# Weakly-supervised detection of diffusion-based image manipulations

Dragoș-Constantin Țânțaru
Bitdefender
dtantaru@bitdefender.com

Elisabeta Oneață
Bitdefender
eoneata@bitdefender.com

Dan Oneață
University Politehnica of Bucharest
dan.oneata@gmail.com

## Abstract

*The remarkable generative capabilities of denoising diffusion models have raised new concerns regarding the authenticity of the images we see every day on the Internet. However, the vast majority of existing deepfake detection models are tested against previous generative approaches (e.g. GAN) and usually provide only a "fake" or "real" label per image. We believe a more informative output would be to augment the per-image label with a localization map indicating which regions of the input have been manipulated. To this end, we frame this task as a weakly-supervised localisation problem and identify three main categories of methods (based on either explanations, local scores or attention), which we compare on an equal footing by using the Xception network as the common backbone architecture. We provide a careful analysis of all the main factors that parameterize the design space: choice of method, type of supervision, dataset and generator used in the creation of manipulated images; our study is enabled by constructing datasets in which only one of the components is varied. Our results show that weakly-supervised localisation is attainable, with the best performing detection method (based on local scores) being less sensitive to the looser supervision than to dataset mismatch.*

## 1. Introduction

Image generation is improving by the day and it is arguably past the point where it is possible to perceptually distinguish between generated (fake) and real content. Generative adversarial models (GAN) [15], normalizing flows [40], denoising diffusion probabilistic models (DDPM) [48]—all provide excellent means for the creation of digital art or entertainment content. However, the advances in image generation come at the cost of also easing malicious use, *e.g.*, by altering reality or spreading misinformation. To counter these harmful effects, deepfake detection methods are developed to discriminate between fake and real samples [37, 38, 49].

Among the classes of generative models, diffusion models are emerging as the dominant paradigm [10], showcasing impressive results on a wide array of tasks including text-controlled image generation [39, 45, 42, 54] or image-to-image translation [44, 42, 32, 54]. Prior work on deepfake detection has naturally mostly considered detecting content generated by GANs [50, 16, 35, 53, 4], however the computer vision community is now starting to consider DDPMs [41, 7]. In this paper, we continue this direction, going one step further to address the task of weakly-supervised deepfake localization.

First, we extend prior approaches to localise the manipulated area and not only label the entire image as fake or real. The binary output of the typical deepfake detection methods provides only coarse and opaque information, especially in the frequent case of local manipulations and forgeries. Second, in contrast to prior work, which addresses localization in a fully-supervised setting [29, 56, 51, 22], we consider a weakly-supervised scenario, where we assume that we only have access to image-level labels and the models are not explicitly trained for localization.

Our work brings the following contributions: 1. We propose a **weakly-supervised** framework for deepfake localization in images that allows to systematically uncover the importance of various factors (**model**, **supervision type**, **dataset**) in the context of weakly-supervised localization of face manipulations. 2. We generate a detailed dataset with locally and fully manipulated images that allows **disentanglement** of different factors in deepfake manipulation localization. 3. We provide **extensive quantitative and qualitative results** as response to fundamental questions in understanding the factors determining the performance of weakly-supervised manipulation detection models. Our analysis provides insights about the models' sensitivity to looser supervision and dataset mismatch.

## 2. Related Work

**Deepfake detection of DDPM content.** Naturally, there is a vast body of work dedicated to the detection of GAN-generated images [49, 37, 33, 38]. However, the deep fake detection community has recently started considering

diffusion-generated images: for example, preliminary works use high-level cues such as inconsistencies in lighting [13] or perspective distortion [14], but end-to-end detection networks were also tested [41, 7]. The later focused on the transferability across classes of generative models (from GAN to DDPM, and vice versa), with the prevailing observation being that detectors do not generalize well across the two types of generators.

