Leveraging Semantic Segmentation for Image Manipulation Detection and Localization

Yuwei Chen Ming-Ching Chang Xin Li
Department of Computer Science, University at Albany, State University of New York, NY, USA
{ychen69, mchang2, xli48}@albany.edu

***Abstract*—Detecting and localizing image manipulation has long been a focus in computer vision. Current state-of-the-art methods primarily identify visual and JPEG compression artifacts. In this study, we propose the integration of semantic segmentation to enhance object awareness and improve the accuracy of spliced object localization. Instead of treating semantic segmentation as an independent submodule, we integrate it as a third branch of an end-to-end model. This approach balances visual, compression, and segmentation artifacts, reducing overemphasis on any single branch. Extensive evaluations on established datasets show a three percent average IoU increase in performance across five mainstream datasets for image manipulation localization. We also provide insights into how existing digital defense models adapt to new image tampering techniques like generative fill and expand.**

***Index Terms*—image forgery detection, image forgery localization, digital forensics, semantic segmentation,**

I. Introduction

Image manipulation, altering visual content to deceive or mislead viewers, has become increasingly prevalent in today's digital landscape. With the widespread availability of sophisticated editing tools and the ease of sharing images across various platforms, the potential for image manipulation to be used for malicious purposes has escalated. Consequently, the development of robust techniques for detecting manipulated images has emerged as a critical area of research across multiple disciplines, including computer vision, image processing, and forensic analysis.

The detection of image manipulation holds significant implications across various domains. In forensic investigations, accurate manipulation detection can aid in verifying the authenticity of digital evidence, thereby facilitating the identification of tampered images and ensuring the integrity of the internet. Additionally, in journalism and media forensics, detecting image manipulations is crucial for preserving the credibility of visual content and combating the spread of misinformation and fake news.

Numerous mainstream image manipulation defense models have been developed to address this emerging threat, each targeting different types of artifacts associated with manipulated images. Approaches encompass Visual-based [1], statistical [2], [3], deep learning [4], and frequency domain analyses [4], [5]. Visual-based artifacts can include misaligned edges, displaced pixels, and variations in lighting or depth. For instance, models like CAT-net [4] focus on integrating multiple artifact types, such as visual anomalies and compression artifacts, to detect and pinpoint manipulated regions within images.

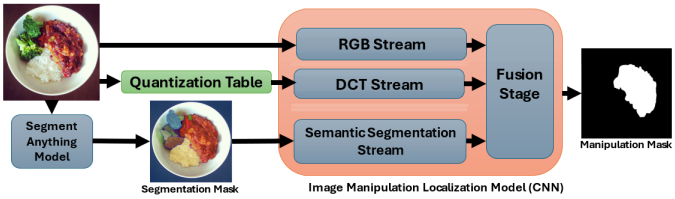

Fig. 1. We integrate *object-aware* semantic segmentation to enhance image manipulation localization. The intuition is based on an observation that image manipulation is frequently done with the whole object entity.

While current mainstream methods have demonstrated effectiveness, they often overlook the integrity of objects when localizing manipulated regions in images. This oversight can result in manipulation regions containing only partial segments of the manipulated object. In this paper, we propose leveraging semantic segmentation to enhance image manipulation localization while maintaining the advantages of capturing visual and compression-type artifacts. Figure 1 overviews our approach to integrating object-aware semantic segmentation with image manipulation detection and localization.

II. Related Works

Image forgery detection and localization have garnered significant attention in recent years due to the proliferation of digital manipulation techniques. Researchers have focused on developing robust methods to identify tampered regions within images, aiming to preserve the integrity and authenticity of digital content. Techniques range from traditional methods like analysis of noise inconsistencies and pixel-level alterations to advanced approaches utilizing deep learning and AI-driven algorithms for more accurate detection and precise localization of forged regions. These efforts are crucial not only for maintaining trust in digital media but also for forensic applications in legal and security domains. Below, we have summarized image manipulation detection methods within the literature.

Pixel-based analysis methods analyze image pixel values and their relationships to identify anomalies indicative of manipulation. Forensic techniques such as Error Level Analysis (ELA) [1] and Noise Analysis [6] leverage pixel-level inconsistencies to detect altered or tampered regions.

Statistical analysis approaches examine statistical properties of images to detect anomalies caused by manipulation. Methods such as Benford's Law [2] and Statistical Moments Analysis [3] exploit statistical irregularities in manipulated images to localize regions with altered content.

