A Deep Learning Approach for Locating Shallow and Deep Image Manipulations

Guerbej Salsabil
Computer Society chapter
IEEE ENIT SB
Tunis , Tunisia
salsabil.guerbej@ieee.org

Mallouli Emna
Computer Society chapter
IEEE ENIT SB
Tunis , Tunisia
mallouliemna@ieee.org

Fathallah Ali
Computer Society chapter
IEEE ENIT SB
Tunis , Tunisia
ali.fathallah@ieee.org

***Abstract*—**With the rise of manipulated visual content, particularly in the forms of ”shallowfakes” created through image editing tools and ”deepfakes” generated by advanced artificial intelligence techniques, the need for effective image localization solutions has become paramount. Existing works often specialize in either shallow- or deep-fake detection, lacking a unified approach. This paper introduces a groundbreaking solution capable of localizing manipulated areas in both shallow- and deep-fake images with high inference accuracy based on the Shallow- and Deep- fake Image Manipulation Localization Using Deep Learning Model , A research made by Junbin Zhang, Hamidreza Tohidypour, Yixiao Wang, and Panos Nasiopoulos [1]. The code and dataset are available at : **https://github.com/dumyysal/Shallowdeepfakesdetection.**

Keywords*—*Image manipulation, Manipulation localization, shallowfakes, Deepfakes

I. Introduction

Images, integral to our daily lives, hold significant informational value. However, the manipulation of images to create false narratives and fraud poses a severe threat , especially when disseminated rapidly on a large scale through the Internet. Especially today, you can easily edit your photos or videos using basic editing tools like Photoshop, where the term "shallow fakes" [2] is applicable. Alternatively, the integration of AI technologies, particularly the intersection of deep learning and computer graphics, has introduced more sophisticated techniques, notably seen in "deepfakes" These advanced manipulations facilitate the seamless swapping or modification of people's identities, particularly their faces in images or videos. The widespread accessibility of these tools, coupled with the prevalence of social media, has led to significant reputational damage for celebrities, executives, and politicians.

In June 2022, a Pennsylvania mother, dubbed the 'Deepfake mom,' received a three-year probation sentence for harassing rivals on her daughter's cheerleading team [3] . Originally accused of employing Deepfake technology to create and disseminate fake videos depicting inappropriate behaviors by her daughter's opponents in March 2021, the prosecutors acknowledged the difficulty of confirming Deepfake video generation without precise evidence and tools [4] . In efforts to safeguard individuals and society from the detrimental consequences of Deepfake misuse, various Deepfake detection approaches have been developed. These approaches predominantly engage in a binary classification task, employing deep neural networks (DNNs) to discern between authentic and manipulated faces. Furthermore, these methods extend their focus to address the challenge of identifying the regions subject to forgery. This involves generating a mask that precisely delineates the manipulated portions within an image.

In this study, we offer an in-depth review of the evolving approaches in shallow Deepfakes localization and manipulation models, considering their reliability perspective. The rest of the paper is organized as follows: In Section II, we provide a detailed introduction to the outlined approach of our model. In Section III, we evaluate the detection performance and justify reliability using selected state-of-the-art models and then exhibits the optimization outcomes. Finally, Section IV concludes with remarks and highlights potential future directions in the research domain.

II. Approcah Outline

A. Network Design

The proposed solution builds upon the UPerNet [5] architecture, originally designed for recognizing multiple visual concepts within a scene simultaneously. In addition to its multi-task framework, our model includes key components like a Pyramid Pooling Module (PPM) head [6] , a detection head, a Feature Pyramid Network (FPN), and a localization head. We've also integrated BayerConv filters to effectively extract noise distribution. The structure of the network is shown in Fig. 1 .