**Local manipulations.** A common setup in deepfake creation is altering a person's face by reenactment, replacement, editing or synthesis using techniques known as face swap, face transfer, facial attribute manipulations or inpainting [37]. These approaches result in local manipulations and are traditionally GAN-based. Increasingly larger and more complex datasets and challenges have emerged [43, 24, 30, 11, 27, 21] and, with these, a considerable effort has been made to expose those types of fakes [12, 33, 19, 1, 3, 55]. However, localizing manipulations has arguably received less attention than detecting whether an image is fake or not. Works that tackle localization rely on local noise fingerprint patterns [56, 29, 17, 34], attention mechanisms [8, 9, 36] or self-consistency checks [23, 2]. Very recent concurrent works proposed a forensic framework for general manipulation localization [17] and a hierarchical fine-grained formulation for image forgery detection [18]. Similar to us they consider diffusion-generated data with local forgeries, but differently they assumed full supervision.

## 3. Methodology

### 3.1. Methods for detection and localization

The task of deepfake detection consists of predicting whether an image is either real or fake. This task is usually framed as a binary classification problem and it is addressed using standard classification networks. In this paper we are interested in evaluating the capabilities of such methods in a weakly-supervised setting: if we assume only image-level labels, can these classifiers be successfully used for *localization* of partially manipulated images?

We identify and investigate three categories of approaches suitable for weakly-supervised localisation. These methods are based either on explanations (`GradCam`), local scores (`Patches`) or attention (`Attention`) (for visual depictions see Figure 1).

**`GradCam`.** While GradCAM explanations were previously used in the deepfake detection literature [55, 47, 52], they were mostly shown as qualitative results and were not evaluated quantitatvely, in terms of how well they localize the input alterations. In this paper we aim to quantify their performance and contrast them with other weakly-supervised localization methods. Concretely, we endow the Xception [6] network with localization capabilities by applying Grad-CAM [46] on the activations produced by block no. 12, the one before the last downsampling operation.

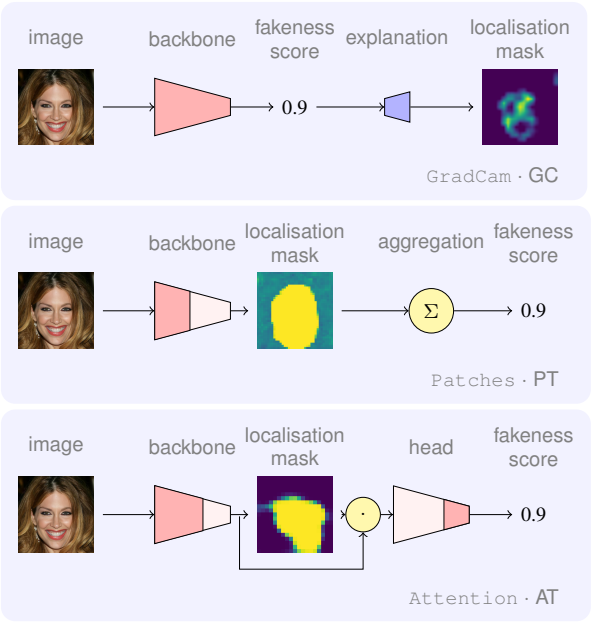


Figure 1. Overview of the three types of approaches proposed for the detection and localization of deepfakes. Each method is able to produce a fakeness score (for detection) and a mask (for localization); the mask is obtained either explicitly (for the first model) or implicitly (for the second and third models).

**`Patches`.** We use Patch–Forensics[4], which is a truncated classification network: it takes the feature activations after several layers and projects them to a patch-level score using $1 \times 1$ convolutions. At train time the loss is computed for each patch, while at test time, we average the per-patch softmax predictions to produce an image-level detection score. The authors experiment with two backbones (Xception [6] and ResNet [20]) and vary the depth of the selected trunk. We choose the Xception backbone truncated after the second block of layers, as this combination yielded good performance in the original work.

