Supplementary material – Leveraging Representations from Intermediate Encoder-blocks for Synthetic Image Detection

Christos Koutlis and Symeon Papadopoulos

Feature fusion using Feature Pyramid Networks (FPN) [1]. Although RINE's novelty lies in the use of *intermediate* CLIP features, rather than the (intentionally simple) fusion mechanism, we perform two experiments to assess the effectiveness of more complex fusion mechanisms such as that of Feature Pyramid Networks. In the first experiment ("CLIP w/ FPN" in Table 1) we replace $\mathfrak{Q}_1$, $\mathfrak{Q}_2$, and TIE modules by an FPN. In the second ("RN50 w/ FPN" in Table 1) we train a ResNet50 (pre-trained on ImageNet) with FPN. Incorporating FPN increases GPU memory consumption from 7GB to 28GB, and training time from 8min to 36min (1 epoch). Results are a lot worse than RINE in terms of ACC (69.6 vs. 91.5) and a little worse in terms of AP (97.1 vs. 98.8). Training of ResNet50 with FPN converges after 10 epochs. It consumes 5GB of GPU memory during training and needs 3min/epoch (31min in total). However, it still results in worse performance than RINE.

Fair comparison with UFD [2]. One could argue that the performance gain of RINE compared to UFD may be questionable, since training with more classes of ProGAN-generated images may make the model overfit to GAN-generated images and hurt its generalization capabilities. Thus, we additionally trained RINE on 20 classes (kept the 4-class configuration; performed no tuning), and UFD on 4 classes. The results are provided in Table 1 at lines "RINE 20-class" and "UFD 4-class". RINE 20-class roughly preserves its 4-class instance performance (90.7 ACC, 98.1 AP) without any further tuning. No overfitting on GANs is observed. The frozen CLIP features are likely robust enough to prevent a lightweight MLP from overfitting. UFD maintains its performance in the 4-class setting as well, exhibiting no significant performance increase.

Considering simpler backbones. In order to assess the importance of the backbone choice (CLIP in RINE's case) we perform two experiments using the ImageNet-pretrained ViT and Wang's detector as backbones, respectively. Table 1 presents the results at lines "RINE w/ ViT" and "RINE w/ Wang". The performance significantly decreases with these backbone choices.

	Generative Adversarial Networks								Low level vision		Perceptual loss			Latent Diffusion			Glide				
description	Pro-GAN	Style-GAN	Style-GAN2	Big-GAN	Cycle-GAN	Star-GAN	Gau-GAN	Deep-fake	SITD	SAN	CRN	IMLE	Guided	200 steps	200 CFG	100 steps	100 27	50 27	100 10	DALL-E	AVG
										ACC											
CLIP w/ FPN	99.1	81.8	75.0	90.0	86.1	68.4	99.3	55.1	72.8	50.7	52.6	60.7	53.1	77.6	54.8	78.9	58.2	58.4	57.6	61.9	69.6
RN50 w/ FPN	89.7	69.3	63.7	58.4	79.9	62.3	77.1	52.2	70.3	60.3	80.0	90.4	61.8	63.9	53.5	64.9	64.5	65.0	64.2	52.8	67.2
RINE 20-class	100.0	90.9	94.2	99.4	99.3	99.6	99.7	72.4	92.8	60.3	92.7	96.9	73.8	98.5	94.7	98.7	82.5	88.5	84.7	95.3	90.7
UFD 4-class	99.6	82.0	72.7	94.8	99.1	96.4	99.4	71.4	63.6	56.2	65.0	82.6	72.0	94.6	73.4	94.9	75.5	75.9	74.8	87.4	81.6
RINE w/ ViT	81.0	55.4	58.1	64.4	73.2	62.9	66.9	56.0	62.2	49.8	66.5	67.5	54.9	56.0	47.3	56.3	55.5	55.7	57.0	54.9	60.1
RINE w/ Wang	85.6	63.7	60.0	58.8	77.5	62.5	73.5	52.3	76.1	56.2	61.6	79.2	59.4	65.4	54.0	65.6	63.6	63.2	62.8	52.8	64.7
RINE 4-cl. (ours)	100.0	88.9	94.5	99.6	99.3	99.5	99.8	80.6	90.6	68.3	89.2	90.6	76.1	98.3	88.2	98.6	88.9	92.6	90.7	95.0	91.5
										AP											
CLIP w/ FPN	100.0	99.4	98.6	99.8	99.3	99.7	100.0	93.8	84.7	85.7	99.1	99.9	90.2	99.8	97.1	99.9	98.9	98.8	98.6	98.5	97.1
RN50 w/ FPN	96.8	79.6	73.6	65.1	85.5	87.3	88.1	56.8	82.2	70.8	95.2	98.1	69.6	73.1	55.8	74.1	70.9	72.5	71.7	54.2	76.0
RINE 20-class	100.0	99.8	100.0	100.0	100.0	100.0	100.0	97.5	97.8	84.2	98.1	99.9	95.4	99.9	99.3	99.9	96.7	98.1	96.9	99.4	98.1
UFD 4-class	100.0	96.7	98.7	99.3	99.9	99.7	100.0	85.5	64.0	76.7	94.6	99.0	88.0	99.4	92.3	99.3	94.1	94.5	93.8	97.7	93.6
RINE w/ ViT	91.0	57.6	64.0	71.4	83.3	71.8	76.2	59.4	77.1	52.2	68.8	72.1	56.9	57.5	45.7	58.3	55.1	55.7	56.4	54.5	64.3
RINE w/ Wang	94.5	76.8	73.6	66.3	85.1	87.1	85.6	58.2	80.3	60.0	87.2	96.3	65.9	77.9	56.7	77.4	76.3	77.5	75.6	57.0	75.8
RINE 4-cl. (ours)	100.0	99.4	100.0	99.9	100.0	100.0	100.0	97.9	97.2	94.9	97.3	99.7	96.4	99.8	98.3	99.9	98.8	99.3	98.9	99.3	98.8

Table 1: Further experimental results. Accuracy (ACC) and average precision (AP) are reported.

References

[1] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)

[2] Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24480–24489 (2023)