgery detector (TruFor), a portrait manipulation detection model, and a text alteration detection model. Our system uses domain-adaptive thresholding for robust decision-making across synthetic and real-world identity documents, achieving second place overall in the challenge. Extensive ablation studies highlight the contribution of each component and the importance of domain calibration.

1. Introduction

The proliferation of deepfake technologies and advanced image manipulation tools has led to a surge in document fraud, posing significant challenges for identity verification systems. In particular, the detection of forged identity documents—whether through face-swapping, textual alteration, or other digital manipulations—remains a critical concern for both industry and research communities.

Recent competitions and benchmarks [5, 6, 11] have highlighted the diversity and complexity of forgeries encountered in real-world scenarios, ranging from synthetic "fantasy" IDs to genuine documents with subtle tampering. Effective detection requires robust, adaptable models capable of generalizing across domains and manipulation types.

In this paper, we present Incode's solution for the ICCV 2025 DeepID challenge on document forgery detection. Our approach integrates multiple specialized detectors and leverages domain-adaptive calibration to address the unique characteristics of both synthetic and real-world ID datasets. We provide a detailed analysis of our method, experimental results, and ablation studies to illustrate the strengths and limitations of our system.

Model	F1 on Fantasy	F1 on Private	Aggregate F1
Sunlight	**0.9906**	0.7191	**0.8006**
Ours	0.8679	**0.7534**	0.7878
AG	0.9581	0.7108	0.7850

Table 1. Final results on the challenge test sets. Our model ranked second overall.

2. Method

2.1. Overview

Our proposed solution integrates three complementary forgery detection models (Fig. 1):

1. **TruFor** [3]: A robust general-purpose image manipulation detector.
2. **Portrait-specific Detector**: Focused explicitly on detecting face-swapping and portrait manipulations.
3. **Text-specific Detector**: Designed to detect alterations within textual fields.

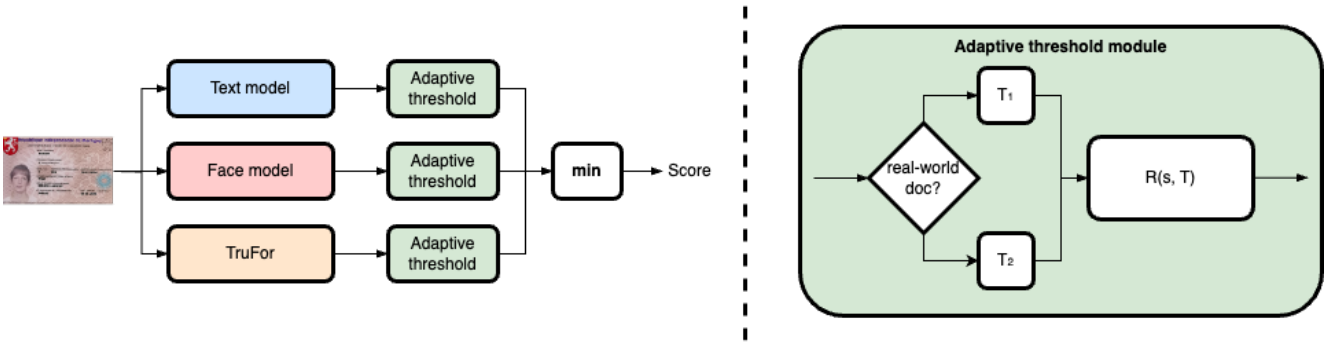
Each model outputs a confidence score $s \in [0, 1]$, where 0 indicates a forged document and 1 indicates an authentic one. To unify these scores, we apply a threshold-based piecewise-linear rescaling function:

$$R(s, T) = \begin{cases} \dfrac{s}{2T}, & 0 \le s < T, \\ \dfrac{1}{2} + \dfrac{s - T}{2(1 - T)}, & T \le s \le 1. \end{cases}$$

This function maps scores below a given threshold T to $[0, 0.5)$ and scores above T to $[0.5, 1]$, effectively calibrating each detector's sensitivity. The final decision score is the minimum of the three rescaled scores:

$$\text{Score} = \min\big(R_{\text{TruFor}},\, R_{\text{Portrait}},\, R_{\text{Text}}\big).$$