Deep learning based approaches have emerged as a powerful tool for image manipulation localization, enabling the development of highly accurate and robust detection models. Convolutional Neural Networks (CNNs) [4] trained on large-scale datasets have shown promising results in localizing manipulations. Generative Adversarial Networks (GANs) have been employed for adversarial image manipulation detection.

Frequency domain analysis techniques operate on image transform domains, such as Fourier or Discrete Cosine Transform (DCT), to detect manipulations. Methods like DCT-based Manipulation Detection [4] and Frequency Analysis of Sensor Pattern Noise (SPN) [5] utilize frequency domain representations to identify manipulations introduced during image editing.

Hybrid methods [7] combine multiple techniques, such as pixel-level analysis, frequency analysis, and deep learning, to achieve enhanced performance in manipulation localization. These methods leverage the complementary strengths of different methodologies to improve detection accuracy and robustness.

III. Method

In our effort to enhance image manipulation localization through semantic segmentation, we chose Cat-Net v2 [4] as our baseline model. Cat-Net was selected for its user-friendly nature and comprehensive codebase, including training and model code. The existing model architecture features two branches designed to capture visual defects and JPEG artifacts. We propose adding a third branch dedicated to semantic segmentation. Solely capturing visual and JPEG artifacts may result in manipulations regions that only cover the spliced object partially. By adding a semantic segmentation branch, we aim to enhance the localization to encompass the manipulated object entirely. We have included this third branch to provide the model with a sense of physical object awareness, thereby enhancing the detection of entity based manipulated.

We adopt much of the lower-level design from HRNet [8] due to its suitability for digital forensics applications. HRNet maintains high-resolution representations throughout the network while preserving lower-level details. This characteristic enables us to capture comprehensive images without losing the intricate details crucial for forensic examinations. HRNet uses stride-2 convolution for downsampling feature maps and avoids pooling layers. Although pooling benefits many computer vision tasks, it is less suitable for tasks requiring subtle signal discernment. Pooling tends to enhance content but suppress noise-like signals, which are significant indicators of tampered regions in image manipulation localization.

Segmentation stream vs. separate module: The decision to incorporate semantic segmentation as a third stream in the model, rather than a separate module, was driven by several key considerations. Primarily, this integration aims to prevent semantic segmentation from dominating the localization process, which could obscure other valuable detection cues. Although semantic segmentation has improved significantly, it has flaws and could cause over-detection if used as a standalone module for completing object regions. By merging it with visual and JPEG artifact detection streams within a unified model framework, the method effectively balances its influence, reducing the risk of over-reliance on any single type of artifact. This architecture, which merges all three streams at the fusion stage, allows for adjustable weighting of each stream's input during training and inference time. This flexibility enhances the model's image manipulation localization performance by fine-tuning semantic segmentation's impact based on specific needs and scenarios rather than relying on fixed rules for segment completion. A detailed illustration of the proposed model architecture can be seen in Figure 2.

A. Model Architecture

Our proposal involves incorporating an additional semantic segmentation stream alongside the existing RGB and DCT streams within CAT-Net [4], as shown in Figure 2. The segmentation stream will focus on delineating the segmentation of pre-existing entities within the image scene. Following the CAT-Net paradigm, we adopt HR-Net for processing the input image across multiple resolutions using up-scaling and down-scaling techniques. Each resolution within this framework is fully connected at every two layers, allowing us to preserve individually captured artifacts at each image resolution while simultaneously sharing features across alternate layers.

Pipeline: The initial stage involves pre-processing the input image, which includes two key procedures: extracting the Discrete Cosine Transform (DCT) table for the DCT stream and utilizing a semantic segmentation model to derive a series of segmentation of the image scene. The segmentation masks of all entities are then consolidated to reconstruct the image in the form of a semantic segmentation labeled image. This image contains semantic class labels for each image pixel. Each label maps that pixel to its corresponding entity segmentation class. We have selected the Meta Segment Anything Model (SAM) [9] to produce the semantic segmentation masks. We have chosen SAM based on its accurate performance in various image scenes.

After obtaining the reconstructed segmentation labeled image and the DCT table of the input image, the original RGB image is sent to the RGB stream. In contrast, DCT table of the Y-channel and corresponding coefficients serve as input for the DCT stream. Simultaneously, the segmentation image is fed to the newly incorporated segmentation stream. Upon completion of processing by all three streams, the resultant feature spaces are fused during the fusion stage.

