Deep Fake Localization

Mihail-Dănuț Dogaru, Eduard-Valentin Dumitrescul and Stefan Smeu
University of Bucharest, Romania

mihail-danut.dogaru@s.unibuc.ro, eduard-valentin.dumitrescul@s.unibuc.ro and ssmeu@bitdefender.com

1. Introduction

The task consists in detecting and localizing areas in an image that were manipulated and analyzing how results transfer from one deep fake generation method to another.
The main difficulty comes from the complexity of the data and the high amount of computational resources needed to process this data.

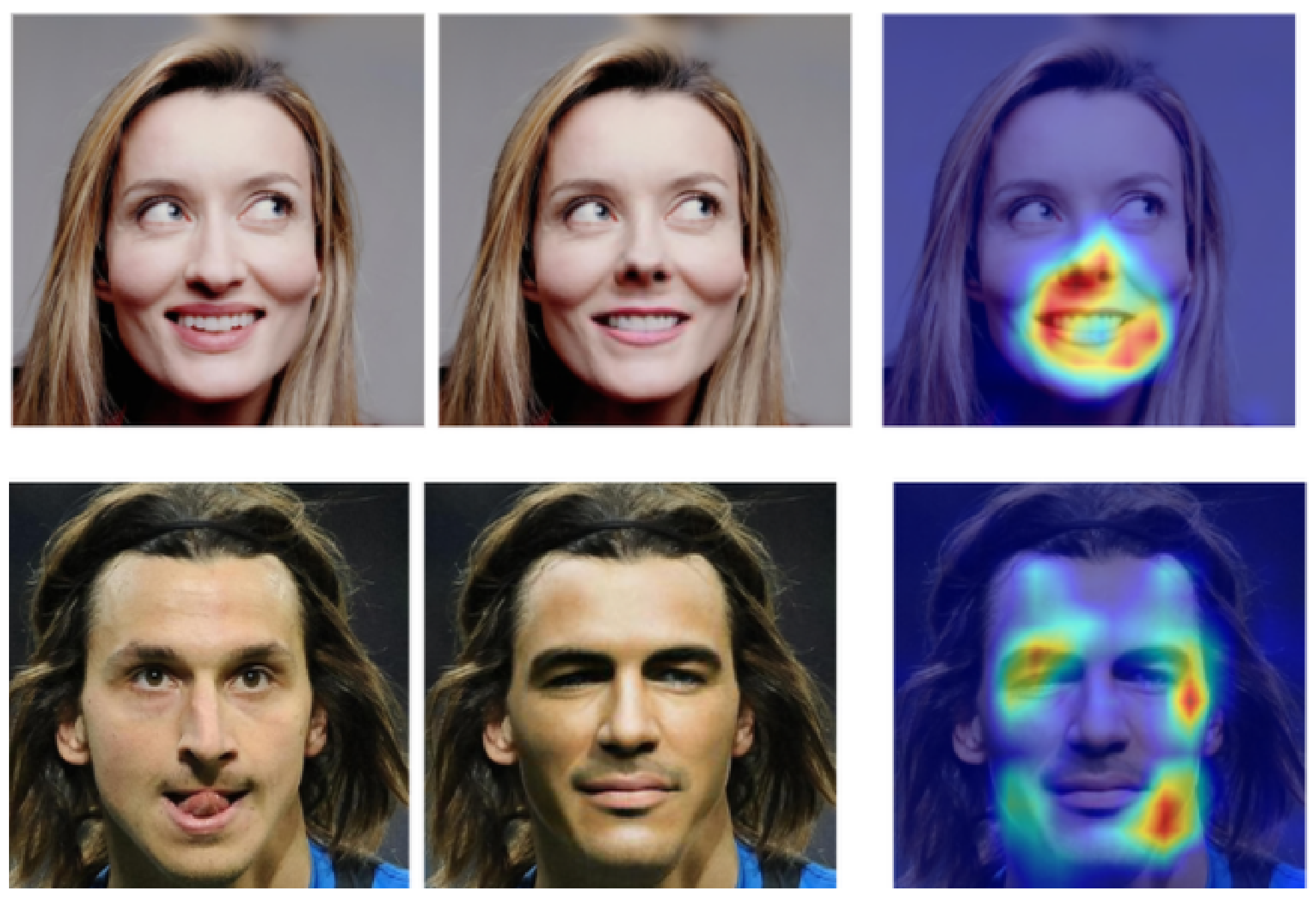

Figure 1: Task Example

2. Dataset and Preprocessing

Dataset Description:
Our dataset consists of the images in CelebAHQ dataset, each of which has been manipulated using 3 different masks and 4 different deep fake generation methods (Lama, Ldm, Pluralistic, Repaint) in order to create new images.
This results in 3600 256x256 real images and 10800 256x256 generated images for each deep fake generation method.
Data preprocessing:
We computed the contour for each mask to comply to the model's requirements. The images were not altered in any way.