Domain-Adaptive Thresholds. Initial experiments demonstrated distinct performance characteristics between synthetic (fantasy) and real-world IDs. Thus, we employed



(a) Model architecture

(b) Adaptive thresholds

Figure 1. Overview of our method. (a) An ensemble of three detectors addresses different forgery types. (b) Domain-adaptive thresholds are applied to scores from each detector to optimize performance on Fantasy and Private datasets.

a lightweight binary classifier to categorize IDs into fantasy or real-world, setting adaptive thresholds accordingly:

$$\begin{aligned} \text{Fantasy:} \quad & T_{\text{TruFor}} = 0.45,\ T_{\text{Portrait}} = 0.99,\ T_{\text{Text}} = 0.90, \\ \text{Real-world:} \quad & T_{\text{TruFor}} = 0.40,\ T_{\text{Portrait}} = 0.90,\ T_{\text{Text}} = 0.70. \end{aligned}$$

2.2. Portrait Forgery Detector

This detector follows a two-stage pipeline. First, a Faster R-CNN [8] model localizes and crops portrait regions from the document. Then, each cropped portrait is passed to a DINOv2 [4] encoder fine-tuned with contrastive learning to distinguish between authentic and manipulated faces. The final score s_{portrait} is the softmax probability for the "authentic" class. Training used a proprietary dataset of 100k samples and public Fantasy data from the challenge.

2.3. Text Forgery Detector

For text forgery detection, we leverage Florence-2 Large [14], formulating the task as a single-token QA problem. The model predicts whether the image includes altered text by outputting tokens `"Yes"` or `"No"`. The probability of the first token being `"No"` is computed by applying the softmax function to the output logits and set as s_{text}.

Training leveraged 10k annotated images combined with synthetic forgeries generated by various text manipulation methods (SRNet [13], STEFANN [9], MOSTEL [7], RS-STE [2], DBEST [10]) and the public Fantasy training data from the challenge.

3. Experiments

Evaluation is conducted on two distinct test sets: **Fantasy** (synthetic documents) and **Private** (real-world IDs), each containing manipulated and genuine samples. We report the F1 score on each set, as well as the aggregate F1.

3.1. Results

Table 1 shows the final performance. Our method achieved second place overall with consistent performance on both test sets—ranking second on the Private set and third on the Fantasy set. Unlike other competitors who performed well on either the Fantasy or the Private sets but not both, our approach maintains strong results across domains, demonstrating effective generalization to both synthetic and real-world document forgeries.

3.2. Ablation

We conduct a series of ablation experiments to evaluate the contribution of each component in our system. Table 2 summarizes the F1 scores for different combinations of modules on the Fantasy and Private test sets.

Combined models with domain-adaptive thresholds. Our best-performing configuration includes all three detectors (TruFor, Portrait, and Text) and uses domain-specific thresholds for fantasy and real-world documents. This setup achieves the highest aggregate F1 of 0.7878, showing strong performance across both domains. Notably, the Fantasy F1 score improves by nearly 5 pp compared to a shared-threshold version, confirming the importance of domain-aware calibration.

Effect of removing domain adaptation. Replacing domain-adaptive thresholds with shared values results in a noticeable F1 drop on the Fantasy set (from 0.8679 to 0.8182), with marginal gain on the Private set. This indicates that threshold tuning plays a key role in handling domain shift between synthetic and real IDs.

Effect of individual detectors. To assess the contribution of each detector, we disable one component at a time:

Version	F1 on Fantasy ↑	F1 on Private ↑	Aggregate F1 ↑
Domain-adaptive thresholds	**0.8679**	0.7534	**0.7878**
TruFor + Portrait + Text (simple thresholds)	0.8182	**0.7555**	0.7743
TruFor + Portrait	0.8072	0.7299	0.7531
TruFor + Text	0.8543	0.7050	0.7498
Portrait + Text	0.6732	0.7281	0.7116
Baseline (TruFor)	0.8067	0.6622	0.7055
MMFusion + Portrait	0.7463	0.7309	0.7355
All fraud	0.6882	0.6553	0.6651
All genuine	0.0771	0.0934	0.0885
Fine-tuned VLM	0.4680	0.6554	0.5992

Table 2. Ablation study results showing the contribution of each model component and setup.

