Contrasting Deepfakes Diffusion via Contrastive Learning and Global-Local Similarities

Baraldi L.*[2], **Cocchi F.***[1,2], Cornia M.[1], Baraldi L.[1], Nicolosi A.[3], Cucchiara R.[1]

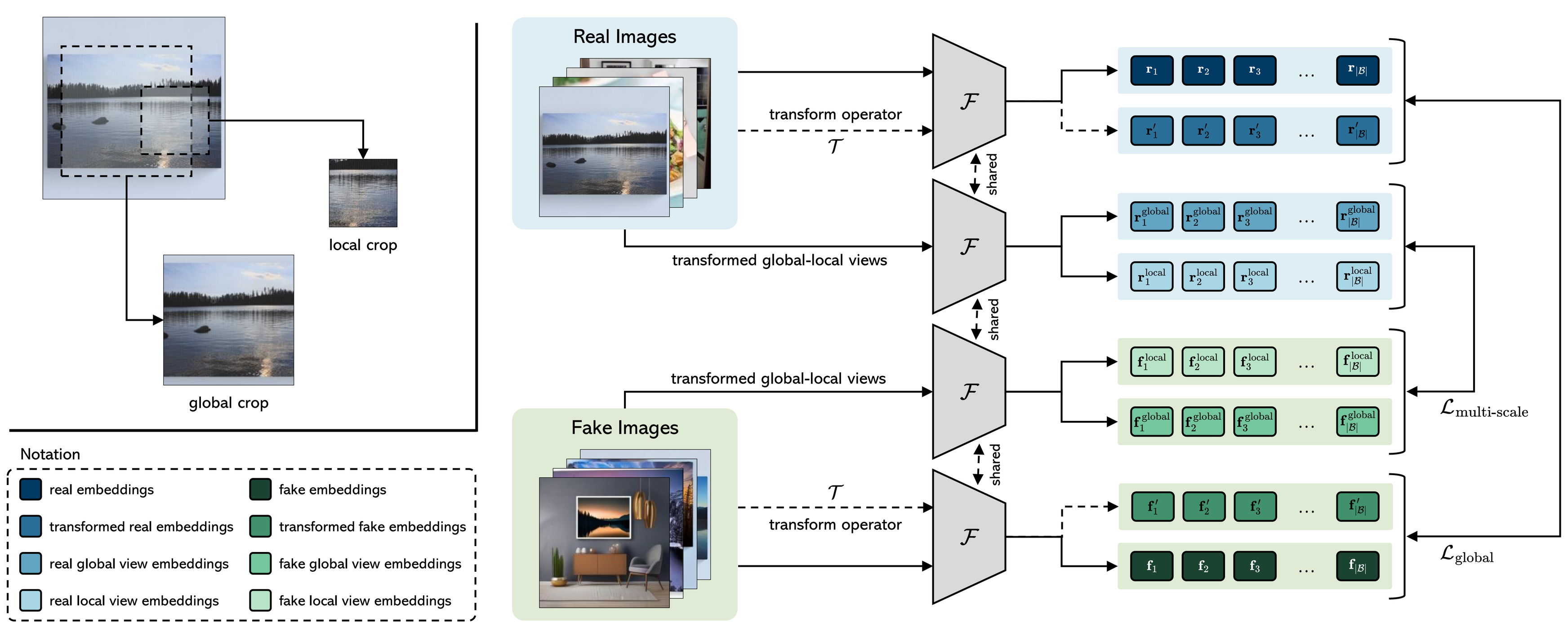
University of Modena and Reggio Emilia, Italy

[1]{name.surname}@unimore.it, [2]{name.surname}@phd.unipi.it, [3]{name.surname}@leonardo.com

Abstract

Separating authentic content and AI-generated images is increasingly difficult. Solutions using foundation models like CLIP are not ideal for deepfake detection, lacking specialized training and local image features. We propose **Co**ntrastive **D**eepfake **E**mbeddings (**CoDE**), an embedding space tailored for deepfake detection, trained via contrastive learning with global-local similarities on an in-house dataset of 9.2 million generated images.

