

Platform Drift in Social-Media Image Forgery Detection and Localization: An Application-Level Analysis

Yahya AL-Nabhani

17221781@siswa.um.edu.my

University of Malaya

Amirrudin Kamsin

University of Malaya

Mohamad Nizam

University of Malaya

Khalil Al Ruqaishi

University of Nizwa

Research Article

Keywords: image forensics, forgery localization, online social networks, platform drift, robustness, deep learning

Posted Date: August 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9513365/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Platform Drift in Social-Media Image Forgery Detection and Localization: An Application-Level Analysis

Yahya Al-Nabhani^{1,2}, Amirrudin Kamsin¹, Mohamad Nizam¹, and Khalil Al Ruqaishi³

¹Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia

²Information Technology Department, Ministry of Education, Nizwa 611, Oman

³Computer Sciences Section, University of Nizwa, Al-Dakhiliyah 616, Oman

Corresponding author: Yahya Al-Nabhani (17221781@siswa.um.edu.my)

Abstract

Online social networks apply platform-specific upload–download pipelines such as re-compression, resizing, chroma modification, and metadata rewriting that can distort forensic traces and reduce the reliability of image manipulation detection and localization. This paper presents a platform-drift robustness evaluation for a forensic system derived from the published REFORGE framework and a standardized visualization panel consisting of input, heatmap, mask, and overlay outputs. All metrics are computed on 20 matched samples per dataset at a fixed operating point in order to isolate platform-induced drift. Using four social platforms (Facebook, WeChat, Weibo, and WhatsApp) and four benchmark datasets (CASIA, Columbia, DSO, and NIST16), we quantify decision stability and three drift metrics: tamper-probability drift (Δp_{fake}), mask-area drift (ΔArea , in percentage points), and connected-component drift ($\Delta \text{Regions}$). Across all evaluated dataset–platform settings, binary decisions remain stable, while confidence and mask geometry exhibit platform- and dataset-dependent drift with rare outliers. We report global trends, dataset-specific sensitivities, and key outliers, and we conclude with deployment recommendations that motivate OSN-consistency regularization and platform-conditioned normalization as practical robustness enhancements.

Keywords: image forensics; forgery localization; online social networks; platform drift; robustness; deep learning

1 Introduction

Online social networks (OSNs) have become a primary channel for image acquisition and dissemination, but OSN upload–download pipelines introduce platform-specific post-processing operations, including JPEG recompression, resizing, chroma subsampling, and metadata rewriting. These transformations can attenuate subtle forensic traces and cause domain shift between training data and deployment conditions. Consequently, OSN robustness is increasingly treated as a first-class requirement for image forgery detection and localization [1].

Our prior work introduced REFORGE, a dual-branch ensemble that combines strong image-level detection with pixel-level localization and reinforcement-learning (RL) refinement to improve mask quality on challenging manipulations [2]. In that study, REFORGE achieved competitive localization performance on standard benchmarks and retained strong performance on OSN-transmitted variants. However, end-task metrics such as IoU and F1 do not fully reveal

how OSN transforms perturb outputs: whether decisions flip, whether confidence drifts, whether masks inflate or deflate, and whether predicted regions fragment or merge.

This paper addresses that gap by presenting a platform-drift analysis built on the same forensic system and the Image Authenticity & Forgery Localization application used in our experiments. We analyze matched sets of the same underlying image across five conditions: the original image and four OSN platforms (Facebook, WeChat, Weibo, and WhatsApp). For each dataset and platform, we quantify decision stability and three drift metrics: probability drift (Δp_{fake}), mask-area drift (ΔArea), and connected-component drift ($\Delta \text{Regions}$). We then identify global trends, dataset-specific sensitivities, and key outliers, and conclude with actionable recommendations for OSN deployment.

Contributions. This paper makes four contributions. First, it defines a drift-oriented OSN robustness protocol for image forgery localization that complements IoU and F1 by reporting Δp_{fake} , ΔArea , $\Delta \text{Regions}$, and decision flip rate. Second, it presents a comprehensive platform-drift study across four datasets and four OSN platforms, reporting means and maxima for each metric. Third, it performs outlier analysis that pinpoints the most drift-prone samples and provides qualitative aligned panels for forensic interpretation. Fourth, it proposes OSN-consistency regularization and platform-conditioned normalization as practical mitigation strategies for deployment.

