Exploring Self-Supervised Vision Transformers for Deepfake Detection: A Comparative Analysis

Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen

Introduction

- Most deepfake detectors use ConvNets as feature extractors.
- The community is hesitant to use large ViTs:
  - Need extensive data to train.
  - Suboptimal generalization with small- and medium-size datasets.
- Recently, ViTs trained with SSL strageries demonstrated remarkable performances:
  - Robust feature extractors → Applicable for deepfake detection?
  - Explicit semantic segmentation → Explanation & localization?
  - Strong transfer learning across various downstream tasks → Better generalizability?

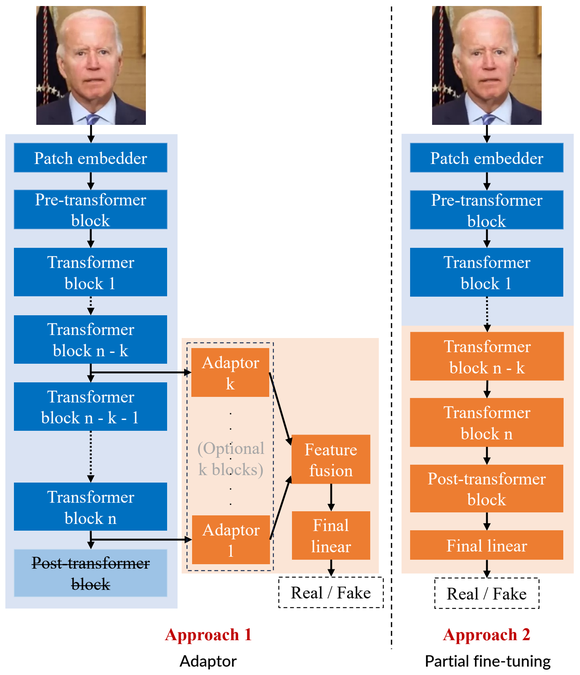
Methodology

Overview:

- Approach 1: Generalized form of the recent DFD approaches.
- Approach 2: Leveraging the strong transferability & segmentation of SSL ViTs.

Comparative study:

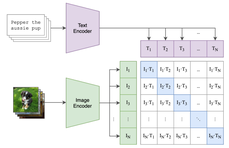
- ConvNets vs. ViTs.
- Supervised vs. unsupervised.
- Frozen backbone with adaptor vs. partial fine-tuning.



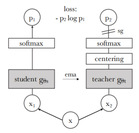
Experimental Design & Results

Backbones

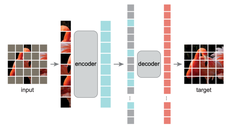
Backbone	Architecture	Way of training	Dataset(s)	Images	Annotations
EfficientNetV2 Large	ConvNet	Supervised	ImageNet-21K	14M	Image classes
DeiT III L/16-LayerScale	Transformer	Supervised	ImageNet-21K	14M	Image classes
EVA-02-CLIP-L/14	Transformer	Supervised	LAION-2B & COYO-700M	2B	Image-text pairs
MAE ViT-L/16	Transformer	Self-supervised	ImageNet-1K	1.3M	Not used
DINO (various versions)	Transformer	Self-supervised	ImageNet-21K	14M	Not used
DINOv2 (various versions)	Transformer	Self-supervised	LVD-142M	142M	Not used



CLIP:
Constrastive image-text pairs
(Radford et al. 2021)



DINO:
Self-distilation with no labels
(Caron et al. 2021)



MAE:
Predicting masked patches
(He et al. 2022)

Datasets

Seen dataset: A mixture of

- **2** real datasets & real parts of DF datasets.
- **6** popular GAN-based deepfake datasets.