Contrastive Deepfake Embeddings (CoDE)



- CoDE is based on a **ViT-Tiny** backbone employing only **5M parameters**.
- Training is conducted via **Info-NCE loss** [6] which is applied to both real and fake images. The global loss $\mathcal{L}_{\text{global}}$ takes into account features representing **global views** of the images. Differently, $\mathcal{L}_{\text{multi-scale}}$ enforce the similarity of features extracted from **local and global crops**.
- Robustness to post-processing techniques is enforced by applying heavy image augmentation during training to enhance robustness.

The Limitations of Foundation Models

- The embedding spaces [1, 2, 3] are not tailored for deepfake detection.
- Models are vulnerable to unseen image processing techniques as proved in [5].
- CLIP smaller backbone is ViT-B (86M parameters), limiting the portability.
- In the future foundation models could be trained on generated images too, leading to the possibility of performance degradation related to data poisoning.

Performance on Seen Generators

Model	w/o Transforms			w/ Transforms						
	Overall	Real	Fake	Overall	Real	Fake	DF-IF	SD-1.4	SD-2.1	SD-XL
Wang *et al.* (RN50 Blur+JPEG 0.5)†	20.7	99.4	1.0	20.8	99.2	1.2	0.9	1.6	1.2	1.4
Wang *et al.* (RN50 Blur+JPEG 0.1)†	21.4	98.7	2.0	21.6	98.2	2.5	2.2	2.8	2.1	2.8
Gragnaniello *et al.*†	21.8	**99.7**	2.3	21.8	**99.5**	2.3	1.4	4.2	1.5	2.1
Corvi *et al.*†	75.9	99.2	70.1	64.1	99.2	55.4	8.1	84.1	76.0	53.3
Ojha *et al.*†	31.0	96.1	14.8	37.7	87.0	25.4	11.3	24.5	19.0	46.8
Wang *et al.* (DIRE)†	79.7	10.0	97.1	76.5	15.8	91.7	89.6	92.4	91.5	93.1
ViT-T (BCE)	97.0	91.4	98.4	93.7	93.8	93.6	92.1	93.5	92.7	96.4
CoDE (Linear)	**98.0**	94.0	99.0	95.7	95.6	95.8	94.8	95.8	94.9	97.5
CoDE (NN)	97.3	89.3	**99.3**	**95.8**	90.5	**97.1**	**96.6**	**97.3**	**96.6**	98.1
CoDE (SVM)	91.3	74.4	95.4	92.5	81.0	95.4	**96.6**	91.7	94.4	**99.0**

- We combine CoDE with various classifiers, including Linear, Nearest Neighbor (NN), and One-Class SVM (SVM). These classifiers are fitted on 10k records (50000 images) of pre-processed images.
- CoDE demonstrates superior performance compared to SoTA detectors on seen generators, excelling in both transformed and non-transformed images. In this setting, CoDE achieves overall accuracies on raw images of 98%, 97.3%, and 91.3% with respectively Linear, NN, and SVM classifiers. Differently, when facing post-processed images CoDE attains accuracies of 95.7%, 95.8%, 92.5% on Linear, NN, and SVM classifiers.

Diffusion-generated Deepfake Detection (D^3) dataset

Existing datasets for deepfake detection suffer from limited generator diversity and insufficient image quantities. To address this, we have introduced the **D**iffusion-generated **D**eepfake **D**etection (**D^3**) dataset, comprising **11.5 million images**.

- Every entry in the dataset includes a prompt, an authentic image, and four images produced by four SoTA diffusion generators.
- Prompts and corresponding real images are taken from LAION-400M [4], while fake images are generated, starting from prompts. To diversify the dataset, images are generated with various aspect ratios, and different encoding and compression methods are used, closely aligning with the encoding distribution of LAION.

