Deepfake Image Detection

Tîrîlă Patric-Gabriel, Mircea Răzvan
University of Bucharest
patric-gabriel.tirila@s.unibuc.ro, andrei-razvan.mircea@s.unibuc.ro

1. Introduction

- **Goal:** Develop robust methods to detect deepfake images and identify their generative source.
- **Context:** Deepfakes generated by diffusion models challenge traditional detection systems.
- **Tasks:**
  - **Task 1:** Cross-generator binary classification (Real vs. Fake).
  - **Task 2:** Generator attribution (Real, LDM, LAMA, Pluralistic, Repaint).
- **Importance:** Improves trust, accountability, and deepfake forensics.

2. Dataset & Preprocessing

Real images: CelebA-HQ
Fake generators: LDM, LAMA, Pluralistic, Repaint

Preprocessing:

- Resize, normalization
- CLIP-based feature extraction

```
=== TRAIN ===
celebhq_real_data   : 9000 images
ldm                 : 9000 images
lama                : 9000 images
pluralistic         : 9000 images
repaint             : 8999 images

=== VALID ===
celebhq_real_data   : 900 images
ldm                 : 900 images
lama                : 900 images
pluralistic         : 900 images
repaint             : 900 images

=== TEST ===
celebhq_real_data   : 900 images
ldm                 : 900 images
lama                : 900 images
pluralistic         : 900 images
repaint             : 900 images
```

3.1 Cross-Generator – Method

Approach:

- Adapted DeCLIP (CLIP-RN50, CLIP-ViT/L-14)
- Binary classifier: Real vs Fake
- Compared with Universal Deepfake Detection (Ojha et al., 2023)

Conclusions:

- Most test scores on LDM are very high LDM-generated fakes are the most detectable, regardless of which generator the model was trained on.
- Most values in the LAMA column are low detection of LAMA-generated images is more difficult.
- Training on Repaint yields good results when testing on other generators models trained on it generalize well.
- On average, both backbones perform similarly, with RN50 achieving the best result (likely due to the small dataset).
- With the RN50 backbone, DeCLIP is significantly better than Universal. On ViT/L, the situation is reversed: Universal achieves slightly better results than DeCLIP.

3.2 Cross-Generator Deepfake Detection - Results

Avg. Precision - DeCLIP + RN50:

Train\Test	LDM	LAMA	Pluralistic	Repaint
LDM		0.47	0.84	0.72
LAMA	0.96		0.85	0.62
Pluralistic	0.74	0.79		0.73
Repaint	0.99	0.94	0.93	

Avg. Precision - DeCLIP + ViT/L:

Train\Test	LDM	LAMA	Pluralistic	Repaint
LDM		0.69	0.82	0.75
LAMA	0.76		0.75	0.65
Pluralistic	0.74	0.59		0.74
Repaint	0.95	0.70	0.93	

Avg. Precision - Universal + RN50:

Train\Test	LDM	LAMA	Pluralistic	Repaint
LDM		0.42	0.78	0.72
LAMA	0.92		0.79	0.56
Pluralistic	0.53	0.72		0.65
Repaint	0.96	0.71	0.82	