3. Model

For the task of deepfake localization in facial images, we used a custom segmentation model based on the YOLOv8 architecture, defined via the `yolo11n-seg.yaml` configuration file. This model, referred to as **YOLOv8n-seg**, is a lightweight, real-time instance segmentation network designed to detect objects inside images.
To evaluate its performance and generalizability, we trained and tested the model for 10 epochs under two experimental setups:

- **Intra-method setting:** Training and testing were performed on data generated using the same deepfake technique.
- **Cross-method setting:** The model was trained on data from three different deepfake generation methods and tested on a fourth, unseen method.

These setups help assess both the model's detection accuracy and its ability to generalize to unseen manipulation techniques, which is crucial for robust deepfake localization.

4. Results

Table 1. Intra-method evaluation results

Dataset	IOU(threshold 0.5)	Accuracy	Precision	Recall	F1
Lama	0.60	0.38	0.73	0.45	0.55
LDM	0.27	0.21	0.46	0.28	0.35
Pluralistic	0.69	0.49	0.70	0.61	0.65
Repaint	0.53	0.35	0.67	0.42	0.52

Table 2. Cross-method evaluation results

Test on	IOU(threshold 0.5)	Accuracy	Precision	Recall	F1
Lama	0.20	0.13	0.41	0.16	0.23
LDM	0.17	0.17	0.35	0.26	0.30
Pluralistic	0.19	0.15	0.42	0.20	0.2
Repaint	0.29	0.20	0.52	0.25	0.34