2 Related Work

Deep learning has become the dominant paradigm for image manipulation detection and localization. Earlier CNN-based approaches such as MVSS-Net emphasize multi-view and multi-scale supervision to capture boundary artifacts and noise-view cues [3]. Recent transformer-centric methods improve long-range modeling and adaptivity, including TruFor, IML-ViT, AdaIFL, and related contrastive or benchmark-driven frameworks [4, 5, 6, 7]. Reinforcement-learning formulations have also been explored for forgery localization under sparse supervision [8].

Robustness under post-processing and real-world degradations remains a central challenge. Prior work studied robustness to common corruptions and OSN transmission, often reporting end-task performance drops [1, 9]. In contrast, our focus is not only whether performance changes under OSN transmission, but how outputs drift in terms of confidence, mask size, and region topology—quantities that directly affect practical forensic workflows and human interpretation. From a deployment perspective, domain generalization and normalization-based adaptation offer practical mechanisms to mitigate test-time shifts by controlling feature statistics, which supports our recommendations on OSN-consistency regularization and platform-conditioned normalization [10].

3 Method Overview: REFORGE and the Forensic Application

This paper evaluates a forensic pipeline based on REFORGE [2] and a downstream application that visualizes model outputs for human-in-the-loop analysis. REFORGE combines an image-level classifier to estimate the probability of manipulation, a segmentation model to predict pixel-level tamper masks, and an RL-based post-processing stage that refines the predicted mask to reduce false positives, enforce spatial coherence, and better align boundaries. The application produces a standardized panel consisting of the input image, a heatmap of predicted tamper likelihood, a binarized mask, and an overlay for qualitative inspection.

Figure 1 summarizes the end-to-end evaluation workflow used in this study.

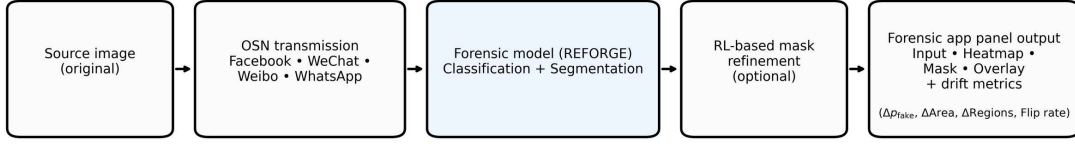


Figure 1: End-to-end OSN robustness evaluation workflow. The same source image is transmitted through OSN platforms, then processed by the REFORGE-based forensic pipeline to produce a standardized visualization panel and drift metrics.

4 OSN Robustness Evaluation Protocol

4.1 Datasets and OSN Platforms

We evaluate four benchmark datasets commonly used in manipulation localization: CASIA, Columbia, DSO, and NIST16. For each dataset, we analyze a matched set of $N = 20$ samples selected from the forensic application outputs. Each sample has five versions: the original source image and four OSN-transmitted variants (Facebook, WeChat, Weibo, and WhatsApp).

4.2 Drift Metric Definitions

Let p_0 and p_k be the predicted tamper probabilities for the original image and OSN platform k , respectively. Let M_0 and M_k be the binarized masks for the original and platform k , and let $\text{Area}(M)$ denote the fraction of pixels labeled as manipulated. Let $\text{Regions}(M)$ be the number of connected components (8-connectivity) in M . We define

$$\Delta p_{\text{fake}}(k) = p_k - p_0, \quad (1)$$

$$\Delta \text{Area}(k) = 100 (\text{Area}(M_k) - \text{Area}(M_0)) \text{ [pp]}, \quad (2)$$

$$\Delta \text{Regions}(k) = \text{Regions}(M_k) - \text{Regions}(M_0). \quad (3)$$

For each dataset–platform pair, we report mean absolute drift for p_{fake} and area, signed mean drift for regions, maxima over samples, and binary decision flip rate.

Table 1: Platform-wise drift statistics computed on 20 matched samples per dataset. Δp_{fake} denotes the change in predicted tamper probability. ΔArea is measured in percentage points (pp) of predicted mask area, and $\Delta \text{Regions}$ is the change in the number of connected components in the binary mask. Flip rate is the fraction of samples whose binary decision changes after OSN transmission.