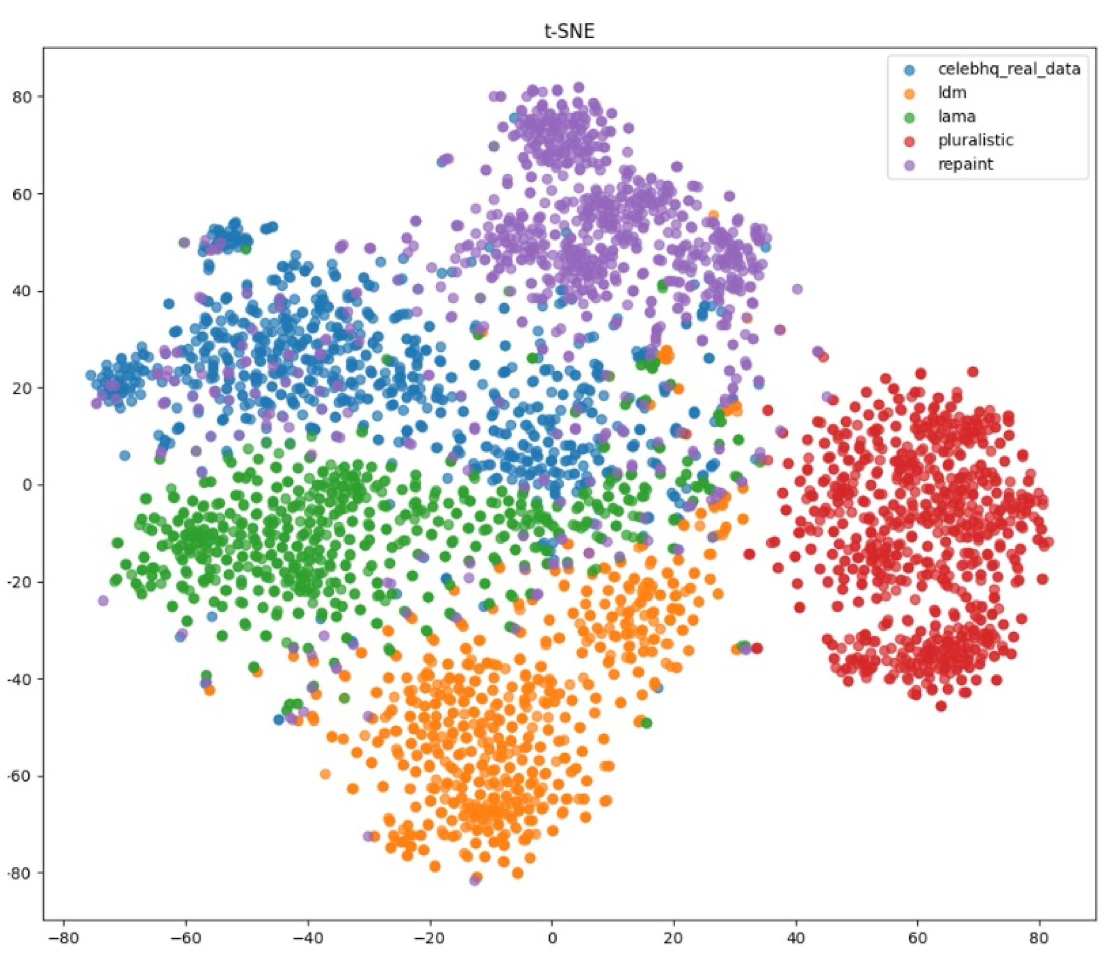
Avg. Precision - Universal + ViT/L:

Train\Test	LDM	LAMA	Pluralistic	Repaint
LDM		0.75	0.83	0.76
LAMA	0.77		0.77	0.67
Pluralistic	0.77	0.63		0.75
Repaint	0.93	0.72	0.89	