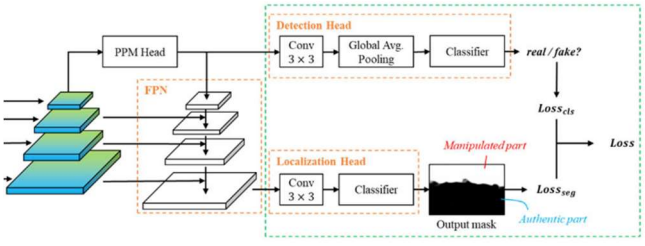

Fig. 1. Part of The network’s Structure

To help the network better understand differences between authentic and manipulated areas, similar to prior works, we've added BayerConv to enhance its capabilities.

This addition improves the model's ability to discern subtle variations between genuine and manipulated content.
We have implemented a loss function designed to quantify the disparity between the model's predictions and the actual ground truth image.

B. Dataset Construction

Our dataset is split into two parts: one with shallowfake images and another with deepfake images.

For shallow fakes, we incorporated datasets such as CASIAv1 [8], CASIAv2 [8], Columbia [9], COVERAGE [10] , and NIST 16 [11]. The resulting test set comprises 1,832 real images and 2,259 fake images, encompassing all three types of shallowfake manipulations (slicing, copy-move, and inpainting)

For deep fakes, no pre-existing image dataset with ground truth masks of manipulated areas was available. Consequently, we constructed the dataset ourselves. For detailed insights into the dataset construction process, readers are encouraged to check the original paper [1].

III. TESTS AND RESULTS

A. Reproducing The Outcomes

We conducted a comparative analysis involving our network, both with and without BayerConv, against the state-of-the-art MVSS-Net [7]. The training of each network was carried out under three distinct configurations: (1) using only the shallowfake training set, (2) using only the deepfake training set, and (3) using both the shallow- and deep-fake training sets. Following the training phase, we evaluated the model's performance using a deep fake dataset.

We reported three different F1 values:
(1) pixel-level, which represents the accuracy of output masks compared to the ground-truth masks for localization task.
(2) image-level metric, denoted as f 1, assesses the agreement between the model's binary output results (real/fake) and the ground truth in the context of a detection task. The f1 score is the harmonic mean of sensitivity and specificity. Sensitivity (Sen) gauges the model's capability to correctly identify positive instances among the actual positives, while specificity (Spe) measures the model's aptitude for correctly identifying negative instances among the actual negatives.
(3) a combined F1 value, which is the harmonic mean of the pixel-level F1 and image-level F1.

We set the detection threshold as 0.5 for these three F1 values and reported best threshold in the testing phase values. We also reported image-level "Area under the Receiver Operating Characteristic (ROC) Curve" (AUC) and And G-mean values which is a valuable metric that takes into account the balance between true positives and true negatives, offering insights into the model's overall effectiveness in handling imbalanced datasets. A higher G-mean indicates better performance in simultaneously optimizing sensitivity and specificity.

TABLE I. EVALUATION RESULTS OF THE **MVSS-NET** NETWORK EXPRIMENTS . IN THE SECOND COLUMN , "S" MEANS SHALLOWFAKE DATASET , AND "D" MEANS DEEPFAKE DATASET. WE HIGHLIGHT THE BEST RESULTS USING RED UNDERLINE , AND THE SECOND BEST RESULTS USING BLACK UNDERLINE .

	Training Set		Test Set = Deepfakes							
					F1					
						Image-Level				
	S	D	Best Thre-hold	G-Mean	Pixel - Level	Sen	Spe	f 1	Combined	AUC
Exp 1	■		0.768627	0.527	0.2395	0.5675	0.4662	0.5119	0.3263	0.5313
Exp 2		■	0.827451	0.936	0.8964	0.9263	0.9402	0.9332	0.9144	0.9774
Exp 3	■	■	0.623529	0.938	0.9033	0.9354	0.9366	0.9360	0.9194	0.9687

TABLE II. EVALUATION RESULTS OF **OURS W/O BAYERCONV** NETWORK EXPRIMENTS . IN THE SECOND COLUMN , "S" MEANS SHALLOWFAKE DATASET , AND "D" MEANS DEEPFAKE DATASET. WE HIGHLIGHT THE BEST RESULTS USING RED UNDERLINE , AND THE SECOND BEST RESULTS USING BLACK UNDERLINE.