Dataset	Platform	Flip (%)	mean $ \Delta p_{\text{fake}} $	max $ \Delta p_{\text{fake}} $	mean $ \Delta \text{Area} $ (pp)	max $ \Delta \text{Area} $ (pp)	mean $\Delta \text{Regions}$	max $ \Delta \text{Regions} $
CASIA	Facebook	0	0.000217	0.003013	0.969	11.270	0.20	4
	WeChat	0	0.003813	0.059452	2.498	15.114	0.35	5
	Weibo	0	0.003484	0.069483	0.166	0.750	0.10	1
	WhatsApp	0	0.001164	0.011722	1.280	7.876	-0.25	3
Columbia	Facebook	0	0.000004	0.000040	0.843	3.112	-0.25	5
	WeChat	0	0.000003	0.000016	0.116	0.506	0.60	7
	Weibo	0	0.000004	0.000063	0.318	1.457	0.60	5
	WhatsApp	0	0.000003	0.000042	0.643	2.264	-0.35	4
DSO	Facebook	0	0.000001	0.000003	0.143	0.487	-0.05	8
	WeChat	0	0.000004	0.000018	0.394	1.103	-2.20	9
	Weibo	0	0.000001	0.000003	0.525	2.210	-2.40	11
	WhatsApp	0	0.000003	0.000015	0.407	1.030	-5.05	12
NIST16	Facebook	0	0.000000	0.000002	0.043	0.230	-0.50	3
	WeChat	0	0.000000	0.000001	0.125	1.851	-0.60	4
	Weibo	0	0.000001	0.000008	0.221	2.338	-0.85	5
	WhatsApp	0	0.000000	0.000002	0.037	0.271	-0.70	5

Table 2: Dataset-level drift summary aggregated across platforms (means over the four OSN variants).

Dataset	mean $ \Delta p_{\text{fake}} $	max $ \Delta p_{\text{fake}} $	mean $ \Delta \text{Area} $ (pp)	max $ \Delta \text{Area} $ (pp)	mean $\Delta \text{Regions}$
CASIA	0.002170	0.069483	1.228	15.114	0.10
Columbia	0.000004	0.000063	0.480	3.112	0.15
DSO	0.000002	0.000018	0.367	2.210	-2.42
NIST16	0.000001	0.000008	0.106	2.338	-0.66

5 Results and Analysis

5.1 Global Trends

Across all 16 dataset–platform combinations, the decision flip rate is 0%, indicating that OSN transmission does not change the final binary decision for the evaluated samples under the chosen operating point. Nevertheless, the soft outputs and the mask geometry do drift.

Averaged over all dataset–platform settings, the mean absolute probability drift is 0.000544, while the largest observed $\max|\Delta p_{\text{fake}}|$ reaches 0.069483. For mask geometry, the overall mean absolute area drift is 0.546 pp, with the largest observed $\max|\Delta \text{Area}|$ reaching 15.114 pp. Region topology exhibits a negative signed mean drift (-0.71), suggesting that OSN processing tends to merge or suppress small fragmented components in the predicted mask, while extreme cases can still create or remove many regions ($\max|\Delta \text{Regions}| = 12$).

Figure 2 highlights platform-specific patterns observed in Table 1. In particular, WeChat yields the largest mean $|\Delta \text{Area}|$ for CASIA (2.498 pp) and the largest $\max|\Delta \text{Area}|$ overall (15.114 pp). For confidence drift on CASIA, Weibo produces the largest $\max|\Delta p_{\text{fake}}|$ (0.069483), while WeChat yields the largest mean $|\Delta p_{\text{fake}}|$ (0.003813). Across Columbia, DSO, and NIST16, the near-zero mean $|\Delta p_{\text{fake}}|$ indicates decision-confidence stability, whereas the predominantly negative $\Delta \text{Regions}$ suggests that OSN processing tends to merge or suppress small components rather than fragment them.