A final merge operation is conducted across all image resolutions to generate a single heatmap indicating the probable manipulation regions within the image. During fusion, one can fine-tune the threshold at the fusion stage to balance the contribution of the three input streams to more accurately

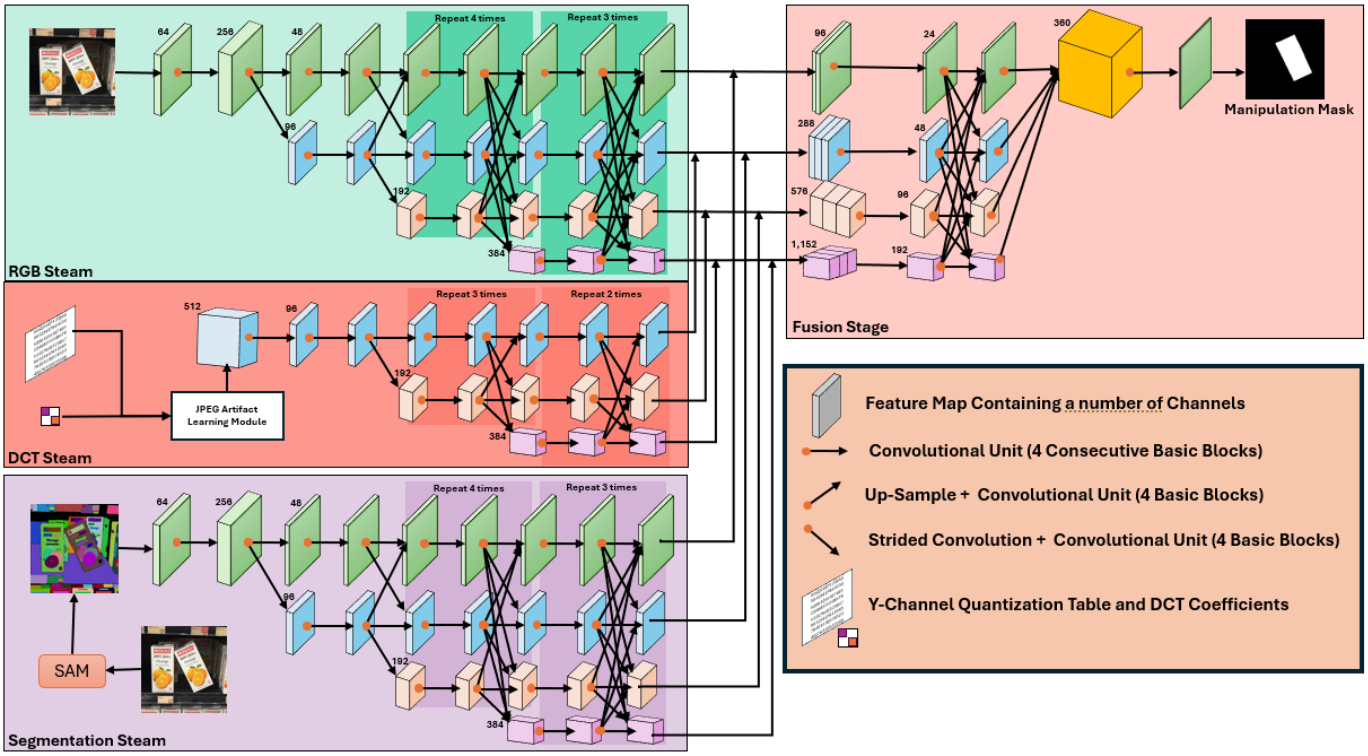

Fig. 2. Detailed illustration of the architecture of the proposed model reveals a structured approach comprising three distinct input streams, each targeting specific categories of forgery artifacts. This design aims to bolster the model's performance and resilience by enabling it to assess multiple types of forgery artifacts prior to generating a manipulation mask. The RGB Stream focuses on capturing visual artifacts like misalignment, while the DCT Stream specializes in detecting compression artifacts. Concurrently, the Segmentation Stream employs semantic segmentation to enhance object awareness within feature space. These streams converge at the fusion stage, where the captured artifacts are integrated into a single manipulation localization mask.

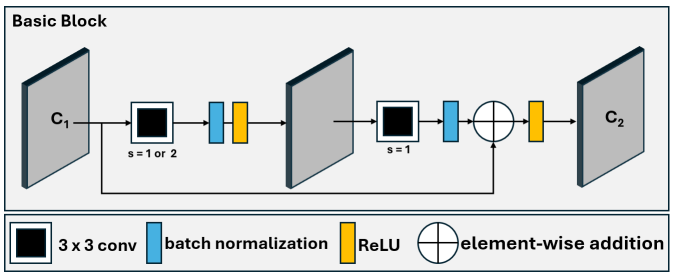

Fig. 3. Detailed view of a convolutional unit shown in Fig. 2. A single convolutional unit contains 4 consecutive basic blocks.