**`Attention`.** We start from [8], which augments an Xception [6] backbone with a learned attention mask to aggregate internal feature maps. The attention mask is used for localisation. To stabilize training, we replace the L1 loss with the binary cross-entropy loss (CE). Our final loss is

$$L = \mathrm{CE}(y, \hat{y}) + \lambda \mathrm{CE}(y, \max \hat{\mathbf{m}}), \tag{1}$$

where $y$ is the true image label, $\hat{y}$ is the predicted fakeness score, $\hat{\mathbf{m}}$ is the estimated localization mask, and $\lambda$ is a term that balances the two losses and is set through cross-validation.

| Type | Generation | Model | Dataset | Num. images | | |
|---|---|---|---|---|---|---|
| | | | | Train | Test | Val |
| Full | Diffusion | P2 | CelebA-HQ | 9K | – | 1K |
| Full | Diffusion | P2 | FFHQ | 9K | – | 1K |
| Local | Diffusion | P2–Repaint | CelebA-HQ | 30K | 8.5K | 3K |
| Local | Diffusion | P2–Repaint | FFHQ | 30K | – | 3K |

Table 1. Details of our proposed dataset with locally manipulated images and fully generated ones.

### 3.2. Datasets

To train and evaluate our models, we use real images and two types of fake images: fully synthesized and locally manipulated images.

**Real data.** We use two face datasets containing high-quality images: CelebA-HQ and FFHQ. CelebA-HQ [25] consists of 30K images that were selected and processed from the CelebA dataset [31]; we keep the original splits for training, validation and testing. FFHQ [26] consists of 70K images that have been crawled from Flickr and were automatically aligned and cropped.

**Fake data: Full-image synthesis.** We use the perception-prioritized (P2) diffusion method of Choi *et al*. [5] to sample fully-synthetic images. For both datasets we sample 10K images (9K for training and 1K images for validation). We refer to these datasets as P2/CelebA-HQ and P2/FFHQ, respectively.

**Fake data: Local manipulations.** We generate two locally inpainted datasets using the Repaint method [32] on the CelebA-HQ and FFHQ datasets. Since Repaint works on top of a pretrained full-image diffusion model, we use the pretrained P2 models, trained on CelebA-HQ and FFHQ, respectively. We refer to the resulting datasets as P2–Repaint/CelebA-HQ and P2–Repaint/FFHQ. The inpainted regions correspond to various face attributes (skin, hair, eyes, mouth, nose, glasses). For P2–Repaint/CelebA-HQ these annotations were manually labeled [25], while for P2–Repaint/FFHQ these are obtained using a pretrained face segmentation method [28]. Given an image (corresponding to the identity of a person) we generate multiple inpaintings by randomly sampling masks corresponding to these face attributes. For smaller parts (eyes, mouth, nose), we dilate the masks with a kernel of random size, up to 15 pixels. The masks occupy on average $18.3\%$ of the image surface.

## 4. Experiments

Our experiments evaluate the proposed methods with different levels of supervision, gradually changing the dataset and the generators in order to quantify their importance in manipulation localization. We investigate the performance using three levels of supervision:

- **Setup `A` (label & full)** is a weakly-supervised setup in which we have access to fully-generated images as fakes and, consequently, only image-level labels.
- **Setup `B` (label & partial)** is a weakly-supervised setup in which we have access to partially-manipulated images, but only with image-level labels (no localization information). This means that while an image may be labelled as "fake", not all of its regions are fake.
- **Setup `C` (mask & partial)** is a fully-supervised setting, in which we have access to ground-truth localization masks of partially-manipulated images.

All reported results are evaluated on the P2–Repaint/CelebA-HQ dataset. For localisation we measure intersection over union (IoU) and pixel-wise binary classification accuracy (PBCA), using a binarization threshold of $0.5$ on the predicted localisation masks. We also measure image-level detection in terms of average precision (AP); for this setting we also use real images from the CelebA-HQ test set.

