SeeABLE: Soft Discrepancies and Bounded Contrastive Learning for Exposing Deepfakes – Supplementary Material

Nicolas Larue[1,2], Ngoc-Son Vu[1], Vitomir Struc[2], Peter Peer[2], Vassilis Christophides[1]
[1]ETIS - CY Cergy Paris University, ENSEA, CNRS, France
[2]University of Ljubljana, Slovenia
nicolas.larue@ensea.fr

Abstract

In the main part of the paper, we introduced SeeABLE, a novel state-of-the-art deepfake detector learned in a one-class learning setting. We evaluated the proposed detector in rigorous experiments in cross-dataset and cross-manipulation scenarios over multiple datasets and in comparison to 12 state-of-the-art competitors, and presented a number of ablation studies to demonstrate the impact of various model components. In this supplementary material, we now provide: (i) additional technical details on SeeABLE, related to: (a) the definition of the evenly distributed prototype used with the Bounded Contrastive Regression (BCR), and (b) the geometric constraints used with the auxilary guidance loss, (ii) visual examples of the generated local image perturbations (i.e., the soft discrepancies) in the spatial and frequency domain, (iii) additional ablations, (iv) qualitative results with examples of face images generated diffusion-based (generative) models, and (v) information on the reproducibility of SeeABLE with links to relevant (open-access) repositories.

1. Hard-prototype generation

The main idea behind SeeABLE is to generate local image perturbations (soft discrepancies) and then map the different perturbations to a set of so-called *hard-prototypes* that can later be used to derive an anomaly score for deepfake detection. One of the key components in this framework are the hard-protoypes, which are defined in a way that ensures that they are evenly (i.e., equidistantly) distributed on an n-dimensional hypercube. This setup not only results in an optimal separability between the different prototypes (in terms of average between prototype distance), but also leads to highly desirable characteristics when used with contrastive learning objectives, as theoretically and empirically demonstrated in [2].

In SeeABLE, we compute the hard prototypes in accordance with the algorithm from [5]. Here, the n-dimensional prototypes $\{\boldsymbol{p}_1 \dots \boldsymbol{p}_K\}$, for $K \in [2, \dots, n{+}1]$, which serve as the targeted optimal representation of the generated soft discrepancies, are defined as vertices of a regular simplex, i.e.:

$$\boldsymbol{p}_i = \begin{cases} \frac{1}{\sqrt{n}}\mathbf{1}, & \text{if } K = 1 \\ -\frac{1+\sqrt{n+1}}{n^{\frac{3}{2}}}\mathbf{1} + \sqrt{\frac{n+1}{n}}\mathbf{e}_{i-1}, & \text{if } K \in [2, \dots, n+1] \end{cases}, \tag{1}$$

where $\mathbf{1} \in \mathbb{R}^n$ is a vector of all ones and $\mathbf{e}_{i-1}$ is a one-hot encoded vector of all zeros with a '1' at the $(i-1)^{th}$ position, and $i = 1, \dots, K$.

2. Guidance loss and geometric contraints

SeeABLE is learned by minimizing a learning objective that consists of a weighted combination of the proposed Bounded Contrastive Regression (BCR) loss and a *Guidance loss* that encourages the model to localize the generated soft-discrepancies, while taking *geometric constraints* into account. Specifically, the guidance loss, defined in Eqs. (12) and (13) of the main paper, uses a three-scale penalty definition for the learning procedure: (1) the lowest penalty of 2^{-2} is assigned if the predicted and true location of the soft-discrepancy overlap, (2) a higher penalty of 2^{-1} is assigned if the predicted and true location stem from a (horizontally) mirror-symmetric region of the face (e.g., regions 6 and 7 in Fig. 1, or regions 9 and 12, etc.), and (3) the highest penalty, proportional to the graph-distance d_{graph} between the predicted and true soft-discrepancy location is assigned in all other cases, i.e., $2^0 \times d_{graph}$.

To define the graph-based distance d_{graph} for the guidance loss, we transform the submask scheme into a graph representation, as illustrated in Fig. 1 for a 4×4 grid strategy. Here, each of the N_{loc} patches is represented by a red node, with two nodes being connected by an edge if they share a common border. An adjacency matrix $\mathbf{A}$ can then be constructed from this graph, where $\mathbf{A}[i,j] = d$ if node i is connected to node j, and $\mathbf{A}[i,j] = 0$ otherwise, with d representing the length of the edge between i and

j. However, in SeeABLE, we are not concerned with the edge length and fix it to $d = 1$. The graph-based distance d_{graph} is thus defined as the total (minimum) number of edges that need to be traversed to connect nodes i and j. We note that this definition is also applicable to other competing submask-generation schemes considered in the main part of the paper.