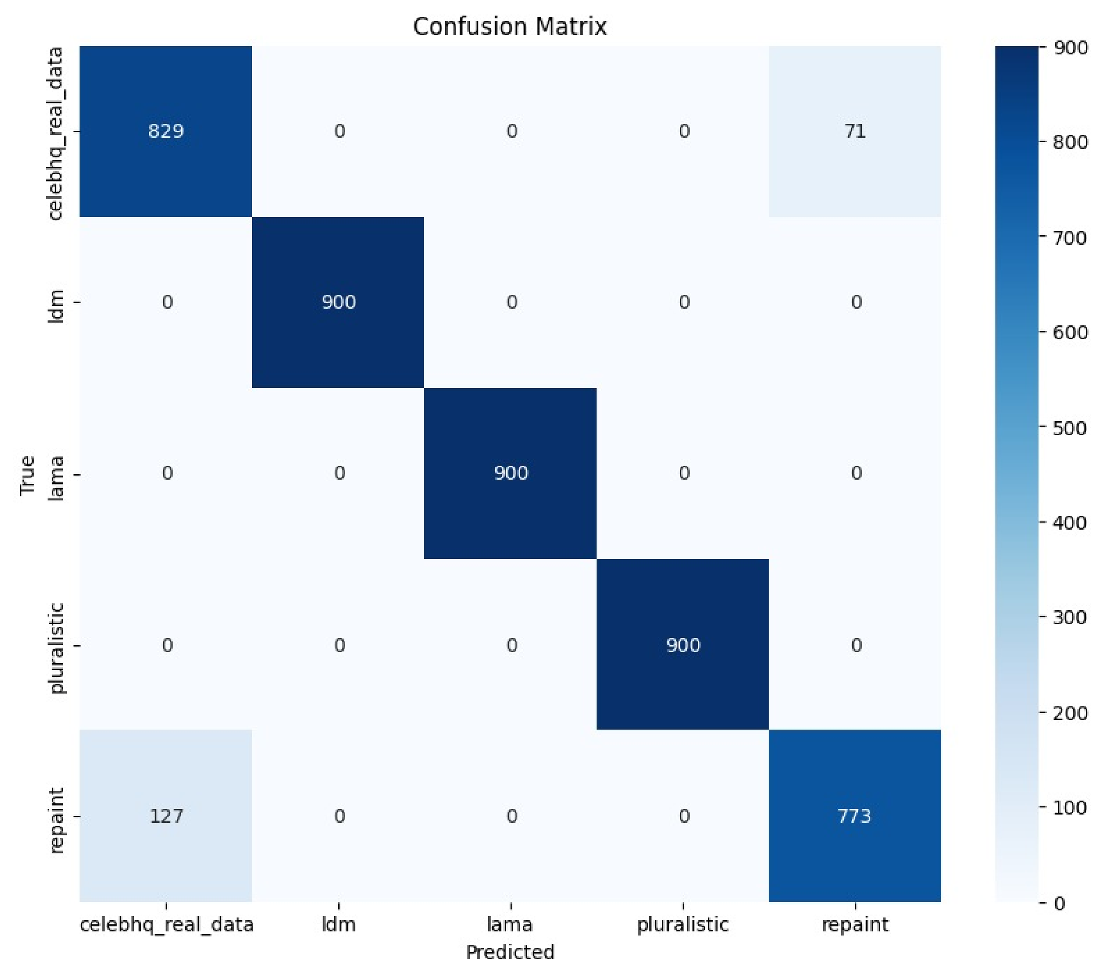
4. Model Attribution

Task: Identify the image's origin (Real, LDM, LAMA, Pluralistic, Repaint) using multiclass classification.

ViT-L/14 – Accuracy: 95.6%
Per-Class: Real 92%, LDM 100%, LAMA 100%, Pluralistic 100%, Repaint 86%
Note: Accurate but computationally expensive.

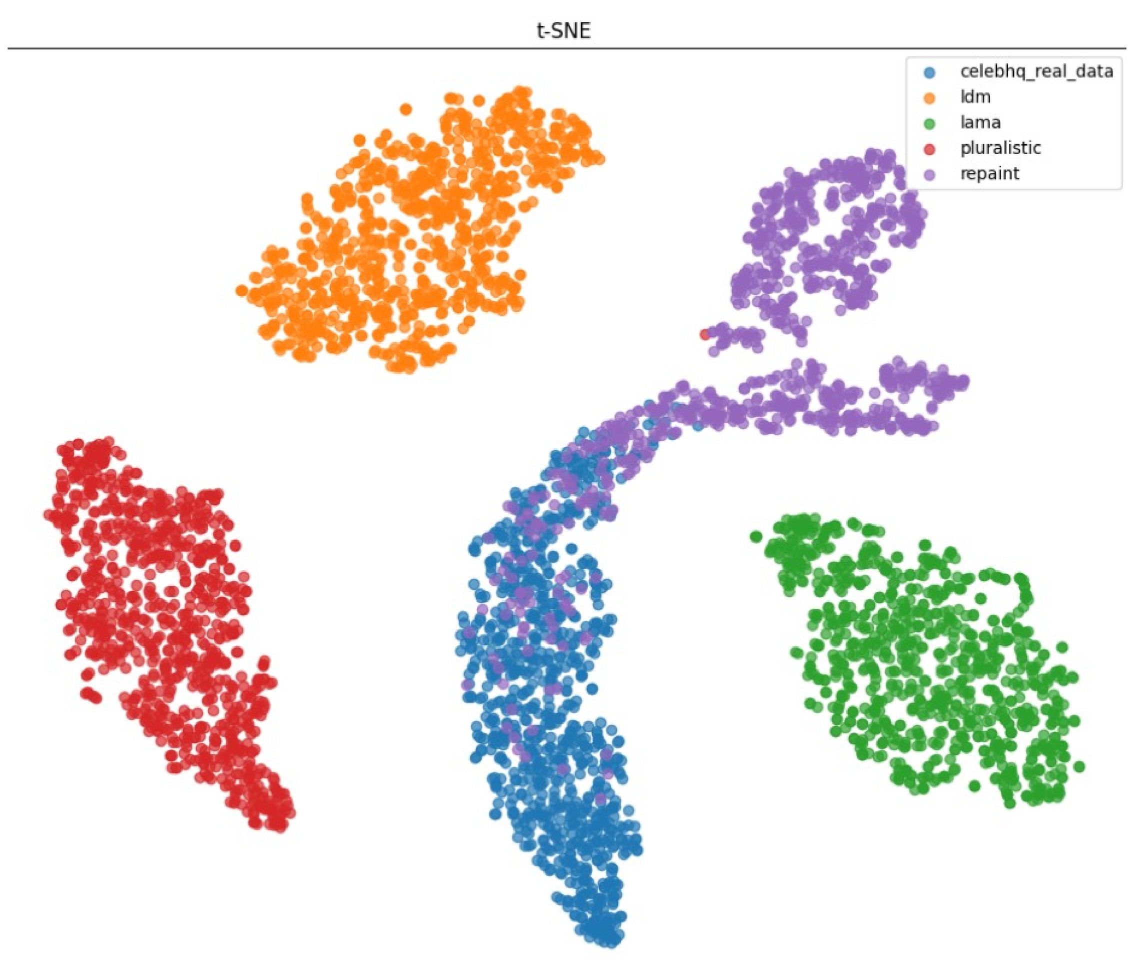


t-SNE (ViT-L/14)