TruFor + Portrait: Removing the Text detector weakens performance on the Fantasy set but maintains good results on the Private set (F1 0.7299). This highlights the Text model's importance for synthetic data.

TruFor + Text: This pairing excels on the Fantasy set (F1 0.8543) but underperforms on the Private set, where face manipulations are more prevalent. This suggests TruFor alone is insufficient for detecting sophisticated facial forgeries in real-world documents.

Portrait + Text: Removing TruFor causes a drastic drop in the Fantasy F1 score while maintaining performance on the Private set. This demonstrates TruFor's critical role in detecting general artifacts common in synthetic images.

TruFor alone. Using only TruFor yields a decent Fantasy F1 (0.8067) but a significantly lower Private F1 (0.6622). This disparity suggests TruFor is effective at spotting low-level manipulation artifacts in synthetic images but struggles with face-specific tampering in real IDs.

Memory-efficient model — MMFusion. We tested MMFusion [12] + Portrait model as a memory-efficient alternative to TruFor. This setup limits image resolution to 2000px, reducing GPU memory usage, but underperforms on the Fantasy set (F1 0.7463), likely due to the loss of fine-grained artefacts from downscaling.

Naïve baselines. The "All fraud" baseline (score 0.0 for every image) surprisingly achieves a moderate Private F1 of 0.6553. In contrast, the "All genuine" baseline performs poorly. This confirms a heavy class imbalance in the test sets, with an estimated fraud-to-genuine ratio of roughly 3:1.

Fine-tuning Florence-2. Lastly, we experimented with fine-tuning the Florence-2 VLM on the challenge's Public train set. Despite high training accuracy, the fine-tuned model performs poorly at test time (Aggregate F1 of 0.5992), suggesting strong overfitting and insufficient generalization. This highlights the importance of diverse, external data sources when training foundation models for visual forgery detection.

3.3. Train-test analysis

To understand how well performance on the Public train set predicts performance on the private test sets, we analyzed the correlation between F1 scores across these datasets. We used a subset of our models that were not trained on the challenge's training data to ensure a fair comparison.

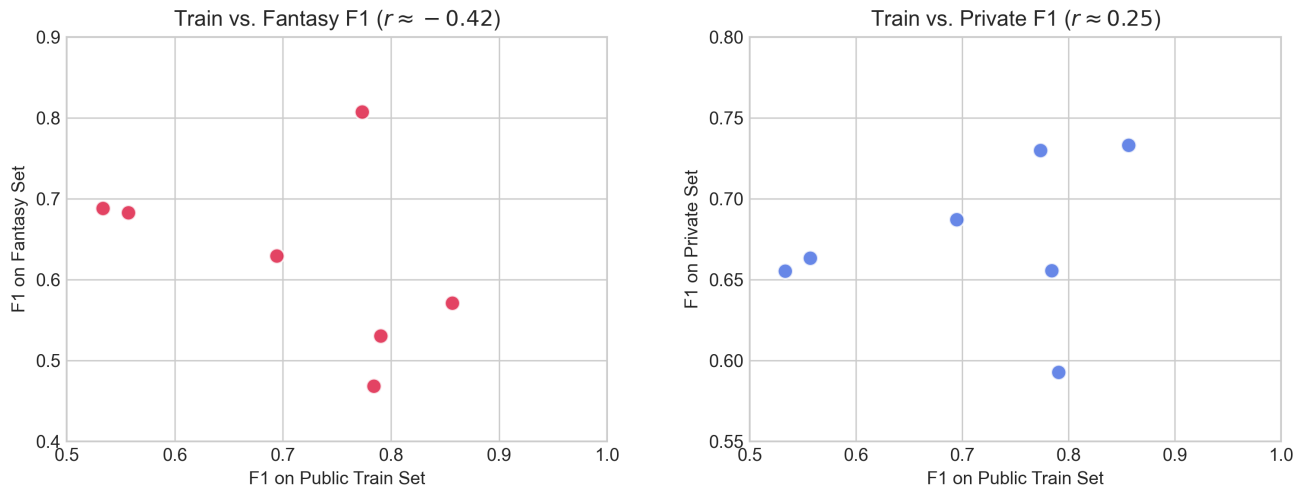
Figure 2 shows scatter plots of the F1 scores. We observe a weak positive correlation (Pearson's $r \approx 0.25$) between performance on the Public train and the Private sets, suggesting the former is not a strong predictor for the latter.