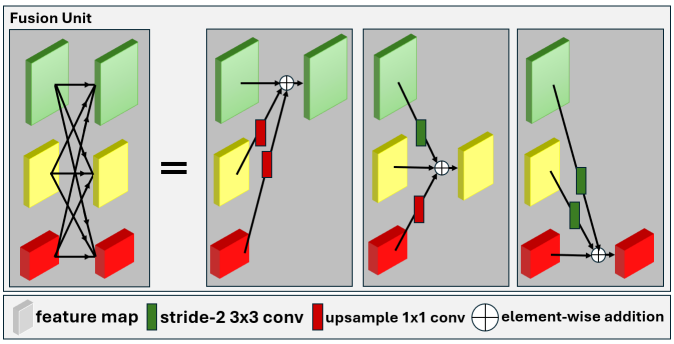

Fig. 4. Detailed breakdown view of the fusion fully connected layers in Fig. 2.

delineate manipulated regions. This heatmap can then be converted to a binary mask using a discrete threshold. We have chosen a threshold of 0.5. A detailed illustration of the model architecture is provided in Figure 2.

Basic Blocks: Figure 3 provides a detailed view of the internals of the convolutional units used in the three streams in Figure 2. These convolutional units are broken down into four consecutive basic blocks. Each basic block consists of two sequences of 3×3 convolution, batch normalization, and ReLU. This design helps stabilize the model during training.

Fusion Stage: The primary objective of the fusion stage involves two key components. First, it amalgamates the feature spaces derived from the three input streams. Second, it integrates and leverages multiple image resolutions once more before the final condensation process to generate the image manipulation heatmap. Figure 4 illustrates this process. The fully connected layer shown on the left side corresponds to the three subsequent stages delineated on the right side of Figure 4. Each stage is partitioned into two phases: (1) standardizing all feature spaces to uniform dimensions via up-scaling and down-scaling operations, and (2) performing element-wise addition across all three standardized feature spaces. This process is iteratively applied to feature spaces corresponding to each of the three image resolution sizes. Subsequently, a final consolidation step is executed by concatenating feature spaces from all three resolution sizes, culminating in the generation of the final heatmap.

Training: We utilized the splicing subset of the DEFACTO dataset [15] to train the new segmentation stream. The DEFACTO dataset is a large-scale image manipulation

Dataset	Year	# Manipulated Images	# Pristine Images	Image Size	Format	Manipulation Method
Columbia(Color) [10]	2006	180	183	757×568 - $1,152 \times 768$	TIF	Random
CASIA v1 [11]	2013	921	800	374×256	JPEG	Manual
Coverage [12]	2016	100	100	400×486	TIF	Manual
NIST16 [13]	2016	564	875	500×500 - $5,616 \times 3,744$	TIF	Manual
DSO-1 [14]	2013	100	100	2000×1500	PNG	Manual

TABLE I
DETAILS OF MAINSTREAM IMAGE MANIPULATION LOCALIZATION DATASETS.

localization dataset derived from the MSCOCO dataset. We chose DEFACTO because it was the only large-scale image manipulation dataset available to us at the time of training.

IV. EXPERIMENTAL EVALUATION

This work is part of the DARPA Semantic Forensics (SemaFor) program development, and the evaluation is performed as a SemaFor Evaluation 5 contest done by an independent evaluator institution. Our method underwent evaluation in the task of image manipulation detection and localization on a curated set of manipulated images that tries to mimic manipulations one would see in a real-world scenario. Furthermore, we also analyze and compare the image manipulation localization results on five mainstream public datasets, including CASIAv1 [11], COVERAGE [12], Columbia [10], NIST16 [13], and DSO-1 [14]. These datasets were selected for their diverse coverage of temporal spans and manipulation techniques. Details for these datasets can be found in Table I.

Mainstream datasets evaluation details: To assess and compare our proposed model against mainstream models, we selected five diverse datasets featuring various types of manipulations and image scenes. Each method is tasked with generating a heat map, subsequently converted into a binary mask using a consistent threshold of 0.5 across all participants to ensure fairness. An Intersection over Union (IoU) score is computed based on the overlap between the generated manipulation binary mask and the ground truth mask. The average IoU score is then generated by averaging all image-level IoU scores to derive the overall IoU for each dataset.