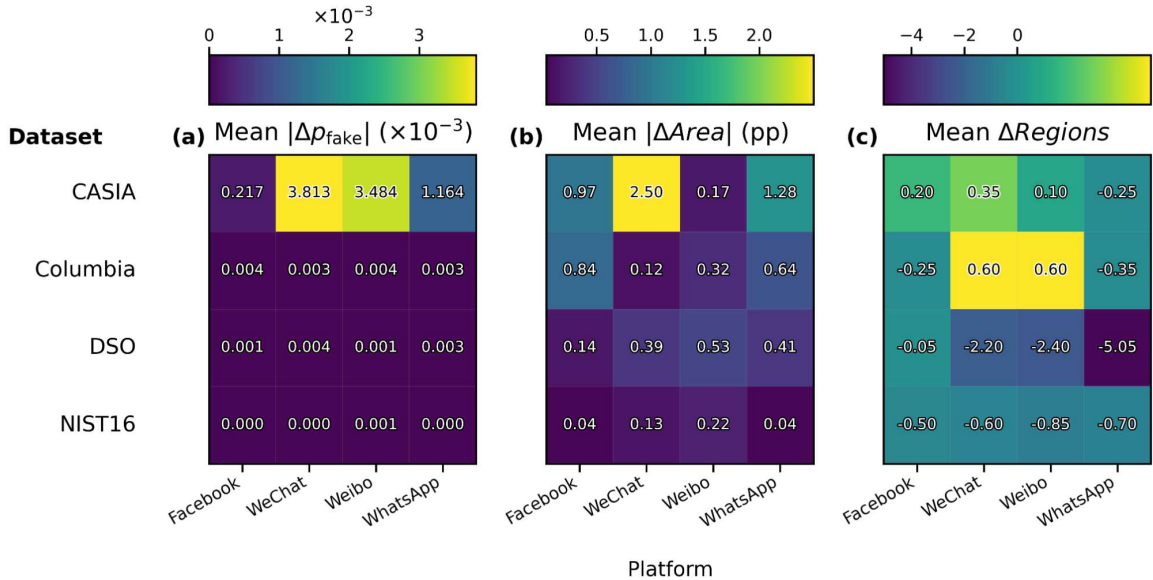


Figure 2: Mean drift heatmaps by dataset (rows) and OSN platform (columns). (a) Mean absolute confidence drift $|\Delta p_{\text{fake}}|$ (scaled by 10^{-3}). (b) Mean absolute mask-area drift $|\Delta \text{Area}|$ in percentage points (pp). (c) Mean change in the number of connected regions $\Delta \text{Regions}$ (positive: fragmentation; negative: merging). Values correspond to Table 1.

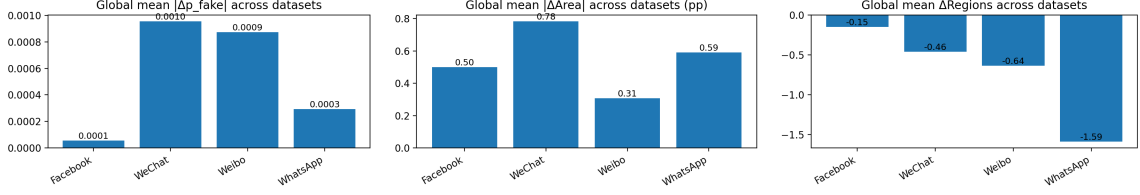


Figure 3: Global platform-averaged drift across datasets: (a) mean $|\Delta p_{\text{fake}}|$, (b) mean $|\Delta \text{Area}|$ in percentage points (pp), and (c) mean $\Delta \text{Regions}$ (negative indicates region merging/suppression).

5.2 Dataset-Specific Findings

CASIA. CASIA exhibits the largest drift among the evaluated datasets. The platform-averaged mean $|\Delta p_{\text{fake}}|$ is 0.002170, with the maximum reaching 0.069483. Mask size is more sensitive: mean $|\Delta \text{Area}|$ is 1.228 pp and reaches a worst-case max $|\Delta \text{Area}|$ of 15.114 pp. Region drift is near zero on average (0.10), implying that the dominant effect is mask inflation or deflation rather than systematic fragmentation; nevertheless, the largest observed $|\Delta \text{Regions}|$ is 5.

Columbia. Columbia shows highly stable probabilities with mean $|\Delta p_{\text{fake}}|$ of 0.000004 and max $|\Delta p_{\text{fake}}|$ of 0.000063. Area drift remains moderate (mean 0.480 pp; max 3.112 pp), while topology can vary (max $|\Delta \text{Regions}| = 7$), suggesting that fine-grained boundary details are more susceptible than global confidence.

DSO. DSO displays small probability and area drift but a pronounced negative mean region drift (-2.42). This indicates that OSN processing tends to suppress small connected components, effectively merging fragmented regions or eliminating tiny false positives. Despite this average stabilization, DSO also contains the largest topology outlier (max $|\Delta \text{Regions}| = 12$).

NIST16. NIST16 is the most stable dataset under the evaluated OSN pipelines, with mean $|\Delta p_{\text{fake}}|$ of 0.000001 and mean $|\Delta \text{Area}|$ of 0.106 pp. The maximum observed area drift (2.338 pp) occurs in a small number of samples and motivates explicit outlier monitoring in deployment.

