Weakly-supervised deepfake localization in diffusion-generated images
(Supplementary material)

1. Dataset details

Table 1 presents the dataset information for each of the three setups (`A`, `B`, `C`) in terms of the number of samples and their provenance for each split (train, validation, test). These details are relevant for the experiments in §5.1 and §5.2 in the main paper.

These data splits are built from real images (coming from either CelebA-HQ or FFHQ) and fake images (generated by either P2 or Repaint–P2, which were trained on either CelebA-HQ or FFHQ). For the weakly-supervised scenarios (`A` and `B`) we train on 9k real and 9k fake images, the fake images being generated by P2. For the fully-supervised scenario we use the same numbers of locally-inpainted samples for both Repaint–P2/CelebaHQ and Repaint–P2/FFHQ: 30K train and 3K validation samples.

Our evaluation is always carried on data derived from CelebA-HQ and even for the detection task we use partially-manipulated images (Repaint–P2/CelebA-HQ), since our focus is weakly-supervised localization.

2. Additional qualitative results

We present additional visual results that paint a more complete image of the performance of the proposed models in different training setups. Firstly, in Figure 1 we show visual results for all three methods `Patches`, `Attention`, `GradCam`, on the three identified training scenarios: `A`, `B`, `C`. We notice that `Patches` performs the best in all setups. In Figure 2 we show additional results when using the same and different datasets for training and for testing. The level of performance degradation is larger for smaller masks.

3. Comparison to other pretrained localization models

We compare to five pretrained models for detection and fully-supervised localization (see Figure 3). MantraNet and PSCC are trained on data forged with copy-move, splicing, removal and enhancement operations. Noiseprint relies on noise-removal techniques and learns to distinguish whether the input patches come from the same source (have similar noise residual patterns). HiFi-Net and TruFor are recent approaches (CVPR'23). The former is trained on diffusion and GAN images (with local and full manipulations) to produce hierarchical attributes. The latter is an improved version of Noiseprint, which is also trained on images from more recent manipulation techniques (GAN).

Visual results in Figure 3 indicate that Noiseprint and Hi-Fi net struggle the most to recover the inpainted regions. The activations obtained with MantraNet seem reasonable, but the network lacks the confidence and hence the small numerical results under a standard threshold of 0.5. PSCC and TruFor generally seem to find the manipulated region but they tend to under or over-segment it. Similarly, `Patches` is mostly correct in localizing the fake area but lacks precision. Unlike other methods, `Patches` has only been trained to localize forgeries of faces. The competitive results obtained with `Patches` on COCO Glide dataset suggest that it is a suitable method to perform analysis in more challenging weakly-supervised scenarios.

4. Additional results with PSCC

Table 2 presents results for PSCC trained in all three scenarios. To ensure a fair comparison, we have trained the PSCC method similarly to `Patches`. In particular, (i) we have initialized the model from scratch (random weights), and (ii) for scenarios `A` and `B`, which provide only a label, we have broadcasted the label to a image-sized matrix to obtain the mask, which is needed as target. However, in the inherent noisy training setup of configuration `B`, we have observed that the model did not converge. Instead, we were able to train in this scenario by starting from the checkpoint provided by the authors. In the paper, we report results by training from scratch.

			train				valid				test loc.		test det.			
			real		fake		real		fake		fake		real		fake	
	sup.	generator	src.	num.	src.	num.	src.	num.	src.	num.	src.	num.	src.	num.	src.	num.
A	label	full	d	9k	P2/d	9k	d	900	P2/d	900	R.P2/CA	8.5k	CA	900	R.P2/CA	900
B	label	partial	d	9k	R.P2/d	9k	d	900	R.P2/d	900	R.P2/CA	8.5k	CA	900	R.P2/CA	900
C	mask	partial	N/A	N/A	R.P2/d	30k	N/A	N/A	R.P2/d	3k	R.P2/CA	8.5k	N/A	N/A	N/A	N/A

Table 1. Datasets used for each of our setups in terms of number of samples (num.) and their provenance (src.) for each of the real and fake parts as well as for each of the splits. We use d to denote one of the two datasets (CelebA-HQ or FFHQ), while R.P2 stands for Repaint–P2 and CA for CelebA-HQ. Note that the evaluation is always carried out on data derived from CelebA-HQ.

			IoU (%)		PBCA (%)	
	sup.	generator	SC	FT	SC	FT
A	label	full	10.7	6.0	71.5	79.8
B	label	partial	–	18.4	–	21.3
C	label	partial	89.0	93.9	98.8	99.5

Table 2. Localization performance by initializing the training of PSCC either from scratch (SC) or by finetuning the author's checkpoints (FT). We observe similar results for both types of initialization, with the exception of scenario B for which the model did not converge when training from scratch. The models are tested on the Repaint–P2/CelebA-HQ test set.

Figure 1. Soft localization maps produced by the three proposed approaches using different level of supervision. `Patches` can accurately detect the manipulations after having seen only fully generated fake images (scenario `A`) or locally-inpainted images with only image-level supervision (scenario `B`). Both `Attention` and `GradCam` struggle in scenarios `A` and `B`. All methods recover the manipulated region in the fully supervised scenario, `C`. This suggests that operating at a patch level is better suited for recovering local manipulations than either using a `GradCam` or `Attention`.