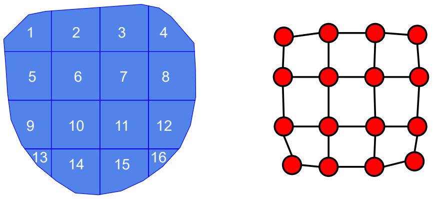

Figure 1. **Illustration of the 4×4 grid-based submask generation scheme and corresponding graph.** The graph definition on the right is used to define the graph-based distance used for the definition of the geometric constraint and respective guidance loss.

3. Examples of soft-discrepancies

To illustrate the impact of the generated soft discrepancies on the visual appearance of the perturbed faces, we show in Figure 2 a number of examples. Here, the first row presents the original images, the second row shows the locally perturbed faces and the last row shows the absolute difference between the two. Note how the soft discrepancies are hardly visible, but still allows learning a highly capable deepfake detector. In Figure 3, we present additional examples across a wider and more diverse set of images, but in addition to the real and perturbed faces and their difference, we also show the blending masks in the third row. Observe how different local areas of the face are targeted by the soft discrepancies, leading to subtle perturbations that are often imperceptible to the human visual system but detectable by SeeABLE.

4. Additional ablations

In the main part of the paper, we show several ablation studies to explore the impact of the different components of SeeABLE on the detection performance. In this section, we now add to these results with two additional ablation experiments that investigate: (i) the impact of spatial and frequency-domain perturbation on the detection task, and (ii) the effect of different grid configurations in the submask-generation scheme.

4.1. Spectral vs. spatial perturbations

Let the complete set of local perturbations, utilized to generate the soft-discrepancies for SeeABLE, be denoted as $\mathcal{P}$ and let this set consist of perturbations being applied in either the spatial or the frequency domain, i.e., $\mathcal{P} = \{\mathcal{P}_{spatial}, \mathcal{P}_{freq}\}$. The set of perturbations (transformations) used in either domain was defined in the main part of the paper in Section 4.1. In Table 8, we investigate the impact of each group of perturbations on the detection performance of SeeABLE on FF+ HQ. The AUC score is again reported as a performance indicator.

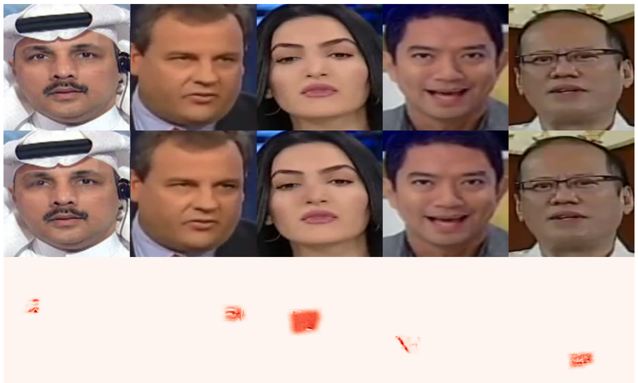

Figure 2. **Illustration of the visual impact of the generated soft discrepancies.** The first row shows examples of real faces, the second row shows the perturbed versions and the last row shows their absolute differences.

$\mathcal{P}$	Test set - AUC (in %)				
	DF	F2F	FS	NT	Avg.
$\{\mathcal{P}_{spatial}\}$	96.4	94.1	95.8	94.8	95.3
$\{\mathcal{P}_{freq}\}$	99.6	97.3	96.9	93.2	96.7
$\{\mathcal{P}_{spatial}, \mathcal{P}_{freq}\}$	**99.2**	**98.8**	**99.1**	**96.9**	**98.5**

Table 8. **Ablation results with respect to the augmentation type used.** Shown are AUC scores (in %) on the FF++ HQ dataset.

As can be seen, the results clearly show that the different perturbation types are complementary to each other. By combining spatial and frequency-domain perturbations, SeeABLE obtains better results than with either of alone. We note again at this point that the number and type of perturbations considered during training (N_{type}) defines the number of prototypes used for the regression task of SeeABLE. In turn, the results in Table 8 also illustrate the impact of changing the number of prototypes when learning the detection model.

4.2. Sensitivity to the grid size

In the main part of the paper, we showed that the grid-based strategy to submask generation yielded the best performance among the evaluated schemes. In Table 9. we now explore the impact of different configurations of this grid. Specifically, we investigate the use of 3×3, 4×4, and 5×5 grids in SeeABLE and observe strong performance across all of these configurations with the highest results

