Deepfake Localization Across Generative Models Using Deep Learning

Ciuperceanu Vlad-Mihai

Thesis Supervisor: Assoc. Prof. Cristian Rusu

Faculty of Mathematics and Computer Science
University of Bucharest

July 2025

Introduction

- **Deepfakes** have become increasingly widespread on social media in recent years, often used for *disinformation*, with a 245% worldwide increase in generated images in 2024 compared to 2023 [2].
- Our focus will be on the **deepfake localization task**, making **pixel-level** predictions and monitoring **Intersection over Union (IoU)**.
- **Applications** include fake news prevention and forensic analysis.
- Two prominent generative methods are based on Generative Adversarial Networks (**GANs**) and the state-of-the-art **diffusion** models.

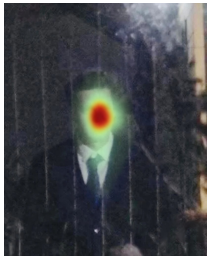


Figure: Localization example

Figure: GAN structure [3] and diffusion process [1].

Objectives and Personal Contribution

- Two **major challenges** in tackling deepfake localization remain in:
  – limited generalization to unseen, out-of-domain data
  – lack of datasets with ground-truth masks required in training
- Our **main objectives** are to explore localization performances and address the issues.
- **Personal contribution** includes:
  – *Architecture design and adaptation*: leveraging features from pretrained large models like CLIP [4] and SAM 2 [5] for localization as in DeCLIP [6].
  – *Related tasks*: showing effectiveness in tasks of detection and classification.
  – *Multi-model approach*: combining benefits in two-step models.
  – *Cross-domain setups:* analyzing knowledge transferability between domains.
  – *Limited cross-domain exposure analysis*: quantifying how samples seen from one domain influences performance on the others and proving that a limited amount of diverse samples seen can lead to strong results through complementary knowledge.

Datasets

• **Dolos** [7] is the deepfake localization dataset used in our study, focusing on robustness through multiple generative scenarios, each containing 9,000 samples of size 256x256 for training, 900 for validation and 900 for test, along with their masks.
• Generator names translated to the names of the domain-specific datasets (first two GAN-related, next two diffusion-based): **Pluralistic**, **LaMa**, **Repaint-P2-9k**, **Latent Diffusion Model (LDM)**.
• For out-of-domain (OOD) generalization, we created **custom cross-domain datasets**:

Train on 3 domains, test on 4th

Extracting a third of each dataset, we obtained 4 datasets where models were exposed to 3 domains, testing on the held-out 4th one.

train_all_4

Created similarly, using all 4 thirds chosen previously, resulting in a dataset exposing the model to all generators.

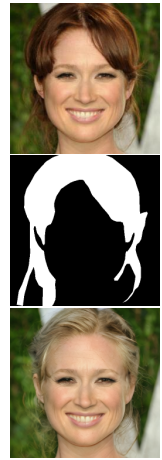
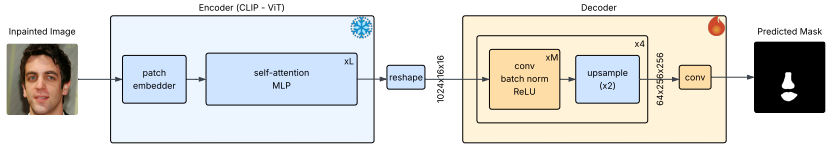
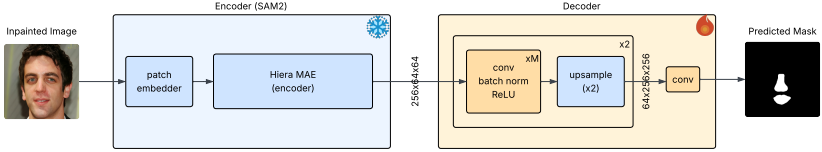


Figure: Examples of real, fake, and inpainting mask images from the dataset.