### 4.1. Quantifying the localization abilities

**What type of localization approach is better for detecting local manipulations?** To answer this question we evaluate all three proposed approaches for manipulation localization in the three setups described above. To exclude other factors of variation we maintain the image generator and the source dataset fixed, that is for scenario `A` we train on P2/CelebA-HQ while for scenario `B` and `C` we use P2–Repaint/CelebA-HQ. Real data from CelebA-HQ is used in both scenario `A` and `B`; for the scenario `C` (fully supervised), real data is not needed. Results for both detection and localization are shown in Table 2. Among the three methods, the `Patches` method shows the best performance across all settings. However, as the training conditions progressively match those of testing, there is an improvement in IoU for all three methods, all converging to a very strong performance in the fully supervised scenario, `C`. This result confirms that deepfake detection models work very well in-domain, but their performance degrades when there is a mismatch between training and testing.

**Are localizations qualitatively different across methods?** Figure 2 shows examples of the localization maps produced by our three detection methods in three previously introduced scenarios. We notice that `Patches`, is able to partially recover the manipulated areas even in the initial setup, `A`, and improves considerably in the subsequent ones. `GradCam` and `Attention` both struggle in scenarios `A` and `B`, but their outputs are different: the former seems to produce weaker activations, which are also spread through irrelevant

| | sup. | generator | IoU (%) GC | IoU (%) PT | IoU (%) AT | PBCA (%) GC | PBCA (%) PT | PBCA (%) AT | AP (%) GC | AP (%) PT | AP (%) AT |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | label | full | 16.8 | **64.9** | 9.7 | 83.1 | **96.7** | 83.4 | 67.3 | **95.3** | 79.3 |
| B | label | partial | 21.5 | **37.7** | 23.2 | 85.1 | **79.8** | 86.3 | **94.4** | **95.3** | **94.4** |
| C | mask | partial | 83.7 | **84.5** | 70.3 | 96.8 | **98.7** | 97.6 | – | – | – |

Table 2. Evaluation of the three proposed localization techniques (`GradCam` GC, `Patches` PT, `Attention` AT) using different levels of supervision: image-level label on full images (`A`), image-level label on locally manipulated images (`B`) and fully-supervised masks (`C`). The generator is P2 and the dataset is CelebA-HQ. We evaluate both localization (using IoU and PBCA) and detection (using AP). `Patches` systematically outperforms the other two methods under all metrics and all scenarios.

| | sup. | generator | CelebA-HQ IoU | CelebA-HQ PBCA | CelebA-HQ AP | FFHQ IoU | FFHQ PBCA | FFHQ AP |
|---|---|---|---|---|---|---|---|---|
| A | label | full | 64.9 | 96.7 | 97.0 | 25.1 | 88.9 | 84.4 |
| B | label | partial | 37.7 | 86.3 | 95.3 | 23.3 | 64.4 | 75.2 |
| C | mask | partial | 84.5 | 98.7 | – | 32.3 | 89.2 | – |

Table 3. Evaluation of `Patches` on the Repaint/CelebA-HQ dataset using two training datasets: CelebA-HQ and FFHQ. When the source dataset does not match the target dataset, we observe a consistent drop in performance across all scenarios. This is more evident in scenario `B` where only image-level supervision is available for locally-manipulated images.