SemaFor evaluation 5 dataset comprises 140 manipulated images and 140 pristine images, with dimensions ranging from $2,000 \times 2,000$ to $4,000 \times 4,000$ pixels. The manipulation methods used are generative fill and generative expand. Generated fill involves selecting a region of the image and synthesizing a new object to splice into that region. Generative expand similarly generates new content, but expands the image into a larger size instead of modifying an existing region. This evaluation set was created to mimic realistic manipulations in real-world image scenes.

SemaFor evaluation 5 task details: To evaluate our proposed method along with other state-of-the-art (SoTA) methods against image manipulation datasets that contain generative fill and expand manipulations, we participated in the SemaFor evaluation image manipulation detection task. Due to the nature of the SemaFor program, all unpublished performer method names are replaced with a performer ID. The positive detection rate at a five percent false alarm rate is evaluated using two equations provided below. The true positive rate (TPR), which represents the positive detection rate, is calculated using the TPR equation. Conversely, the false positive rate (FPR), which denotes the false alarm rate, is determined using the FPR equation. The variables TP, FN, and TN denote true positive, false negative, and true negative, respectively. These measures are critical because a high positive detection rate at a low false alarm rate is essential in digital forensics applications. False detection not only wastes limited resources but also damages the trustworthiness of the model.

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN} \tag{1}$$

A. Evaluation on mainstream datasets

Table II shows a quantitative comparison of our proposed method against 14 SoTA methods. We have observed that the incorporation of a third branch dedicated to semantic segmentation within our proposed framework has yielded an average performance enhancement of 3% relative to the baseline of CAT-NET v2 [4]. Our approach secured the second position overall among existing state-of-the-art methods while TruFor achieved first place based on the average IoU metric across all five datasets. Notably, our method achieved first place in the CASIA and Columbia datasets. However, in the remaining public datasets, including Coverage, NIST16, and DSO-1, our method secured the second position. This is primarily due to the considerable performance disparity between our baseline Cat-Net v2 and TruFor. Therefore, while we observed a substantial performance improvement over our baseline, the disparity between Cat-Net v2 and TruFor remained too substantial for these specific datasets.

In general, our observations indicate that the proposed modifications have enhanced the localization performance of the baseline Cat-Net v2 model. We selected the Cat-Net v2 model due to its well-established performance and the convenience of its codebase, which includes both training and testing modules. Although we contemplated augmenting the TruFor model with an additional module focused on semantic segmentation to explore the impact of transformer-based architectures atop the CNN Cat-Net framework, the current version of the released code lacks any training components.

Regarding results in Table II, the most notable performance improvements were noted in the COVERAGE and DSO-1 datasets, whereas the least enhancement was observed in NIST16. Various factors may contribute to the relatively lower performance increase in the NIST16 dataset. Overall, it is evident that, among all participating datasets, NIST16 exhibits the lowest average scores across all existing mainstream methodologies. Additionally, the manipulated images within the NIST

Method	Casiav1	Coverage	Columbia	NIST16	DSO-1	Average
ADQ [16]	0.302	0.165	0.401	0.146	0.421	0.287
Splicebuster [17]	0.143	0.192	0.565	0.174	0.372	0.289
EXIF-SC [18]	0.106	0.164	0.798	0.227	0.442	0.347
CR-CNN [19]	0.481	0.391	0.631	0.300	0.289	0.418
RRU-NET [20]	0.408	0.279	0.575	0.154	0.312	0.346
ManTraNet [21]	0.180	0.317	0.508	0.172	0.412	0.318
SPAN [22]	0.112	0.235	0.759	0.228	0.233	0.313
AdaCFA [23]	0.128	0.183	0.403	0.106	0.235	0.211
IF-OSN [24]	0.553	0.304	0.753	0.330	0.470	0.482
MVSS-NET [25]	0.528	0.514	0.729	0.320	0.358	0.490
PSCC-NET [26]	0.520	0.473	0.604	0.113	0.458	0.434
Noiseprint [27]	0.137	0.229	0.513	0.196	0.439	0.303
CAT-NET v2 [4]	0.752	0.381	0.859	0.308	0.584	0.577
TruFor [28]	0.737	**0.600**	0.859	**0.399**	**0.930**	**0.705**
Proposed Method	**0.761**	0.460	**0.872**	0.310	0.630	0.607

TABLE II
IMAGE LOCALIZATION IOU RESULTS FROM EXISTING STATE OF THE ART METHODS ON FIVE PUBLICLY AVAILABLE MAINSTREAM DATASETS.