Methods



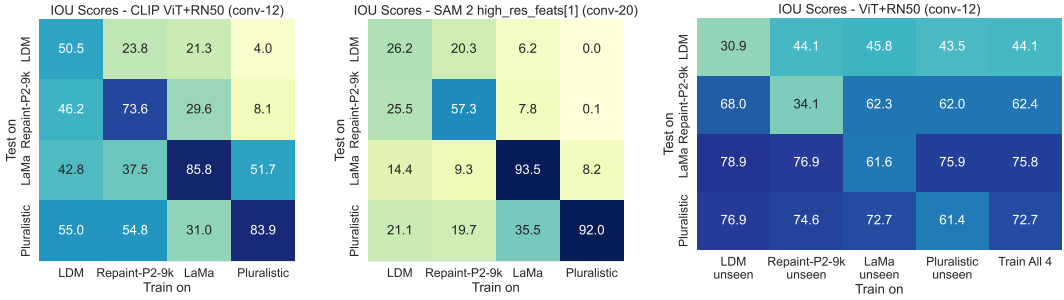
(a) DeCLIP architecture: ViT, RN50, ViT+RN50 fusion encodings



(b) SAM 2-based architecture: *image_embed*, *high_res_feats[0]*, *high_res_feats[1]* encodings

Figure: **Encoder-decoder** model architectures for localization, combining **self-supervised representations** from **frozen pretrained encoders** with a **trainable decoder**.

Results and Analysis



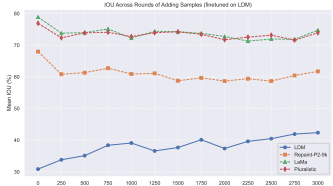
(a) CLIP ViT+RN50 (b) SAM 2 *high_res_feats[1]* (c) CLIP ViT+RN50

Figure: Results of **single-domain** (a,b) and **cross-domain** (c) training.

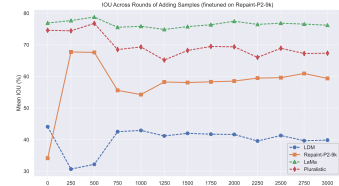
- Diffusion knowledge can be **transferred** to "simpler" domains; vice versa not much.
- These indicated **feature predilection** through a bias towards generators, while highlighting **complementarity** in representations and possible learned knowledge.
- **Near-top performance** on ViT+RN50 in a setup **seeing samples of all domains**.

Cross-domain Training - Exposure Quantitative Analysis

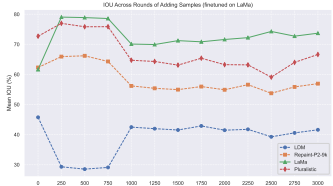
- How does the **number of samples seen** from one domain **influence performance** over the others?
- We will **fine-tune** for 10 epochs the cross-domain models over a gradually increasing number of samples from the unseen dataset.
- Increases on new domains with small drops on seen ones means that **light exposure to more domains** can greatly improve results.



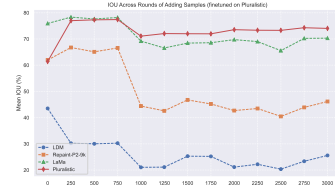
(a) Fine-tuning on LDM



(b) Fine-tuning on Repaint-P2-9k



(c) Fine-tuning on LaMa



(d) Fine-tuning on Pluralistic

Figure: Fine-tuning IoUs on test sets plotted by each round of adding batches of 250 samples until reaching a third of 3,000 images.

Cross-domain Training - Exposure Quantitative Analysis

Data seen from domains	LDM	Repaint-P2-9k	LaMa	Pluralistic
9,000 each per model	50.5	73.6	85.8	83.9
3,000 each per model	40.8	58.0	67.9	75.8
3,000 each (*train_all_4*)	44.1	62.4	75.8	72.7
1,000 LDM, 500 rest	38.7	51.5	68.8	65.4
1,750 LDM, 750 rest	41.1	53.0	71.2	68.7

Table: IoUs of CLIP-based models: considering all 4 domain-specific models, then on combined training.

- Setups of only 2,500 and 4,000 samples achieved high results over a singular model **exposed to significantly less data** (72% and 55% less than initial training setups, respectively 79% and 66% less than in *train_all_4*).
- **Similar results**, with improvements on LDM and LaMa, are seen over thirds-training.
- This implies that sample **diversity** matters more than the sheer quantity of singular, domain-specific data, **tackling both main issues of this task**.