Real DF-IF SD-1.4 SD-2.1 SD-XL

`Soft top Jeep CJ5 convertible Vinyl 19551975`

Real DF-IF SD-1.4 SD-2.1 SD-XL

`Christ Church College`

Performance on Unseen Generators

Model	Guided	LDM			GLIDE			DALL-E			Midjourney	Avg
		200	200 (CFG)	100	100 (27)	50 (27)	100 (10)	v1	v2	v3		
Wang *et al.* (RN50 0.5)†	52.3	51.1	51.4	51.3	53.3	55.6	54.3	52.5	50.9	49.8	50.1	52.4
Wang *et al.* (RN50 0.1)†	62.0	53.9	55.3	55.1	60.3	62.7	61.0	56.1	66.2	50.2	52.2	57.7
Gragnaniello *et al.* †	54.1	58.0	61.1	57.5	56.9	59.6	58.8	71.7	57.1	50.1	50.9	57.8
Corvi *et al.*†	52.1	**99.3**	**99.3**	**99.3**	58.0	59.1	62.3	**89.4**	49.6	82.9	**98.3**	77.2
Ojha *et al.*†	**69.5**	94.4	74.0	95.0	78.5	79.1	77.9	87.3	60.1	53.5	53.9	74.8
Wang *et al.* (DIRE)†	56.7	62.6	61.3	62.2	63.2	63.4	63.1	63.0	63.4	60.7	62.3	62.0
CoDE (Linear)	53.5	92.5	95.6	91.9	71.7	75.4	72.9	63.1	71.4	86.7	84.0	78.0
CoDE (NN)	53.5	92.7	96.1	92.5	73.8	76.9	74.0	67.0	74.3	88.6	86.8	79.6
CODE (SVM)	54.6	91.0	90.4	90.9	77.2	78.8	77.6	76.1	80.2	**91.0**	89.7	**81.6**

- When detecting unseen diffusion models, not encountered during training, CoDE outperforms competitors, achieving average accuracies of 79.6% with the NN classifier and 81.6% with the SVM classifier.

Ablation Study on Training Losses

Model	w/o Transforms			w/ Transforms						
	Overall	Real	Fake	Overall	Real	Fake	DF-IF	SD-1.4	SD-2.1	SD-XL
w/ $\mathcal{L}_{\text{global}}$ only (real $\rightarrow$ fake)	87.7	74.9	90.9	83.5	75.3	85.6	80.7	84.6	86.5	90.5
w/ $\mathcal{L}_{\text{global}}$ only (pre-trained)	87.3	93.8	85.7	86.2	92.9	84.5	94.0	76.6	76.7	91.0
CoDE (Linear)	**98.0**	**94.0**	**99.0**	**95.7**	**95.6**	**95.8**	**94.8**	**95.8**	**94.9**	**97.5**
w/ $\mathcal{L}_{\text{global}}$ only (real $\rightarrow$ fake)	76.2	75.1	76.5	73.9	75.3	73.5	68.3	70.3	73.9	81.5
w/ $\mathcal{L}_{\text{global}}$ only (pre-trained)	96.1	86.8	98.4	94.2	86.5	96.2	95.3	96.0	95.4	98.0
CoDE (NN)	**97.3**	**89.3**	**99.3**	**95.8**	**90.5**	**97.1**	**96.6**	**97.3**	**96.6**	**98.1**
w/ $\mathcal{L}_{\text{global}}$ only (real $\rightarrow$ fake)	80.1	**86.7**	78.5	74.9	**86.2**	72.1	66.2	67.8	71.2	83.4
w/ $\mathcal{L}_{\text{global}}$ only (pre-trained)	89.2	46.4	**99.9**	89.1	46.0	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**
CoDE (SVM)	**91.2**	74.4	95.4	**92.5**	81.0	95.4	96.6	91.7	94.4	99.0

- CoDE, which incorporates $\mathcal{L}_{\text{multi-scale}}$, reaches better performance compared to only employing $\mathcal{L}_{\text{global}}$. Further, CoDE performs best when trained from scratch.

References

[1] Radford et al. *Learning transferable visual models from natural language supervision.* In ICML, 2021.
[2] Caron et al. *Emerging properties in self-supervised vision transformers.* In CVPR, 2021.
[3] Maxime, et al. *Dinov2: Learning robust visual features without supervision.* In arXiv preprint arXiv:2304.07193, 2023.
[4] Schuhmann et al. *Laion-400m: Open dataset of clip-filtered 400 million image-text pairs.* In NeurIPS Workshop, 2021.
[5] Cocchi et al. *Unveiling the Impact of Image Transformations on Deepfake Detection: An Experimental Analysis.* In ICIAP, 2023.
[6] Van den Oord et al. *Representation Learning with Contrastive Predictive Coding.* In NeurIPS, 2018.

Acknowledgments

elsa European Lighthouse on Secure and Safe AI

Funded by the European Union

ECCV