Deepfake Localization

Vlad-Mihai Ciuperceanu, Sebastian Popa, Stefan Smeu* and Elisabeta Oneata*

University of Bucharest, Romania;
*Bitdefender, Romania
vlad-mihai.ciuperceanu@s.unibuc.ro, sebastian.popa@s.unibuc.ro,
ssmeu@bitdefender.com, eoneata@bitdefender.com

UNIVERSITY OF BUCHAREST
VIRTUTE ET SAPIENTIA

1. Introduction

The rapid advancement of generative AI has made image manipulations increasingly subtle, raising the need for **deepfake localization** — detecting where forgeries occur in an image. Unlike image-level detection, localization provides fine-grained, interpretable results, critical for **applications** such as fake news prevention and forensic analysis.

However, localization is a **challenging task**: manipulated regions are often visually imperceptible, datasets lack pixel-level labels, and models struggle to generalize - especially to unseen generators like latent diffusion models.

To address this, we propose using **SAM2** (*Ravi et al.*, 2024) as a visual backbone, leveraging its segmentation power to capture local forgery cues more effectively than vision-language approaches (e.g., DeCLIP (*Smeu et al.*, 2025)). Our supervised framework aims for precise localization while improving generalization to unseen generators.

We show SAM2 **enhances** accuracy on fine-grained in-domain manipulations and analyze strategies to boost out-of-domain **robustness** against emerging synthetic techniques.

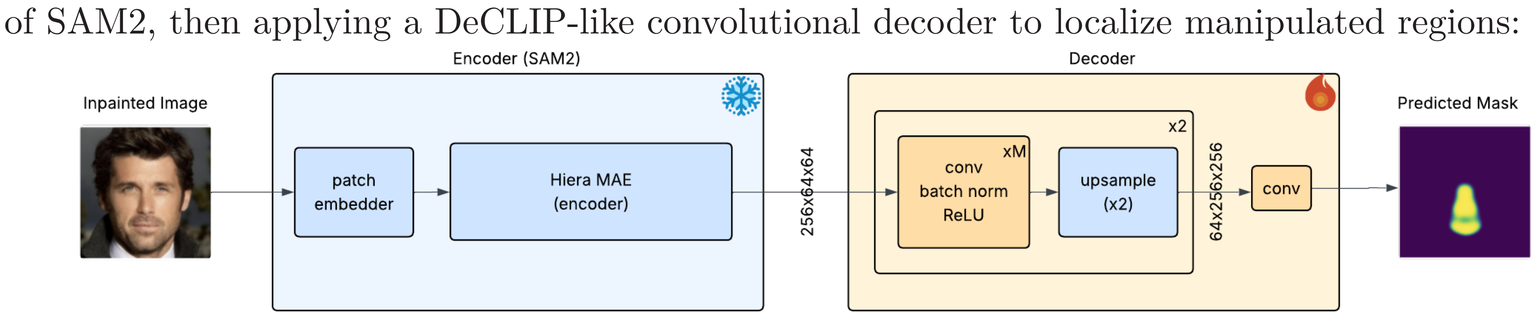
2. Dataset

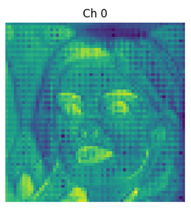
Dolos (*Tantaru et al.*, 2023) is a dataset for localizing subtly manipulated facial images, featuring edits to features like mouth, eyes, and hair via inpainting. It comprises four subsets created using different GAN/diffusion models. Dolos offers forged images and ground truth masks for controlled analysis of specific forgeries and cross-method generalization.

Cross-domain:
To evaluate out-of-domain generalization, we built cross-domain training sets. Each experiment trained and validated on balanced samples from three datasets, testing on a held-out fourth. We randomly selected 3,000 training and 300 validation images from each of the three training sets (9,000/900 total), assessing if training on diverse manipulations enhances generalization to new forgery types with consistent computational cost.