Two-Step Model

Model Configuration	LDM	Repaint-P2-9k	LaMa	Pluralistic
Generalist	44.1 / —	62.4 / —	75.8 / —	72.7 / —
Detector + Generalist	— / 63.1	— / 62.2	— / 78.6	— / 74.2
Classifier + 4 Specialists	49.7 / 72.6	33.4 / 64.5	91.5 / 93.5	75.4 / 85.5

Table: IoU testing Two-Step pipelines, compared to the generalist model. Results are presented both on just the domains and after adding the test set of real samples.

- *First Step* (*detector*/*classifier*) + *Second Step* (*generalist*/*specialists*) pipeline.
- For **detection/classification**, we trained a *linear layer* over the ViT tokens.
- We employed four **"specialist"** models for best performances: ViT+RN50 over diffusion and SAM 2 *high_res_feats[1]* over GANs, with a classifier as the first step.
- Thus, we **compared four specialists with one generalist model**.
- Metric increases in second values appear due to the **correct real-image predictions**.
- Drops in IoU are also attributed to *misclassifications*, resulting in OOD testing.

Qualitative Results

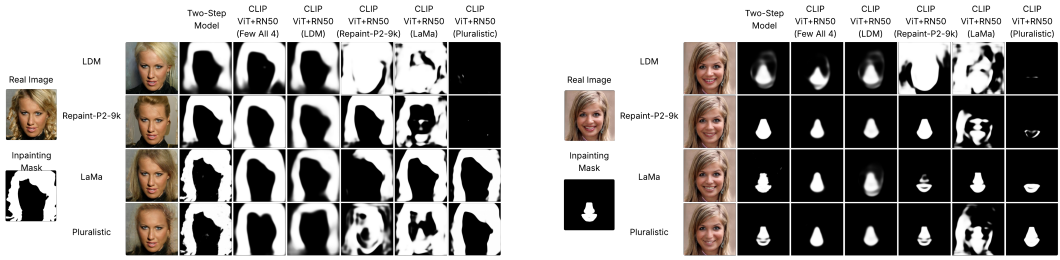


Figure: Qualitative results across domains (demo illustration).

- Model **confidence** visually translates to higher intensity pixels through bigger probabilities in strong white regions, compared to less certain, blurry areas.
- Predictions are **sharper** in GAN-related scenarios than in diffusion, reflecting performance differences.

Conclusion

- **Mean IoU across domains** of 63.7% was obtained on *train_all_4* (56.1% on smallest subset) and 62.5% on Two-Step Model, as the single-domain best was of 48.6%.
- Overall, a **Two-Step Model leads to the best results** if all its components are optimized. This implies a heavy system of multiple models, each trained on a significant amount of samples from various domains.
- A much **lighter alternative is a generalist model**, trained on limited diverse samples, suitable for common real-world scenarios, with less labeled data available.
- **Future work** could involve more generators to conduct similar experiments, possibly employing larger models. Better classifiers in Two-Step pipelines, meta-models and feature fusions should be investigated.

Thank you for your attention!

References

[1] *CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications*. URL: https://cvpr2022-tutorial-diffusion-models.github.io/. accessed: 01.07.2025.

[2] *Deepfake Cases Surge in Countries Holding 2024 Elections, Sumsub Research Shows*. URL: https://sumsub.com/newsroom/deepfake-cases-surge-in-countries-holding-2024-elections-sumsub-research-shows/. accessed: 10.06.2025.

[3] *Google Machine Learning Advanced Courses GAN*. URL: https://developers.google.com/machine-learning/gan. accessed: 01.07.2025.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763. arXiv: 2103.00020 [cs.CV].

[5] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. *Sam 2: Segment anything in images and videos*. 2024. arXiv: 2408.00714 [cs.CV].

[6] Stefan Smeu, Elisabeta Oneata, and Dan Oneata. "DeCLIP: Decoding CLIP representations for deepfake localization". In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2025, pp. 149–159. DOI: 10.1109/WACV61041.2025.00025.

[7] Dragoș-Constantin Țânțaru, Elisabeta Oneață, and Dan Oneață. "Weakly-supervised deepfake localization in diffusion-generated images". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 6258–6268. DOI: 10.1109/WACV57701.2024.00614.