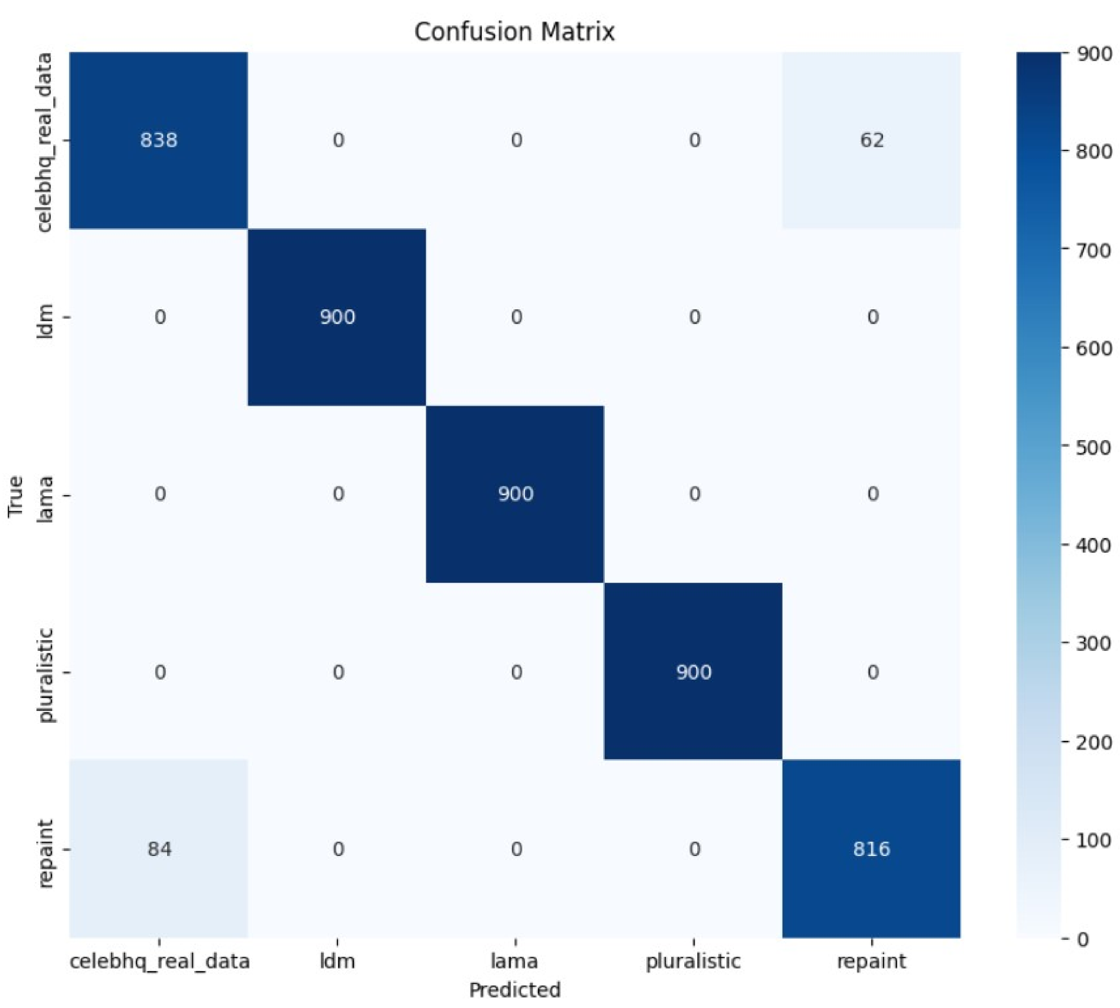


Confusion Matrix (ViT-L/14)

RN50 – Accuracy: 96.76%
Per-Class: Real 93%, LDM 100%, LAMA 100%, Pluralistic 100%, Repaint 90%
Note: Lightweight and high-performing alternative.



t-SNE (RN50)



Confusion Matrix (RN50)