	Training Set		Test Set = Deepfakes							
					F1					
						Image Level				
	S	D	Best Thre-hold	G-Mean	Pixel - Level	Sen	Spe	f 1	Combined	AUC
Exp 4	■		0.128215	0.537	0.2997	0.1742	0.8323	0.2881	0.2938	0.5490
Exp 5		■	0.402647	0.955	0.9218	0.9513	0.9544	0.9529	0.9371	0.9887
Exp 6	■	■	0.381297	0.930	0.8659	0.9195	0.9378	0.9286	0.8961	0.9790

TABLE III. EVALUATION RESULTS OF **OURS WITH BAYERCONV** NETWORK EXPRIMENTS . IN THE SECOND COLUMN , "S" MEANS SHALLOWFAKE DATASET , AND "D" MEANS DEEPFAKE DATASET. WE HIGHLIGHT THE BEST RESULTS USING RED UNDERLINE , AND THE BEST SECOND RESULTS USING BLACK UNDERLINE.

	Training Set		Test Set = Deepfakes							
					F1					
						Image Level				
	S	D	Best Thre-Hold	G-Mean	Pixel - Level	Sen	Spe	f 1	Combined	AUC
Exp 7	■		0.155277	0.474	0.2731	0.1751	0.7713	0.2854	0.3701	0.4672
Exp 8		■	0.525390	0.955	0.9238	0.9504	0.9591	0.9548	0.9390	0.9904
Exp 9	■	■	0.307297	0.945	0.9097	0.9322	0.9550	0.9435	0.9263	0.9877

Effects of training sets. When we looked at all three different networks, we noticed that training with both shallow- and deep-fake datasets didn't cause a big change in accuracy for each set on its own. However, what we did find is that training with both sets allowed the networks to pick up on general features needed to spot both shallow- and deep-fake images.
For example, in Table III Exp 9, where the test set includes a mix of deep-fake images, it is evident that training with both the shallow- and deep-fake sets leads to the highest combined (f 1) score (0.9435) and AUC (0.9877). This outperforms the results obtained by testing the other models.

Effects of network design. Comparing results in Tabs I and II we conclude that using a network based on UPerNet can achieve a better result compared to MVSS-Net, regardless of what training set is used during training. For example, when training using both shallow- and deep-fake training sets, our network (even without the BayerConv) achieves an AUC of 0.9790 (Tab II Exp 6).
Besides, feeding the noise features extracted by BayarConv can further improve the inference accuracy. For networks trained with only deepfake images, with and without BayerConv achieve the same level of combined F1 of 0.9435 and AUC of 0.9877. When training with both test sets (Tab I Exp 3, Tab II Exp 6, Tab III Exp 9)

Qualitative visualization. For visualization purposes, we show some samples of output masks of the MVSS-Net network comparing with our model with BayerConv in Fig.2.

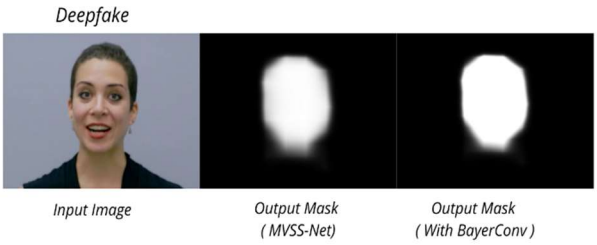

Fig. 2. Samples of network outputs for the two networks.

The output of the network with BayerConv is clear on the edges comparing to MVSS-Net model.

B. *Optimizing Phase*

- **Alternative DataSet**

During the optimization phase, we opted to use alternative new datasets. Our dataset was curated to include both types of images—shallow fakes and deep fakes . The objective was to evaluate how our model would predict and respond to new data. The recorded values above illustrate the outcomes of this experimentation.