Performer ID	PD@0.05 FAR	EER
1	**0.577**	**0.175**
2	0.288	0.289
3	0.158	0.367
4	0.150	0.442
5	0.042	0.468
Proposed	0.319	0.273

TABLE III
SEMAFOR EVALUATION IMAGE MANIPULATION DETECTION TASK RESULTS

dataset often depict synthetic scenes featuring checkerboards and a series of 3D shapes. These highly synthetic scenes are often not included in the training, as most existing datasets train off of natural images.

B. Results from SemaFor IMD Evaluation

As shown in Table III, all participating performers struggled to detect the image manipulations in this task. This indicates that existing state-of-the-art methods still have difficulty adapting to new types of image manipulations, such as generative stable diffusion, generative fill, and generative expand. Notably, generative expansion conflicts with many assumptions made by current image manipulation detection and localization models. Defensive models often make the assumption that the manipulated region contains only a single object. However, generative expansion typically creates a manipulation region in the form of a rectangular region. This region can contain multiple entities or background regions.

This new image manipulation capability requires defensive models to move beyond focusing solely on objects as the manipulation region. This highlights the importance of having multiple input streams that can capture a variety of manipulation-relevant artifacts. By doing so, future defensive models will be better equipped to adjust to new manipulation types, improving their robustness and adaptability in the face of evolving image manipulation techniques.

C. Qualitative Analysis

In our comprehensive evaluation across all five datasets, the prominence of the spliced object within the image emerges as a crucial factor. Most failure instances are associated with spliced objects that differ significantly from those encountered during training. This trend is particularly evident in the NIST16 dataset, where many manipulated objects consist of wooden geometric shapes set against a checkerboard background. Such scenes are vastly different from typical real-world imagery, highlighting the importance of the semantic context of the tampered object during model training.

This observation aligns with the notion that while models strive to capture abstract low-level signatures, the semantic content of the tampered object remains highly important at inference time. This explains why all performers in the SemaFor evaluation struggled to identify the correct tampered region. The generative expand technique creates manipulated regions that are not confined to a specific tampered object; instead, these regions are typically rectangular and completely synthetic. This divergence from traditional manipulation patterns emphasizes the need for models that can adapt to new types of image manipulations by incorporating diverse input streams and focusing on a variety of manipulation artifacts.

Figure 5 showcases the localization results of a proposed image forgery detection method, displayed in grayscale with manipulated areas highlighted in color. Correctly identified manipulated regions are marked in green, while misidentified areas, including false positives and negatives, are shown in red. The figure spans results across five different datasets, each featuring varied types of image scenes. Datasets like DSO-1, CASIA, and Coverage mainly contain natural scenes that mirror real-life environments. In contrast, the NIST16 dataset includes images with synthetic qualities that are less typical of natural settings, making it the most challenging dataset for image manipulation detection among the five. The Columbia dataset, although synthetic, tends to have lower-quality manipulations, which are easier for most defense methods to detect compared to the high challenge posed by NIST16.

An analysis of binary masks produced by the proposed model across multiple datasets revealed a distinct trend: the masks are either characterized by high Intersection over Union (IoU) scores or are nearly blank, indicating a significant polarization in the model's decision-making. This polarization shows that the model is effective in reducing false positives but has the risk of missing actual manipulations (non-detections or mis-detections).

D. Overall Evaluation Analysis

Based on the observed results, integrating a third input stream dedicated to semantic segmentation can significantly improve the performance of CNN-based defensive models in image manipulation localization tasks. Our results indicate an average enhancement of three percent IoU across five mainstream methodologies. However, it is evident that current defensive models struggle to adapt to emerging manipulation techniques. It is foreseeable that the efficacy of defensive methods will continue to improve as the community develops a more extensive array of training and testing resources.

Limitations: While semantic segmentation can enhance image manipulation localization, it can also introduce noise