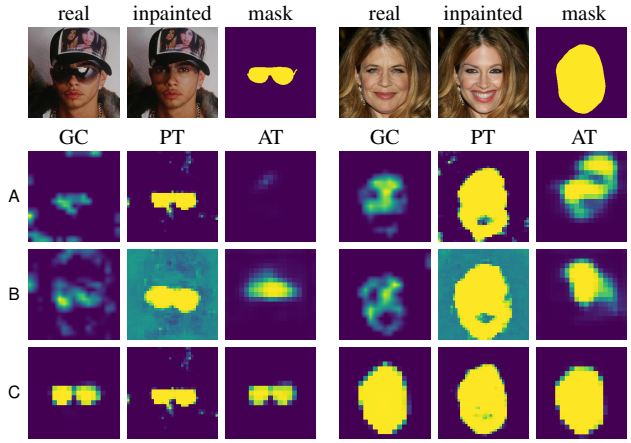


Figure 2. Soft localization maps produced by the three proposed approaches using different level of supervision. `Patches` (PT) can accurately detect the manipulations after having seen only fully generated fake images (scenario `A`) or locally-inpainted images with only image-level supervision (scenario `B`). Both `Attention` (AT) and `GradCam` (GC) 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 GradCam or attention.

areas of the image (this effect is more pronounced in scenario `A`), while the latter one does not detect any manipulations (scenario `A`) or outputs coarse manipulations (scenario `B`).

**What is the expected localization performance when training and testing datasets are different?** To answer this question we designed an experiment in which the training and testing data come from different source datasets, while keeping the generator, the method and the level of supervision the same. Quantitative results are shown in Table 3 for all scenarios under both localization and detection metrics. There is a consistent drop in performance across all scenarios and metrics when the dataset for training the generator is different. A closer look at the soft localization maps reveals a more complete picture (Figure 3): when training on FFHQ, the predictions overlap with the ground-truth even for small regions (nose and mouth), but they are less certain on the boundaries and the masks appear eroded or with holes.

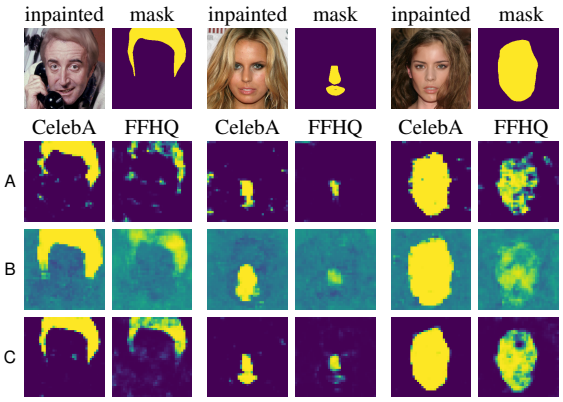


Figure 3. Soft localization maps when using the same and different source datasets for training and testing. For training we use data derived form either CelebA-HQ or FFHQ, while for testing we use data derived from CelebaA-HQ. If training and testing source datasets differe, then the produced maps become less sharp and eroded, especially in the weakly-supervised scenarios, `A` and `B`.

## 5. Conclusions

In this paper, we presented a first look at weakly-supervised localization in the context of diffusion-generated images of faces. We proposed a framework and a dataset that allows to systematically explore the importance of different factors in model performance, such as: choice of detection method, level of supervision, dataset and type of generator used. We summarize our findings: 1. The detection of local manipulations can be performed weakly supervised, even in the most restrictive scenarios. 2. The patch-based method consistently outperforms the other two approaches (explanations or attention) across multiple settings and metrics. 3. The detection performance in one of the weakly-supervised settings (image label, partial manipulations) is strong across all detection methods, suggesting that partially-manipulated images can be used for training deepfake classifiers.

## References


[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: A compact facial video forgery detection network. In *IEEE Workshop on Information Forensics and Security*, pages 1–7, 2018. 2

[2] Susmit Agrawal, Prabhat Kumar, Siddharth Seth, Toufiq Parag, Maneesh Singh, and R. Venkatesh Babu. SISL: Self-supervised image signature learning for splicing detection & localization. In *CVPRW*, pages 22–32, 2022. 2

[3] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. Hybrid LSTM and encoder–decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, 28(7):3286–3300, 2019. 2

[4] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? Understanding properties that generalize. In *ECCV*, pages 103–120, 2020. 1, 2

[5] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *CVPR*, pages 11472–11481, 2022. 3