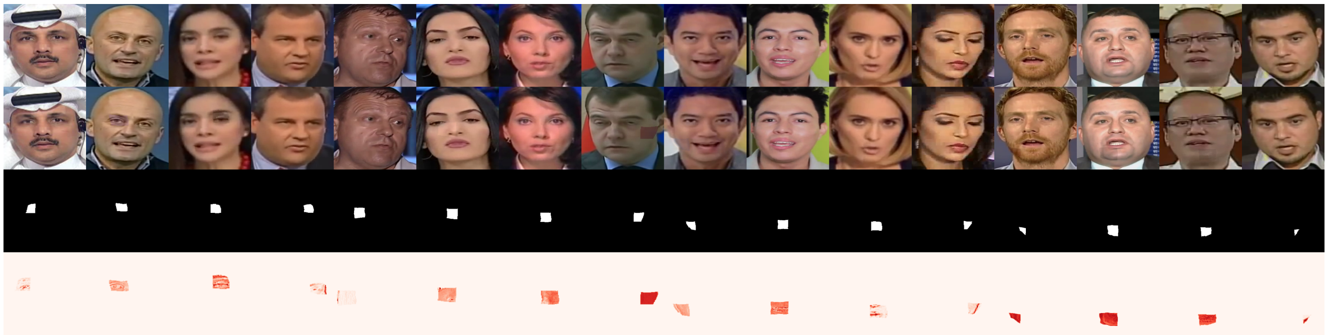

Figure 3. **Impact of the soft discrepancies on the visual appearance of the perturbed faces across a diverse set of examples.** The first and second row show the real and locally perturbed faces, respectively. The third row shows examples of the blending masks generated by the grid-based submask-generation scheme, and the last row shows the absolute differences between the initial and perturbed faces.

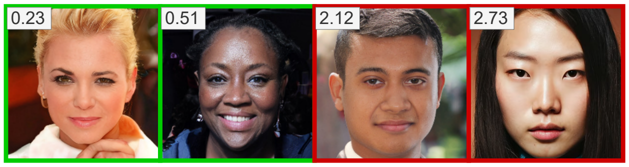

Figure 4. **Visual examples of randomly selected real (in green) and diffusion-generated faces (in red) with corresponding anomaly scores.** The first two images represent real faces from CelebA-HQ [3] and FFHQ [4], respectively, the third image was generated with latent diffusion [7] and the fourth with Midjourney [6].

observed with the 4 × 4 configuration. This particular configuration appears to offer a good trade-off between locality and visual-perturbation-impact to facilitate learning a well performing detection model. We note that the $\mathcal{A}$ notation stands for the mask criteria defined in the main part of the paper, i.e., $\mathcal{A}_1$ (full coverage), $\mathcal{A}_2$ (no overlap), $\mathcal{A}_3$ (balanced size).

	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	Avg. - AUC (in %)
$\text{SM}_{\text{Grid 3x3}}$	✓	✓	✓	74.9
$\text{SM}_{\text{Grid 4x4}}$	✓	✓	✓	**75.9**
$\text{SM}_{\text{Grid 5x5}}$	✓	✓	✓	75.1

Table 9. **Impact of different grid configurations on the performance of SeeABLE on the DFDC dataset.** Results are shown for different submask-generation strategies, with the 4 × 4 grid performing the best among the tested configurations.

5. Diffusion models

The recent proliferation of probabilistic diffusion models has led to the creation of numerous synthetic image datasets [1, 8]. However, there remains a dearth of facial datasets specifically designed for deepfake detection. Additionally, techniques employed for producing high-quality, non-existent facial images with a high degree of realism fall under the umbrella of *entire face synthesis* rather than deepfake generation. Such techniques predominantly utilized Generative Adversarial Networks (GANs), but are now increasingly adopting denoising diffusion probabilistic models for the synthesis task.

Similarly to the majority of work on deepfake detection available in the literature, our paper focuses on the detection of input-conditioned face manipulations, where local regions of the original faces are altered, and not synthesis procedures that generate entire (artificial) face images. Although the intricate details of (entire) face synthesis techniques are beyond the scope of this paper, we showcase the effectiveness of SeeABLE for the detection of synthetically generated full face images in Figure 4. Here, we apply our model to two real face images from the CelebA-HQ [3] and FFHQ [4] datasets and two diffusion- diffusion-generated images - the first generated with latent diffusion [7] and the second with Midjourney [6]. As can be seen from the reported anomaly scores in the corners of the presented images, SeeABLE ensures good separation between the real and synthetic images and produces comparably higher anomaly scores for the synthesized faces, despite the fact that it was not trained specifically for the detection of such types of data.

6. Reproducibility

We note that all of our experiments are fully reproducible. The source code, training scripts, models and learned (model) weights, associated with SeeABLE, are made publicly available here:

- SeeABLE: `https://github.com/anonymous-author-sub/seeable`

The remaining code used in the paper is also available from the official repositories, i.e.,

- Dlib:
  http://dlib.net/
- RetinaFace:
  https://github.com/deepinsight/insightface
- FaceSwap:
  www.github.com/deepfakes/faceswap

References

[1] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *arXiv preprint arXiv:2209.04747*, 2022. 3

[2] Florian Graf, Christoph D. Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In *ICML*, 2021. 1

[3] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 3

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 3

[5] Kenneth Lange and Tong Tong Wu. An mm algorithm for multicategory vertex discriminant analysis. *Journal of Computational and Graphical Statistics*, 17(3):527–544, 2008. 1

[6] Midjourney. https://www.midjourney.com/ Accessed 2023-08-03. 3

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3

[8] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022. 3