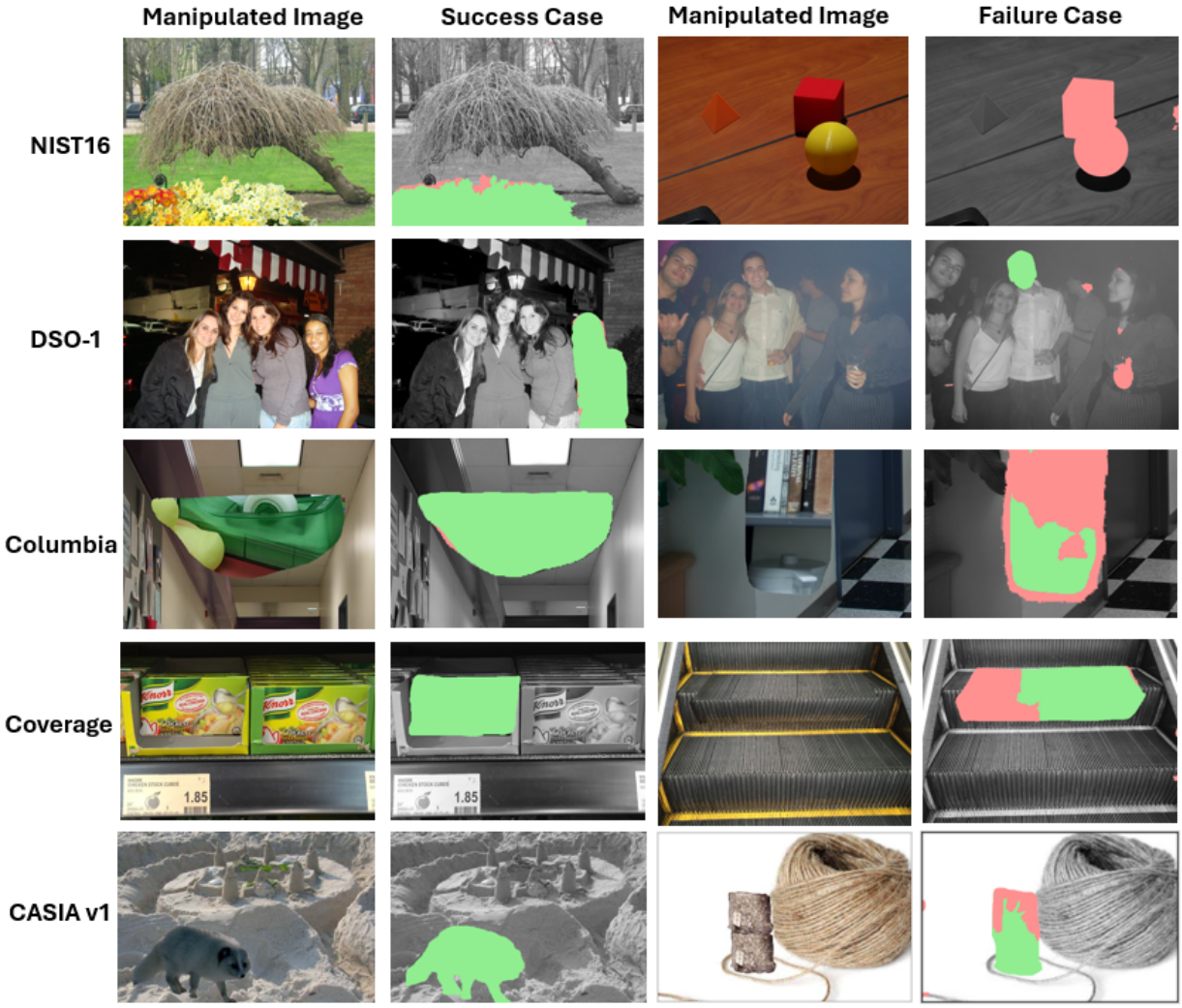

Fig. 5. Qualitative examples showing the proposed method localization results for all five participating mainstream datasets. Green regions represent image regions that have been correctly identified as manipulated. Red regions represent image regions that have been mis-identified. This can take the form of false positive and false negative image regions.

by including irrelevant regions. There is also a risk of incomplete segmentation, where object boundaries are inaccurately defined. This can limit its effectiveness by failing to adequately support partially detected manipulation regions.

V. Conclusion

In this study, we introduced a novel approach to enhance the detection and localization of image manipulation by integrating semantic segmentation into an end-to-end model. This integration incorporates object awareness and balances visual, JPEG compression, and segmentation artifacts, addressing the limitations of current techniques that often overly rely on a single artifact type. Our method significantly improves the robustness and accuracy of spliced object localization, as demonstrated by extensive evaluations on established datasets. Additionally, our research provides valuable insights into the adaptability of existing digital defense models against emerging image tampering techniques like generative fill and generative expand. This adaptability is crucial as image manipulation methods continue to evolve.

Future Works: To minimize the chances of the semantic segmentation introducing additional noise, we would like to explore filtering the semantic segmentation of interest. One plausible approach to filtering irrelevant segmentation is filter segmentation based on entities present in the image caption or description. The assumption is that the image manipulation is semantically significant. Therefore, this newly spliced entity will be present within the image description.