- **Testing phase and results**

TABLE IV. Evaluation results of **ours with BayerConv** network expriments on the new alternative dataset . In the second Column,"S" means shallowfake dataset , and "D" means deepfake dataset

	Training Set		Test Set = Deep And Shallow fakes	
	S	D	Pixel – f1	Image level acc
Exp 7	■		0.6629	0.2921
Exp 8		■	0.6742	0.3843
Exp 9	■	■	0.9565	0.9565

It is evident that our new dataset has yielded noteworthy results, particularly in Experiment 9, where the model was trained using both types of datasets. The image-level accuracy reached 0.9565, showcasing a substantial improvement compared to Experiment 7, where the model was trained solely on the shallow dataset and achieved an accuracy of 0.2921.

Outputs visualization.

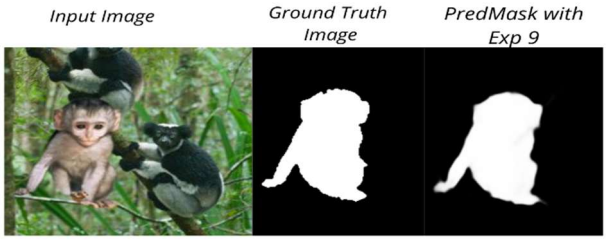

Fig. 3. Samples of testing outputs

The input image in Figure 3 belongs to the shallow type, and it is a novel addition for the model. Despite this, the output mask from Experiment 9 closely aligns with the ground truth image. This observation explains the success of the model's performance when confronted with new data, emphasizing its adaptability and accuracy in handling previously unseen shallow-type images

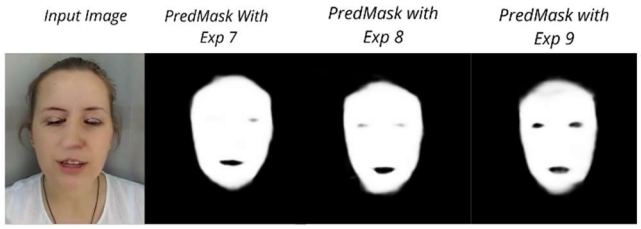

Fig. 4. Samples of testing outputs

In the figure, a series of predicted masks from experiments 7 to 9 is presented. Notably, the masks become progressively more authentic as the model is trained with a broader spectrum of datasets. This trend highlights the positive impact of incorporating diverse data types in the training process, enhancing the model's ability to generate more accurate and realistic predictions.

IV. CONCLUSION

In this paper, we trained and tested the model to replicate its results. Additionally, we conducted an optimization phase where we constructed a new test dataset to evaluate the model's performance. The results demonstrated that our model achieved an impressive accuracy score of 95% on new data. Future work will explore the model's suitability for diverse datasets, aiming to assess its generalizability across different scenarios.

ACKNOWLEDGMENT

This research was made possible, in part, through participation in the 11th edition of the Tunisian Student Young Professional (TSYP). The goal was to contribute to the challenge organized by Computer Society Chapter and Young Professionals. (Https://tsyp.ieee.tn).

REFERENCES

[1] *J*. Zhang, H. Tohidypour, Y. Wang and P. Nasiopoulos, "Shallow- and Deep- fake Image Manipulation Localization Using Deep Learning," 2023 International Conference on Computing, Networking and Communications (ICNC), Honolulu, HI, USA, 2023, pp. 468-472, doi: 10.1109/ICNC57223.2023.10074246.

[2] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in CVPR, 2018..

[3] K. Katro, "Bucks county mother gets probation in harassment case involving daughter's cheerleading rivals." https://tinyurl. com/yh82nz45, 2022, accessed: 2022-09-08.

[4] D. Harwell, "Remember the 'deepfake cheerleader mom'? prosecutors now admit they can't prove fake-video claims," https: //tinyurl.com/ysz42r9r, 2021, accessed: 2022-09-08

[5] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in ECCV, 2018

[6] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, "Pyramid scene parsing network," in CVPR, 2017.

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

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016.

[9] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, "Pyramid scene parsing network," in CVPR, 2017.

[10] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, "Feature pyramid networks for obbject detection," in CVPR, 2017.

[11] J. Dong, W. Wang, and T. Tan, "CASIA image tampering detection evaluation database," in ChinaSIP, 2013.