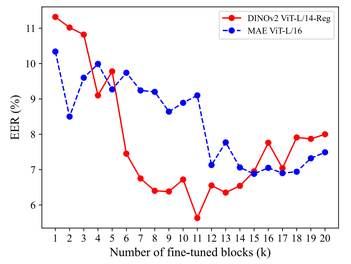
Unseen dataset: Diffusion-based deepfakes dataset (Ţânţaru *et al.*)

Size & structure:

Type	Real	Fake	Total
Training	44,037	55,963	100,000
Validation	13,200	13,000	26,200
Test	10,000	11,000	21,000
Validation (unseen)	1,900	10,700	12,600
Test (unseen)	900	3,600	4,500

Results on seen dataset

- Performances on some real sets are poor because of false positives:
  - Distributions of real videos are diverse.
  - Low-quality deepfake videos → Artifacts are destroyed → Decision boundary moved.
- Approach 2 (partial fine-tuning) is better than approach 1 (frozen backbone with adaptor).
- SSL ViTs > Supervised ViTs > Supervised ConvNets.
- Fine-tuning about ½ of DINOv2 gives the optimal performances.



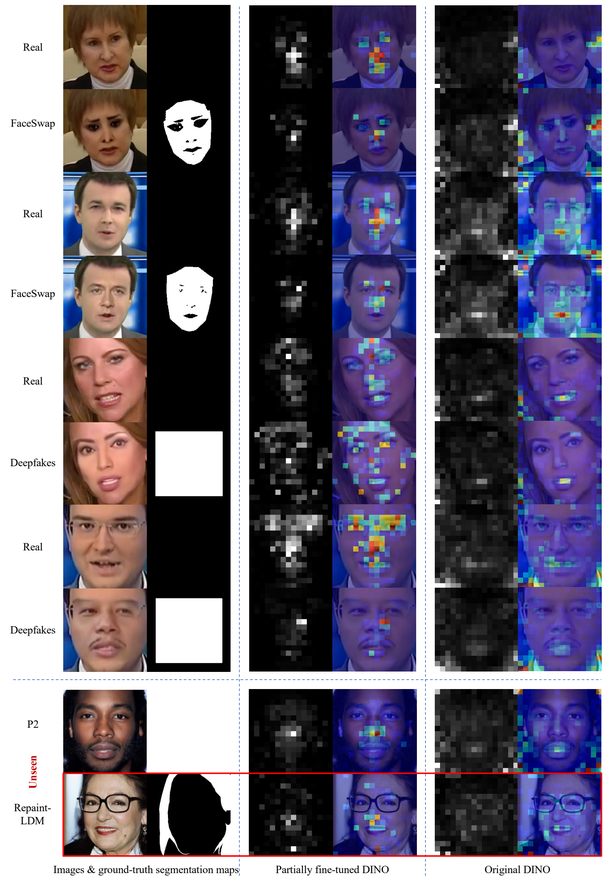
Model	k	FF++ Real	FF++ DF	FF++ F2F	FF++ FS	FF++ NT	FF++ FSh	DFD Real	DFD Fake	Vid-TIMIT	DF-TIMIT	Vox-Celeb2	YT-DF	DFDC Real	DFDC Fake	GANs	Acc.	EER	HTER
Approach 1																			
EfficientNetV2 Large	4	60.90	84.70	83.40	79.50	75.30	83.50	61.00	85.50	59.75	67.90	92.62	67.40	54.00	71.60	95.55	78.50	21.71	21.62
DeiT III L/16-LayerScale	4	38.80	88.00	87.10	81.60	82.20	82.70	37.60	92.00	93.45	39.90	99.68	53.10	57.20	62.20	95.70	79.96	19.77	19.96
EVA-02-CLIP-L/14	4	50.40	97.40	90.80	93.20	80.80	90.30	44.50	97.20	83.30	66.40	99.82	31.90	57.10	83.90	99.90	83.30	16.51	16.77
MAE ViT-L/16	4	47.20	97.50	93.70	94.20	91.00	97.70	54.20	98.40	91.75	71.90	99.70	85.00	53.70	82.90	99.90	88.06	**11.90**	**12.14**
DINOv2 ViT-L/14-Reg	4	72.40	94.60	87.00	89.30	76.50	89.40	69.40	98.30	96.80	46.20	99.70	67.30	77.70	76.00	99.85	87.42	11.98	12.41
Approach 2																			
EfficientNetV2 Large	1	65.00	94.60	89.00	89.30	84.10	90.50	22.90	91.50	99.20	60.40	99.16	62.50	72.40	68.90	97.25	84.75	15.05	15.22
DeiT III L/16-LayerScale	1	56.90	97.00	89.50	89.90	84.10	86.50	33.40	95.00	96.35	18.60	99.08	56.50	66.40	79.00	97.25	82.64	17.21	17.28
EVA-02-CLIP-L/14	1	53.50	97.50	92.70	92.40	83.30	92.40	62.40	98.80	96.55	62.90	99.26	82.20	64.00	84.60	99.00	88.29	11.77	11.77
MAE ViT-L/16	1	68.60	98.10	91.70	94.20	85.70	91.90	64.50	97.20	96.00	71.30	99.70	75.10	72.30	84.70	98.45	89.65	**10.34**	**10.35**
DINOv2 ViT-L/14-Reg	1	75.60	97.20	92.80	94.80	81.60	93.40	30.80	99.60	99.75	62.70	99.74	76.90	74.90	86.10	99.85	88.78	11.32	11.26
MAE ViT-L/16	15	80.50	99.20	94.10	93.50	90.30	95.80	77.60	99.70	99.70	77.30	99.92	85.10	81.50	84.20	99.25	93.16	6.88	6.81
DINOv2 ViT-L/14-Reg	11	85.10	98.60	94.30	95.40	89.70	97.20	67.50	99.60	99.85	92.40	99.86	86.80	88.10	89.00	99.65	94.38	**5.63**	**5.64**