5.3 Panel-derived Mask Stability Under OSN Transmission

In addition to drift metrics, we quantify how spatially consistent the predicted tamper masks remain after OSN transmission. We extract the mask tile from the standardized forensic panel output, binarize it, and compare the original-image mask M_0 with each platform mask M_k using Dice and IoU overlap. To capture boundary sensitivity, we additionally compute boundary IoU (bIoU) on 1-pixel morphological boundaries.

Table 3: Panel-derived mask stability between the original mask M_0 and OSN masks M_k (Dice, IoU, and boundary IoU). Values are aggregated over all platforms and samples per dataset; P05 denotes the 5th percentile (worst 5%).

Dataset	mean Dice	P05 Dice	mean IoU	P05 IoU	mean bIoU	P05 bIoU
CASIA	0.881	0.396	0.828	0.247	0.486	0.012
Columbia	0.942	0.687	0.920	0.523	0.579	0.185
DSO	0.938	0.819	0.887	0.693	0.514	0.285
NIST16	0.978	0.906	0.960	0.828	0.733	0.108

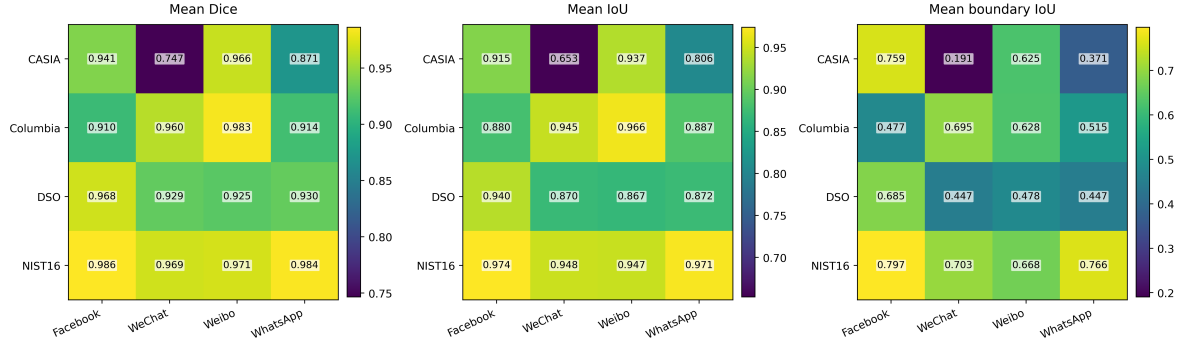


Figure 4: Mean panel-derived overlap stability between the original predicted mask and OSN variants: Dice, IoU, and boundary IoU (bIoU). Each cell is averaged over 20 matched samples per dataset.

Table 4: Top absolute drift outliers (one per drift metric and dataset) among the 20 matched samples and four OSN platforms. “Max $|\cdot|$ ” reports the maximum absolute drift observed for the metric; “Signed” indicates the direction of the drift for that outlier (negative: decrease/merging; positive: increase/fragmentation).

Dataset	Metric	Platform	Sample ID	Max $ \cdot $	Signed
CASIA	Δp_{fake}	Weibo	Sp_D_CNN_A_nat0014_nat0017_0097	0.069483	+0.069483
	$\Delta \text{Area (pp)}$	WeChat	Sp_D_CNN_R_art0006_art0087_0276	15.114	-15.114
	$\Delta \text{Regions}$	WeChat	Sp_D_CNN_R_art0006_art0087_0276	5	+5
Columbia	Δp_{fake}	Weibo	canong3_canonxt_sub_20	0.000063	+0.000063
	$\Delta \text{Area (pp)}$	Facebook	canong3_canonxt_sub_08	3.112	-3.112
	$\Delta \text{Regions}$	WeChat	canong3_canonxt_sub_08	7	+7
DSO	Δp_{fake}	WeChat	splicing-17	0.000018	-0.000018
	$\Delta \text{Area (pp)}$	Weibo	splicing-04	2.210	+2.210
	$\Delta \text{Regions}$	WhatsApp	splicing-02	12	-12
NIST16	Δp_{fake}	Weibo	16	0.000008	+0.000008
	$\Delta \text{Area (pp)}$	Weibo	16	2.338	-2.338
	$\Delta \text{Regions}$	WhatsApp	16	5	-5

