Supplementary - Beyond Deepfake vs Real: Facial Deepfake Detection in the Open-Set Paradigm

Thiru Thillai Nadarasar Bahavan[1], Sachith Seneviratne[1], Sanjay Saha[2],
Ken Chen[1], Sanka Rasnayaka[2], Saman Halgamuge[1]
[1]The University of Melbourne, Parkville [2]National University of Singapore
{bahavant, ken.chen2}@student.unimelb.edu.au, {sachith.seneviratne, saman}@unimelb.edu.au
sanka@nus.edu.sg, contact.sanjaysaha@gmail.com

1. Explainability

The proposed method detects unknown samples by thresholding the model's softmax outputs, rejecting samples with low output scores. In essence, this approach identifies unknowns by recognizing the absence of features required to classify a test sample into any known class, rather than detecting novel features within the image.

This implies that when the model identifies a known facial forgery, it focuses on a single strong discriminative feature. In contrast, when the model encounters an unknown facial forgery, it relies on a broader set of weaker features. By visualizing the class activation maps, we can observe this behavior [3]. For known facial forgeries, the class activations are more concentrated, whereas for unknown facial forgeries, the activations are more diffuse and spread out.

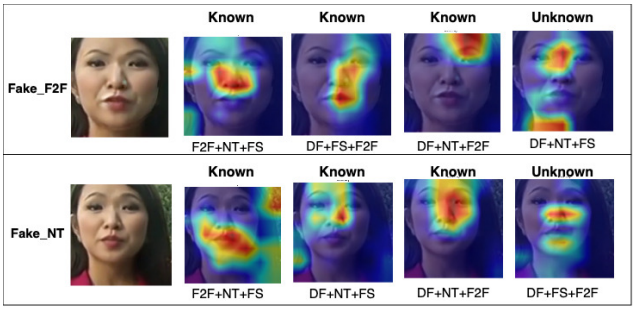

Figure 1. Class activation maps produced for images created by known and unknown facial forgeries for two different cases of unknown test classes. Notably, the images with unknown facial forgeries show multiple clusters of activations.

2. Extended Ablation Studies

2.1. Impact of Pairwise similarity scaling factor in Stage 1

We modify the supervised contrastive loss to focus on improving the clustering of real data samples, which often fail to form compact representations in the feature space as introduced in the methods section. By introducing a pairwise similarity scaling factor, we prioritize real data pairs, encouraging tighter clustering and enhancing the representation learning of real samples. The alpha parameter controls the influence of the real data samples in the loss function. In this section, we explore the effects of various values of the alpha parameter on the overall performance.

When alpha is 0, the algorithm becomes conventional supervised contrastive learning. Table 1 illustrates the effect of the alpha factor. As observed, as alpha increases from zero, we get improvements in the performance: however, it degrades for larger values of alpha. This is potentially due to an overemphasis on real data pairs, which may reduce the model's ability to capture the underlying differences between manipulated samples, leading to less effective generalization. Therefore, careful tuning of the alpha parameter is essential to balance the influence of real data samples and manipulation-related variations.

Table 1. Unknown class detection evaluated using frame-level AUROC for varying alpha values. The best results are highlighted in bold.

Unknown Class	1	1.21	2.25	4
DF	0.7009	**0.7189**	0.7119	0.7076
F2F	0.7129	**0.7399**	0.7219	0.7126
FS	0.6113	**0.6781**	0.6610	0.6580
NT	0.7877	**0.8233**	0.8100	0.7926

2.2. Impact of Labeling Scheme in Stage 1

Table 2 compares the impact of two labeling schemes on model performance during Stage 01 training:

- **Scheme 1:** Samples are labeled based on the specific forgery method used (e.g., FaceSwap (FS), DeepFake (DF), Face2Face (F2F), and NeuralTextures (NT)).
- **Scheme 2:** Samples are labeled only as real or fake.

As shown in Table 2, using a fine-grained, forgery-specific labeling scheme (Scheme 1) significantly improves the model's ability to detect unknown forgeries. This improvement arises because Scheme 1 enables the model to learn highly discriminative, method-specific features unique to each forgery technique, with minimal overlap between features of different methods. In contrast, Scheme 2 captures only general discriminative features shared across all forgery methods, which limits its ability to generalize effectively. The granularity of Scheme 1's feature representation equips the model to better recognize distinctive manipulation patterns, enhancing its ability to classify previously unseen forgery methods as unknown.

Table 2. Unknown class detection evaluated using frame-level AUROC. Each row represents a model trained on known classes and tested on unknowns. Scheme 1 uses forgery-specific labels, while Scheme 2 indicates whether an image is fake or real. The best results are highlighted in bold.

Unknown Class	Scheme 1	Scheme 2
DF	**0.7189**	0.5841
F2F	**0.7399**	0.6784
FS	**0.6781**	0.5672
NT	**0.8233**	0.5432

2.3. Impact of Labelling Scheme in Stage 3

Open-set detection is a rapidly evolving field. Many existing approaches, such as OpenMax [1] and Membership Loss [2], depend on additional unknown data for model fine-tuning. In contrast, softmax thresholding delivers robust performance while maintaining results comparable to closed-set classifiers for known classes.

Table 3 compares the impact of two labeling schemes on model performance during Stage 02 training:

- **Scheme 1:** Samples are labeled based on the specific forgery method used (e.g., FaceSwap (FS), DeepFake (DF), Face2Face (F2F), and NeuralTextures (NT)).
- **Scheme 2:** Samples are labeled only as real or fake.

As observed in Table 3, giving fine-grained forgery-specific information leads to better performance in unknown class detection.

Table 3. Unknown class detection using the frame-level AUROC metric. The row is a model trained on the training classes and then tested on the unknown class. Scheme 1 refers to forgery-specific labels. Scheme 2 identifies whether the image is real or fake. Higher values of AUROC are optimal.

Unknown Class	Scheme 1	Scheme 2
DF	**0.7189**	0.6841
F2F	**0.7399**	0.6784
FS	**0.5672**	0.511
NT	**0.8233**	0.7432

References

[1] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1563–1572, 2016. 2

[2] Pramuditha Perera and Vishal M Patel. Deep transfer learning for multiple class novelty detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11544–11552, 2019. 2

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 1