[6] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017. 2

[7] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. *arXiv preprint arXiv:2211.00680*, 2022. 1, 2

[8] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *CVPR*, pages 5781–5790, 2020. 2

[9] Sowmen Das, Md. Saiful Islam, and Md. Ruhul Amin. GCA-Net : Utilizing gated context attention for improving image forgery localization and detection. In *CVPRW*, pages 81–90, 2022. 2

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 1

[11] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 2

[12] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *CVPR*, pages 9468–9478, 2022. 2

[13] Hany Farid. Lighting (in) consistency of paint by text. *arXiv preprint arXiv:2207.13744*, 2022. 2

[14] Hany Farid. Perspective (in) consistency of paint by text. *arXiv preprint arXiv:2206.14617*, 2022. 2

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[16] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In *ICME*, pages 1–6, 2021. 1

[17] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. *arXiv preprint arXiv:2212.10957*, 2022. 2

[18] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023. 2

[19] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *CVPR*, pages 5039–5049, 2021. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[21] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In *CVPR*, pages 4360–4369, 2021. 2

[22] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. SPAN: Spatial pyramid attention network for image manipulation localization. In *CVPR*, pages 312–328, 2020. 1

[23] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *ECCV*, pages 101–117, 2018. 2

[24] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics 1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, pages 2889–2898, 2020. 2

[25] 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

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

[27] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021. 2

[28] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[29] Ang Li, Qiuhong Ke, Xingjun Ma, Haiqin Weng, Zhiyuan Zong, Feng Xue, and Rui Zhang. Noise doesn't lie: Towards universal detection of deep inpainting. In *IJCAI*, pages 786–792, 2021. 1, 2

[30] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *CVPR*, pages 3207–3216, 2020. 2

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 3

[32] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pages 11461–11471, June 2022. 1, 3

[33] Asad Malik, Minoru Kuribayashi, Sani M Abdullahi, and Ahmad Neyaz Khan. Deepfake detection for human face images and videos: A survey. *IEEE Access*, 10:18757–18775, 2022. 1, 2

[34] Hannes Mareen, Dante Vanden Bussche, Fabrizio Guillaro, Davide Cozzolino, Glenn Van Wallendael, Peter Lambert, and Luisa Verdoliva. Comprint: Image forgery detection and localization using compression fingerprints. *arXiv preprint arXiv:2210.02227*, 2022. 2

[35] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs leave artificial fingerprints? In *IEEE Multimedia Information Processing and Retrieval*, pages 506–511, 2019. 1

[36] Ghazal Mazaheri, Niluthpol Chowdhury Mithun, Jawadul H. Bappy, and Amit K. Roy-Chowdhury. A skip connection architecture for localization of image manipulations. In *CVPRW*, pages 119–129, 2019. 2

[37] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1):1–41, 2021. 1, 2

[38] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M Nguyen. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223:103525, 2022. 1

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[40] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, pages 1530–1538. PMLR, 2015. 1

[41] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022. 1, 2

[42] 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. 1

[43] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *CVPR*, pages 1–11, 2019. 2

[44] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH*, pages 1–10, 2022. 1

[45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1

[46] 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 *CVPR*, pages 618–626, 2017. 2

[47] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *CVPR*, pages 18720–18729, 2022. 2

[48] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, volume 37, pages 2256–2265, 2015. 1

[49] Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020. 1

[50] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *CVPR*, June 2020. 1

[51] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *CVPR*, pages 9543–9552, 2019. 1

[52] Ying Xu, Kiran Raja, Luisa Verdoliva, and Marius Pedersen. Learning pairwise interaction for generalizable deepfake detection. In *WACV*, pages 672–682, 2023. 2

[53] Ning Yu, Larry S. Davis, and Mario Fritz. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *ICCV*, October 2019. 1

[54] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1

[55] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *CVPRW*, pages 1831–1839, 2017. 2

[56] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *CVPR*, pages 1053–1061, 2018. 1, 2