Disclaimer: This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Acknowledgments. This work is supported by the DARPA Semantic Forensics (SemaFor) Program under contract HR001120C0123 and NSF CCSS-2348046. The authors appreciate the computational resource provided by the University at Albany.

REFERENCES

[1] D. Raković, "Error level analysis (ela)," *Tehnika*, vol. 78, pp. 445–451, 01 2023.

[2] F. Pérez-González, G. Heileman, and C. Abdallah, "Benford's law in image processing." vol. 1, 09 2007, pp. 405–408.

[3] A. Popescu and H. Farid, "Statistical tools for digital forensics," vol. 3200, 05 2004.

[4] M.-J. Kwon, S.-H. Nam, I.-J. Yu, H.-K. Lee, and C. Kim, "Learning JPEG compression artifacts for image manipulation detection and localization," *International Journal of Computer Vision*, vol. 130, no. 8, p. 1875–1895, May 2022. [Online]. Available: http://dx.doi.org/10.1007/s11263-022-01617-5

[5] J. Fridrich, D. Soukal, and J. Lukás, "Detection of copy-move forgery in digital images," *Int. J. Comput. Sci. Issues*, vol. 3, pp. 55–61, 01 2003.

[6] A. Popescu and H. Farid, "Exposing digital forgeries in color filter array interpolated images," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3948–3959, 2005.

[7] Z. Zhang, M. Li, and M.-C. Chang, "A new benchmark and model for challenging image manipulation detection," in *AAAI*, 2024.

[8] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," 2020.

[9] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023.

[10] T.-T. Ng, S.-F. Chang, , and Q. Sun, "A data set of authentic and spliced image blocks," *Columbia University ADVENT Technical Report*, 2004.

[11] J. Dong, W. Wang, and T. Tan, "CASIA image tampering detection evaluation database," http://forensics.idealtest.org, 2010.

[12] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, and S. Winkler, "COVERAGE — a novel database for copy-move forgery detection," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 161–165.

[13] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith, and Jonathan, "MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation." IEEE Winter Conference on Applications of Computer Vision (WACV 2019), Waikola, HI, 2019-01-11 00:01:00 2019.

[14] T. Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. Rocha, "Exposing digital image forgeries by illumination color classification," *Information Forensics and Security, IEEE Transactions on*, vol. 8, pp. 1182–1194, 07 2013.

[15] G. MAHFOUDI, B. TAJINI, F. RETRAINT, F. MORAIN-NICOLIER, J. L. DUGELAY, and M. PIC, "Defacto: Image and face manipulation dataset," in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[16] T. Bianchi, A. De Rosa, and A. Piva, "Improved DCT coefficient analysis for forgery localization in JPEG images," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2444–2447.

[17] D. Cozzolino, G. Poggi, and L. Verdoliva, "Splicebuster: A new blind image splicing detector," 11 2015.

[18] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Fighting fake news: Image splice detection via learned self-consistency," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[19] C. Yang, H. Li, F. Lin, B. Jiang, and H. Zhao, "Constrained R-CNN: A general image manipulation detection model," 2020.

[20] X. Bi, Y. Wei, B. Xiao, and W. Li, "RRU-Net: The ringed residual u-net for image splicing forgery detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[21] Y. Wu, W. AbdAlmageed, and P. Natarajan, "ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9535–9544.

[22] X. Hu, Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang, and R. Nevatia, "SPAN: Spatial pyramid attention network for image manipulation localization," in *European Conference on Computer Vision (ECCV),*. Springer, 2020, pp. 312–328.

[23] Q. Bammey, R. G. von Gioi, and J.-M. Morel, "An adaptive neural network for unsupervised mosaic consistency analysis in image forensics," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 182–14 192.

[24] H. Wu, J. Zhou, J. Tian, and J. Liu, "Robust image forgery detection over online social network shared images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 13 440–13 449.

[25] X. Chen, C. Dong, J. Ji, J. Cao, and X. Li, "Image manipulation detection by multi-view multi-scale supervision," 2021.

[26] X. Liu, Y. Liu, J. Chen, and X. Liu, "PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7505–7517, 2022.

[27] D. Cozzolino and L. Verdoliva, "Noiseprint: a CNN-based camera model fingerprint," 2018.

[28] M.-J. Kwon, S.-H. Nam, I.-J. Yu, H.-K. Lee, and C. Kim, "Learning JPEG compression artifacts for image manipulation detection and localization," *International Journal of Computer Vision*, vol. 130, no. 8, p. 1875–1895, May 2022. [Online]. Available: http://dx.doi.org/10.1007/s11263-022-01617-5