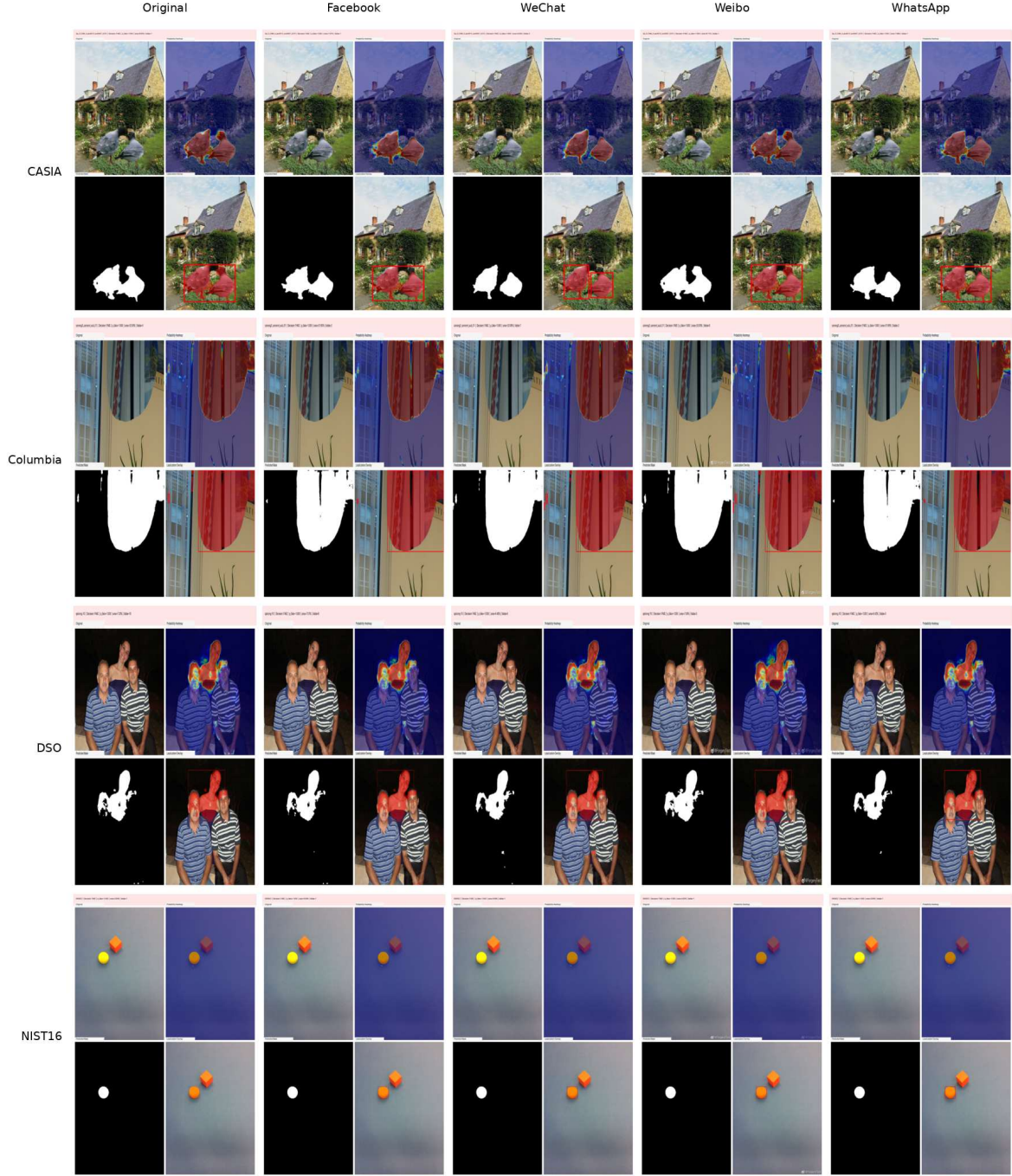


Figure 5: Representative aligned forensic panels (one sample per dataset) across the original and four OSN platforms. Each cell shows the standardized forensic panel (Input/Heatmap/Mask/Overlay) produced by the application. The full 80-page aligned atlas is provided as Supplementary Material S1.