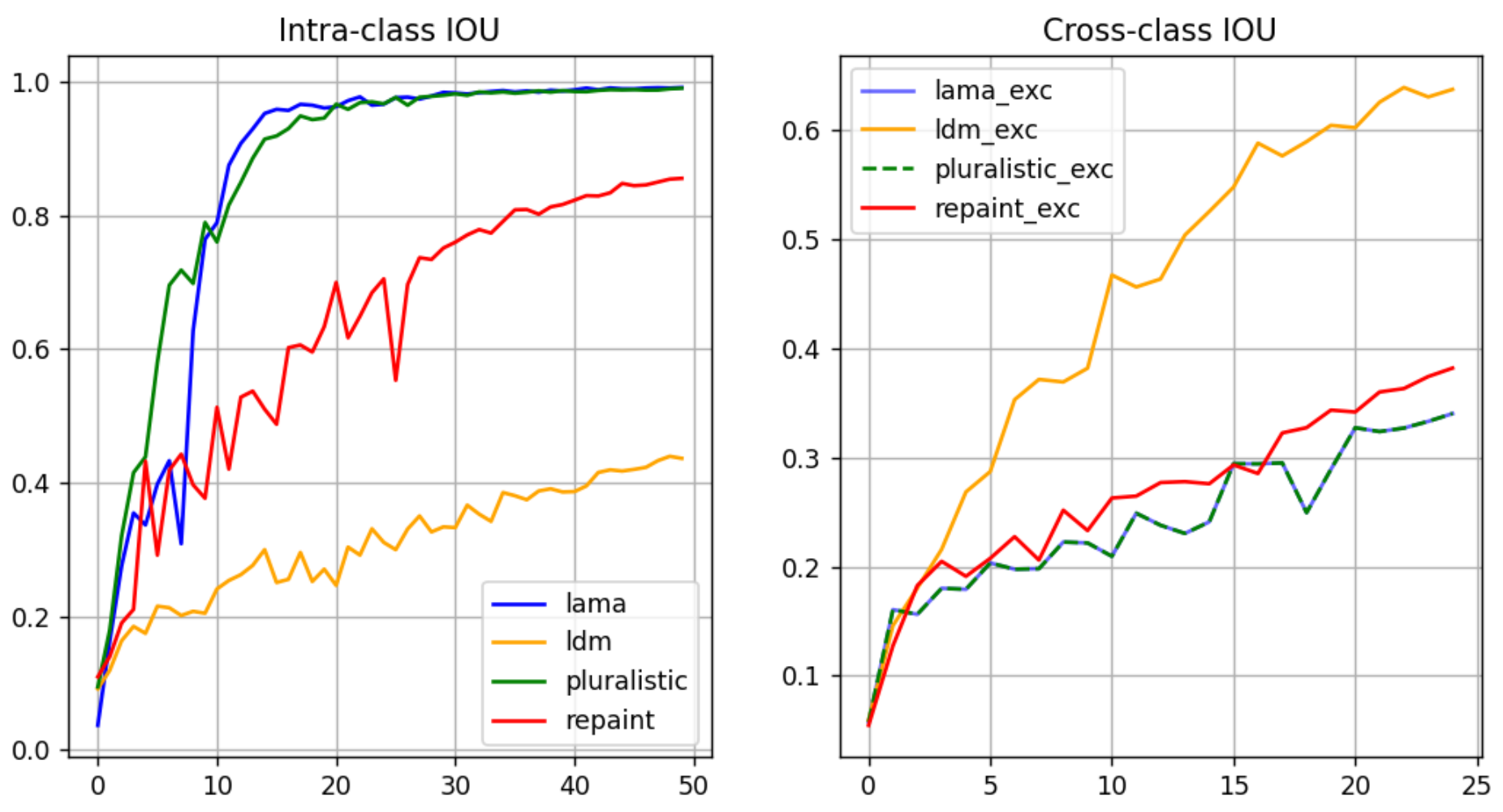

Figure 2: IOU (Intersection Over Union) plots during training.

Observations:

- In the cross-method setting, the model performed poorly, highlighting the difficulty of generalizing across generation methods.
- Hardware limitations restricted training; better results may be possible with more data and epochs.
- Lama and Pluralistic showed similar behavior to their training splits, with minor differences in test results.

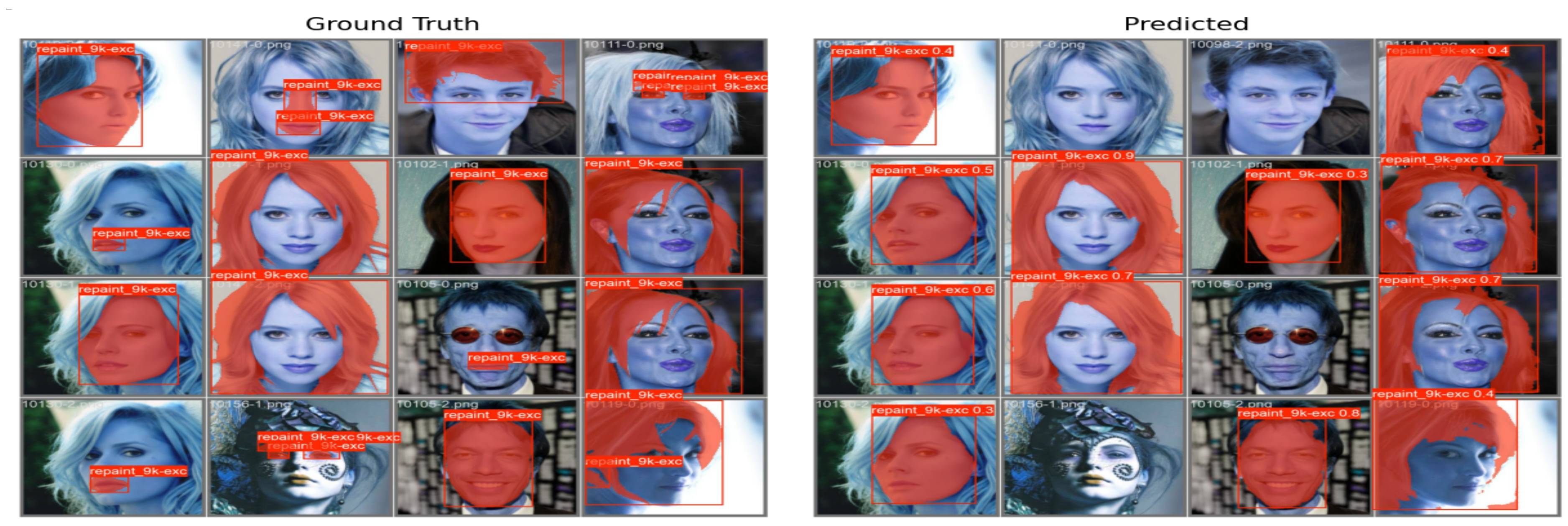

Figure 3: Labels vs Prediction - Cross-Method Repaint

5. Conclusion

The results demonstrate that using the YOLO model may be effective for deepfake localization, particularly when it is trained on data generated by the same method. However, further testing is required to assess its general applicability, as it is unclear whether the model truly learns to detect generated regions or merely memorizes the shape of the masks and bases its predictions on that.
In the future, more diverse masks, a larger dataset, and improved hardware will be necessary to thoroughly evaluate and enhance the model's performance.