Interestingly, despite both the Public train and Fantasy test sets containing synthetic IDs, we found a moderate negative correlation (Pearson's $r \approx -0.42$) between their F1 scores. This unexpected inverse relationship suggests that the forgery techniques differ significantly between the two synthetic datasets. Models optimized for the Public train set's specific artifacts may perform worse on the distinct patterns present in the Fantasy test set.

4. Discussion

Our experiments highlight a key discrepancy between the challenge benchmarks and real-world scenarios. While the challenge datasets contain approximately 75% fraudulent documents, operational fraud detection systems typically encounter fraud rates below 5%. This class imbalance significantly impacts performance metrics.

To quantify this, we evaluated TruFor on the Fantasy training set under two conditions: the original distribution (75% fraud) and a resampled one mirroring a real-world scenario (5% fraud). The F1 score dropped from 0.827 to 0.142 on the realistic distribution. This suggests that fu-



(a) Train F1 vs. Fantasy F1 ($r \approx -0.42$)

(b) Train F1 vs. Private F1 ($r \approx 0.25$)

Figure 2. Correlation between F1 scores on the public train set and the two private test sets (Fantasy and Private). Each point represents a different model configuration.

ture challenges could better align with diverse deployment scenarios by incorporating evaluations with varying class distributions.

Additionally, the Public train dataset contains images exceeding 4000 px resolution, requiring substantial computational resources. Our analysis suggests that such high resolutions offer diminishing returns for fraud detection. In contrast, real-world applications typically operate on more moderate-resolution images due to practical constraints. Future iterations might consider generating synthetic images at more practical resolutions to reduce computational overhead without significantly compromising detection efficacy.

5. Conclusion

In this paper, we presented Incode's second-place solution for the ICCV 2025 DeepID challenge. Our approach integrates three specialized detectors—TruFor, a portrait forgery model, and a text manipulation model—with a domain-adaptive thresholding strategy. This ensemble delivered robust and balanced performance across both synthetic (Fantasy) and real-world (Private) test sets, a key differentiator from other top methods.

Our ablation studies confirmed that each component was crucial and that domain-adaptive thresholds were vital for bridging the gap between datasets. The analysis also revealed important insights into train-test correlations and the impact of class distribution, offering valuable considerations for designing future benchmarks that better reflect diverse operational scenarios.

References

[1] The challenge of detecting synthetic manipulations in id documents (deepid 2025). https://deepid-iccv.github.io/, 2025. Accessed: 2025-08-22. 1

[2] Zhengyao Fang, Pengyuan Lyu, Jingjing Wu, Chengquan Zhang, Jun Yu, Guangming Lu, and Wenjie Pei. Recognition-synergistic scene text editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13104–13113, 2025. 2

[3] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20606–20615, 2023. 1

[4] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[5] Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards robust tampered text detection in document image: New dataset and new solution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5937–5946, 2023. 1

[6] Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Generalized tampered scene text detection in the era of generative ai. *arXiv e-prints*, pages arXiv–2407, 2024. 1

[7] Yadong Qu, Qingfeng Tan, Hongtao Xie, Jianjun Xu, Yuxin Wang, and Yongdong Zhang. Exploring stroke-level modifications for scene text editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2119–2127, 2023. 2

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[9] Prasun Roy, Saumik Bhattacharya, Subhankar Ghosh, and Umapada Pal. Stefann: scene text editor using font adaptive neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13228–13237, 2020. 2

[10] Joshua Santoso, Christian Simon, et al. On manipulating scene text in the wild with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5202–5211, 2024. 2

[11] Juan E Tapia, Naser Damer, Christoph Busch, Juan M Espin, Javier Barrachina, Alvaro S Rocamora, Krištof Ocvirk, Leon Alessio, Borut Batagelj, Sushrut Patwardhan, et al. First competition on presentation attack detection on id card. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2024. 1

[12] Kostas Triaridis, Konstantinos Tsigos, and Vasileios Mezaris. Mmfusion: Combining image forensic filters for visual manipulation detection and localization. *arXiv preprint arXiv:2312.01790*, 2023. 3

[13] Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Editing text in the wild. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1500–1508, 2019. 2

[14] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4818–4829, 2024. 2