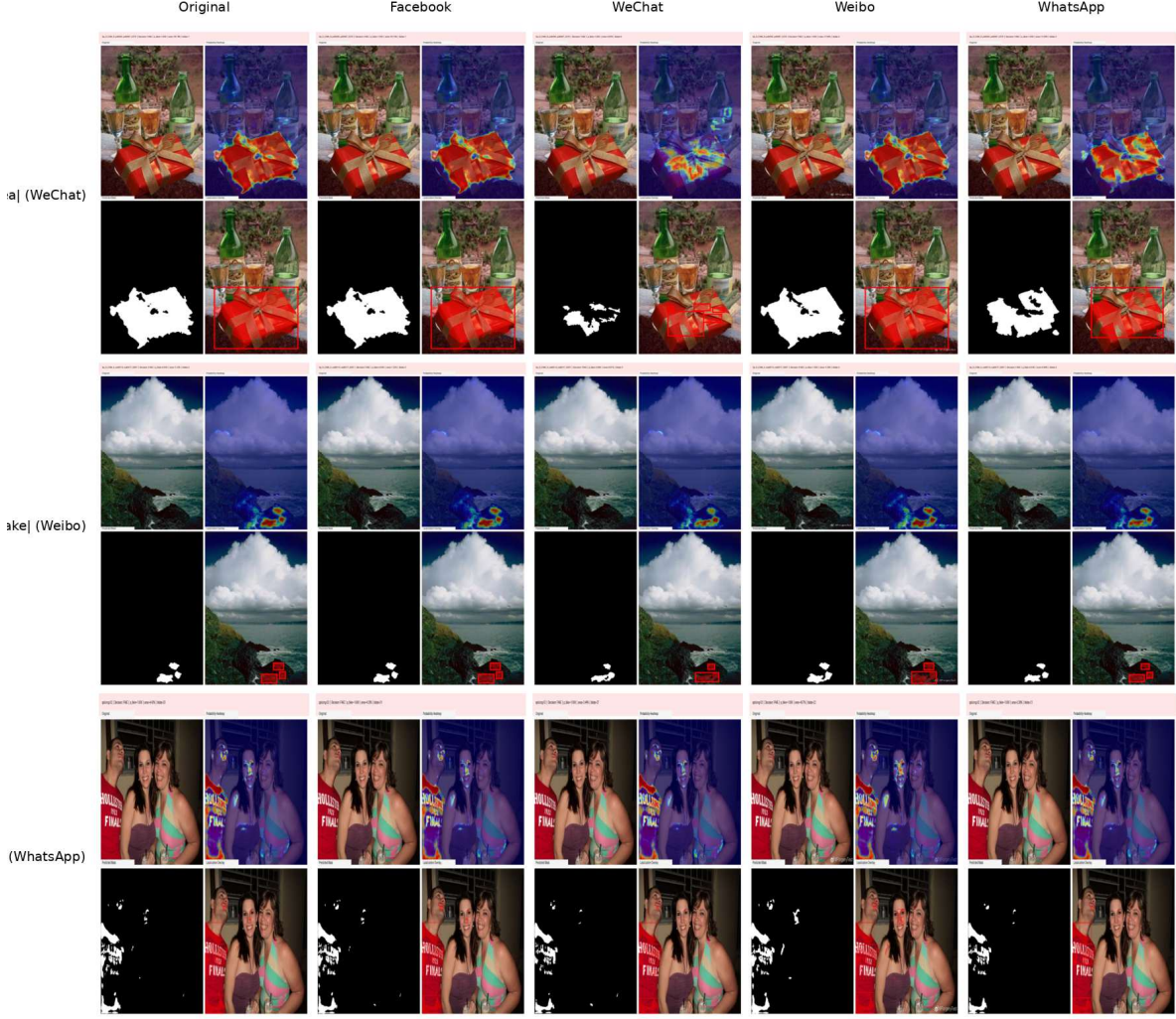


Figure 6: Stress-test outlier panels for the largest observed drift cases. These examples correspond to the strongest confidence, area, and topology outliers identified in Table 4.

Table 5: Representative state-of-the-art comparison on pixel-level localization (IoU/F1). Values for competitors follow the consolidated evaluation protocol reported in AdaIFL and are also summarized in the original REFORGE paper.

Method	DSO-1 (IoU/F1)	NIST16 (IoU/F1)
TruFor [4]	0.270 / 0.634	—
AdaIFL [7]	—	0.393 / 0.706
REFORGE [2]	0.636 / 0.757	0.676 / 0.757

Beyond drift metrics, REFORGE remains competitive on conventional localization benchmarks. As reported in REFORGE [2], the framework attains IoU/F1 of 0.636/0.757 on DSO-1 and 0.676/0.757 on NIST16, outperforming representative transformer baselines such as TruFor and improving over AdaIFL on NIST16 under the consolidated protocol. Importantly, OSN robustness evaluated in REFORGE reports a mean IoU/F1 of 0.675/0.757 across OSN-transmitted variants, suggesting that the main degradation manifests as drift rather than decision flips. Our drift analysis complements these findings by identifying which platforms and datasets contribute most to mask and topology changes.