Results on unseen dataset

Model	k	Threshold	Real	Repaint P2	Repaint LDM	LaMa	Pluralistic	Acc.	TPR	TNR	EER	HTER
Approach 1												
EfficientNetV2 Large	4	0.6355	47.89	52.89	55.78	49.11	56.67	52.47	47.89	53.61	49.22	49.25
DeiT III L/16-LayerScale	4	0.9983	52.89	52.67	56.44	44.44	52.67	51.82	52.89	51.56	47.83	47.78
EVA-02-CLIP-L/14	4	0.0737	63.44	43.00	52.44	31.22	51.22	48.27	63.44	44.47	45.75	46.04
MAE ViT-L/16	4	0.9385	56.00	53.22	62.78	19.33	64.67	51.20	56.00	50.00	47.31	47.00
DINOv2 ViT-L/14-Reg	4	0.0759	60.44	53.00	66.89	43.11	70.89	58.87	60.44	58.47	**40.67**	**40.54**
Approach 2												
EfficientNetV2 Large	1	0.5479	63.00	50.22	53.89	65.78	64.89	59.56	63.00	58.69	39.58	39.15
DeiT III L/16-LayerScale	1	0.9999	56.00	58.56	69.56	39.56	69.56	58.64	56.00	59.31	42.56	42.35
EVA-02-CLIP-L/14	1	0.9999	45.44	71.11	83.44	12.22	82.11	58.87	45.44	62.22	45.20	46.17
MAE ViT-L/16	1	0.1769	65.44	47.11	58.22	13.67	71.00	51.09	65.44	47.50	44.22	43.53
DINOv2 ViT-L/14-Reg	1	0.9980	50.78	70.22	78.22	65.00	86.78	70.20	50.78	75.06	**36.28**	**37.08**
MAE ViT-L/16	15	0.8948	69.89	50.78	68.89	22.56	76.44	57.71	69.89	54.67	37.56	37.72
DINOv2 ViT-L/14-Reg	11	0.7418	70.22	53.22	73.22	93.00	74.56	72.84	70.22	73.50	**27.61**	**28.14**

- Most of the behaviors are similar to those on the seen dataset.
- DINOv2 is the **absolute winner**
  - SSL DINO: employs strong data augmentations → More robust to unseen deepfake.
  - SSL MAE: use little or no data augmentations → Less robust.
- However, generalizability is still a challenging problem!

Visualization



Conclusion & Future Work

Conclusion:

- Backbone: SSL ViTs are better than supervised ConvNets
- An appropriate SSL training strategy is needed (DINOv2 is better than MAE).
- Partial fine-tuning the backbone is better than using adaptors with a frozen backbone.

Future work:

- Improve deepfake localization with the self-attention mechanism (without grond truth).
- Explore SSL on unlabeled deepfake datasets.
- Improve generalizability.

Inter-University Research Institute Corporation
Research Organization of Information and Systems
National Institute of Informatics

nhhuy@nii.ac.jp

honghuy127.github.io

IEEE International Joint Conference on Biometrics