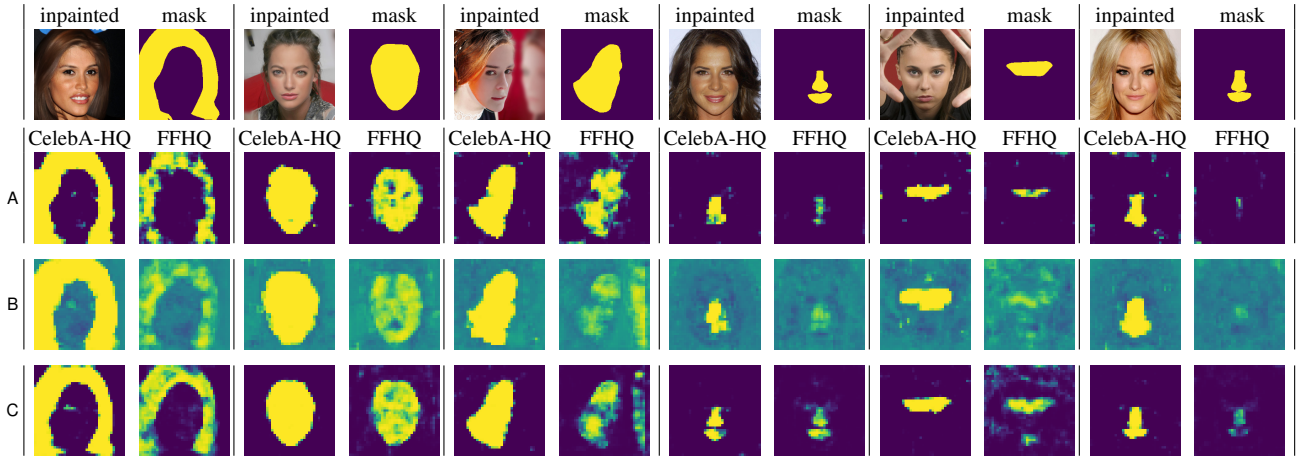

Figure 2. Soft localization maps when using the same and different source datasets for training and testing. For training we use data derived either form CelebA-HQ or FFHQ while for testing we use data derived from CelebA-HQ. With different training and testing source datasets the produced maps become less sharp and eroded, especially in the harder weakly supervised scenarios, A and B. Due to the noisy nature of the training in scenario B the separation between real and fake regions is dimmed.

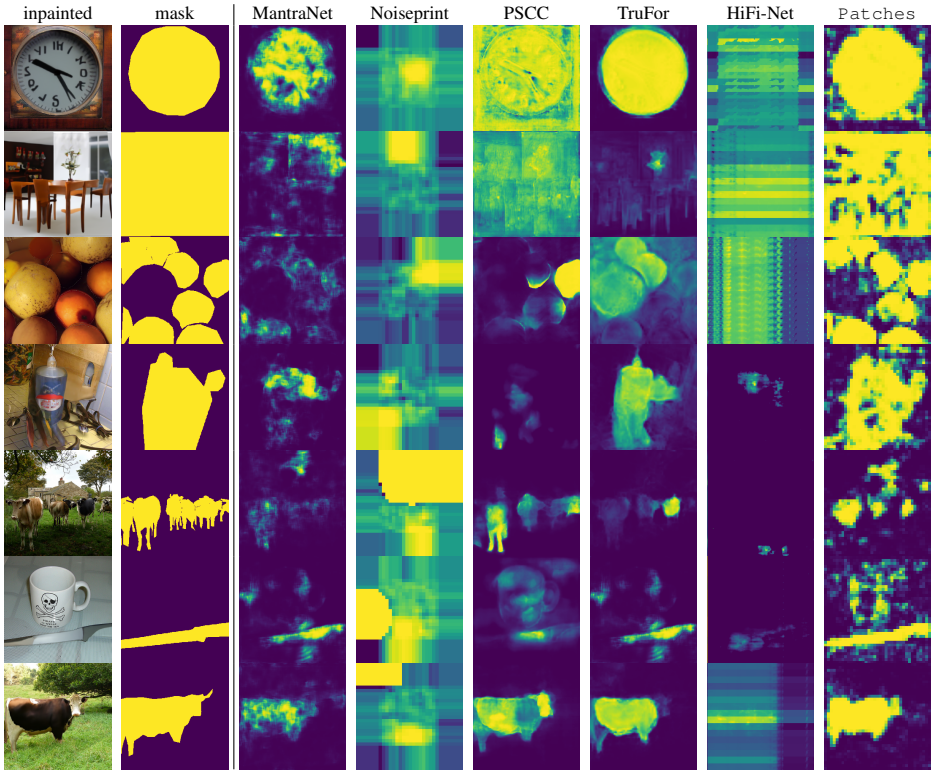

Figure 3. Visual results obtained with five pre-trained methods: MantraNet, Noiseprint, PSCC, TruFor, HiFi-Net and `Patches` on COCO Glide dataset. For these visualizations all methods are trained fully-supervised.