6 Robustness Discussion and Methodological Recommendations

The results indicate a favorable deployment property: decisions are stable under OSN transmission for the evaluated samples (0% flip rate), and mean drifts are small on most datasets. However, the presence of platform-dependent area and topology drift—especially under WeChat recompression and for CASIA samples—has practical implications for forensic workflows. Area inflation can overestimate manipulation extent, while region merging can hide small forged fragments.

We recommend two complementary strategies to reduce OSN-induced drift. **(1) OSN-consistency regularization:** during training, treat OSN transmission as a data-augmentation family and enforce consistency in both manipulation probability and mask geometry across original and OSN-like variants. **(2) Platform-conditioned normalization:** when the deployment platform is known or can be inferred, incorporate a lightweight conditioning signal into normalization layers and calibrate probability and mask thresholds per platform. Together, these strategies target both invariance and adaptation and directly address the drift patterns observed in the quantitative and qualitative analysis.

7 Conclusion

We presented an OSN robustness evaluation centered on platform drift rather than only end-task accuracy. Using four datasets and four OSN platforms, we found that decisions remain stable while confidence, mask area, and region topology can drift in a platform- and dataset-dependent manner. CASIA is the most sensitive dataset, and WeChat induces the largest average mask-area drift. We highlighted key outliers and provided aligned qualitative panels to support forensic interpretation. Finally, we proposed OSN-consistency regularization and platform-conditioned normalization as practical strategies to mitigate drift for real-world deployment.

Acknowledgment

The authors gratefully acknowledge Universiti Malaya and the High Performance Computing facilities provided by the Data-Intensive Computing Centre (DICC) for supporting and supplying the computational resources necessary to conduct the experiments and evaluations presented in this work.

Statements and Declarations

Funding. This work was supported by the University of Malaya’s Data-Intensive Computing Centre (DICC).

Competing Interests. The authors declare that they have no competing interests.

Author Contributions. Yahya Al-Nabhani: Conceptualization, methodology, software, investigation, formal analysis, visualization, writing—original draft. Amirrudin Kamsin: Supervision, conceptualization, writing—review and editing. Mohamad Nizam: Supervision, validation, writing—review and editing. Khalil Al Ruqaishi: Methodology, validation, writing—review and editing.

Data Availability. The figures and source files used to prepare the submission package are included with the submission materials. Supplementary atlas and drift-analysis files are available as supplementary material.

References

- [1] H. Wu, J. Zhou, J. Tian, J. Liu, and Y. Qiao, “Robust image forgery detection against transmission over online social networks,” *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 443–456, 2022.
- [2] Y. Al-Nabhani, A. Kamsin, M. Nizam, and K. Al Ruqeishi, “REFORGE: A robust ensemble for image forgery detection and localization in social network images,” *IEEE Access*, vol. 14, pp. 8773–8788, 2026.
- [3] X. Chen, C. Dong, J. Ji, J. Cao, and X. Li, “Image manipulation detection by multi-view multi-scale supervision,” in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14185–14193.
- [4] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, “TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 20606–20615.
- [5] J. Zhou, A. Y. Alhammadi, Z. Jiang, B. Du, X. Ma, and X. Du, “IML-ViT: Benchmarking image manipulation localization by vision transformer,” *arXiv:2307.14863*, 2023.
- [6] J. Zhou, X. Ma, X. Du, A. Y. A. Hammadi, and W. Feng, “Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning,” in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22346–22356.
- [7] Y. Li, F. Cheng, W. Yu, G. Wang, G. Luo, and Y. Zhu, “AdaIFL: Adaptive image forgery localization via a dynamic and importance-aware transformer network,” in *Computer Vision – ECCV 2024*, 2024, pp. 477–493.
- [8] C. Yan, X. Li, X. Zhang, and C. Yang, “Image forgery localization with sparse reward compensation using curiosity-driven deep reinforcement learning,” *Journal of Visual Communication and Image Representation*, vol. 112, p. 104587, 2025.
- [9] H. Wu, J. Zhou, J. Tian, and J. Liu, “Robust image forgery detection over online social network shared images,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13440–13449.
- [10] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, “Domain generalization: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396–4415, 2023.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [S1AtlasIndex.zip](#)