3. Methods

SAM2 is a segmentation model built on a hierarchical transformer encoder inspired by Masked Autoencoders (MAE), which processes images at multiple spatial resolutions. Thus, it extracts multi-scale features that capture both global context and fine spatial details.
In this work, we used an encoder-decoder architecture, extracting features from the frozen encoder of SAM2, then applying a DeCLIP-like convolutional decoder to localize manipulated regions:



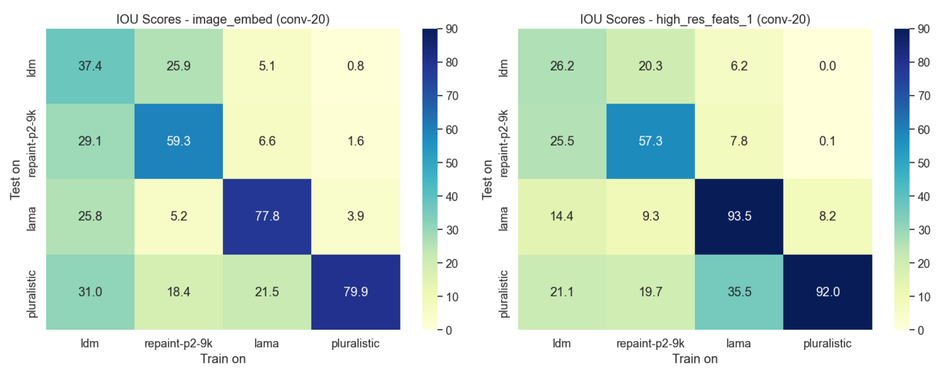


We leverage three key outputs from the frozen SAM2 encoder, extracted from different stages of the hierarchical encoder: *image_embed* (compact global representation of size 256×64×64), *high_res_feats[0]* (early-stage features of size 32×256×256) and *high_res_feats[1]* (intermediate-level features of size 64×128×128).

Training Setups: The model is trained in a fully-supervised manner on each of the 4 deepfake datasets over 20 epochs, using BCE loss and an Adam optimizer of learning rate 0.001. Later, we used our custom cross-domain training sets to investigate how transferable are the learned representations.
Evaluation: To evaluate the model's performance, we employ several standard metrics for localization, such as Intersection over Union (IoU), F1-score and pixel-wise average precision, computed both in-domain (ID) and out-of-domain (OOD) to assess the model's generalization ability.

4. Results

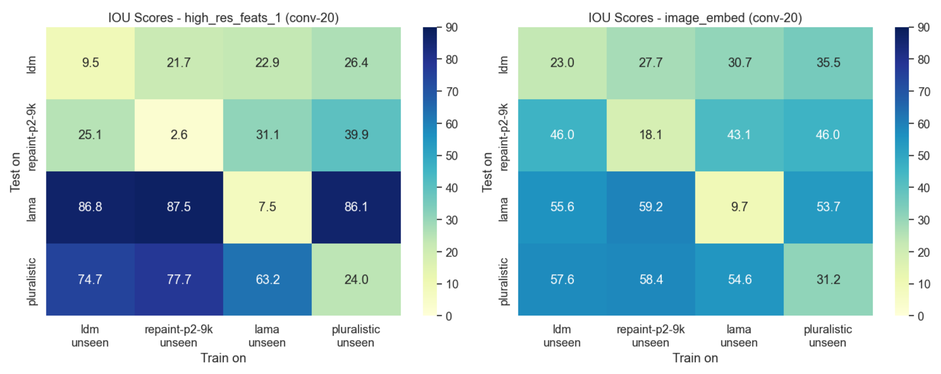
We experimented with decoders of different sizes on *image_embed*, selecting the largest for all further runs. We then extended our analysis to *high_res_feats[0]* and *high_res_feats[1]*, which improved GAN boundary precision (especially the latter), but underperformed on diffusion images.



To summarize our results over different decoder types and feature maps, we computed the mean IOU both ID and OOD, reporting them in this table. The best mean IOUs were achieved using the largest decoder, with *high_res_feats[1]* performing best ID and *image_embed* being best OOD.

Decoder	Features	ID	OOD
Conv-4	*image_embed*	52.93	14.77
Conv-12	*image_embed*	61.57	15.12
Conv-20	*image_embed*	63.62	14.59
Conv-20	*high_res_feats[0]*	60.84	11.01
Conv-20	*high_res_feats[1]*	67.27	14.01

We evaluated generalization with a cross-domain setup: training on three datasets and testing on the fourth, cycling through all combinations, employing *image_embed* and *high_res_feats[1]*. The former worked better on diffusion-based deepfakes, while the latter excelled on GANs.



Finally, we illustrate a few qualitative predictions from the best-performing configurations. In order: LaMa and Pluralistic (GAN-based methods) using *high_res_feats[1]* on the first row, respectively Repaint-P2-9k and LDM (diffusion-based methods) using *image_embed* features on the second one.

5. Conclusion

In summary, our model performed well across diverse deepfake generators, with *image_embed* favoring diffusion-based fakes and *high_res_feats[1]* excelling on GANs. Although cross-domain training struggled on the unseen generator, it yielded strong results on the seen ones, offering a promising direction for improving generalization.