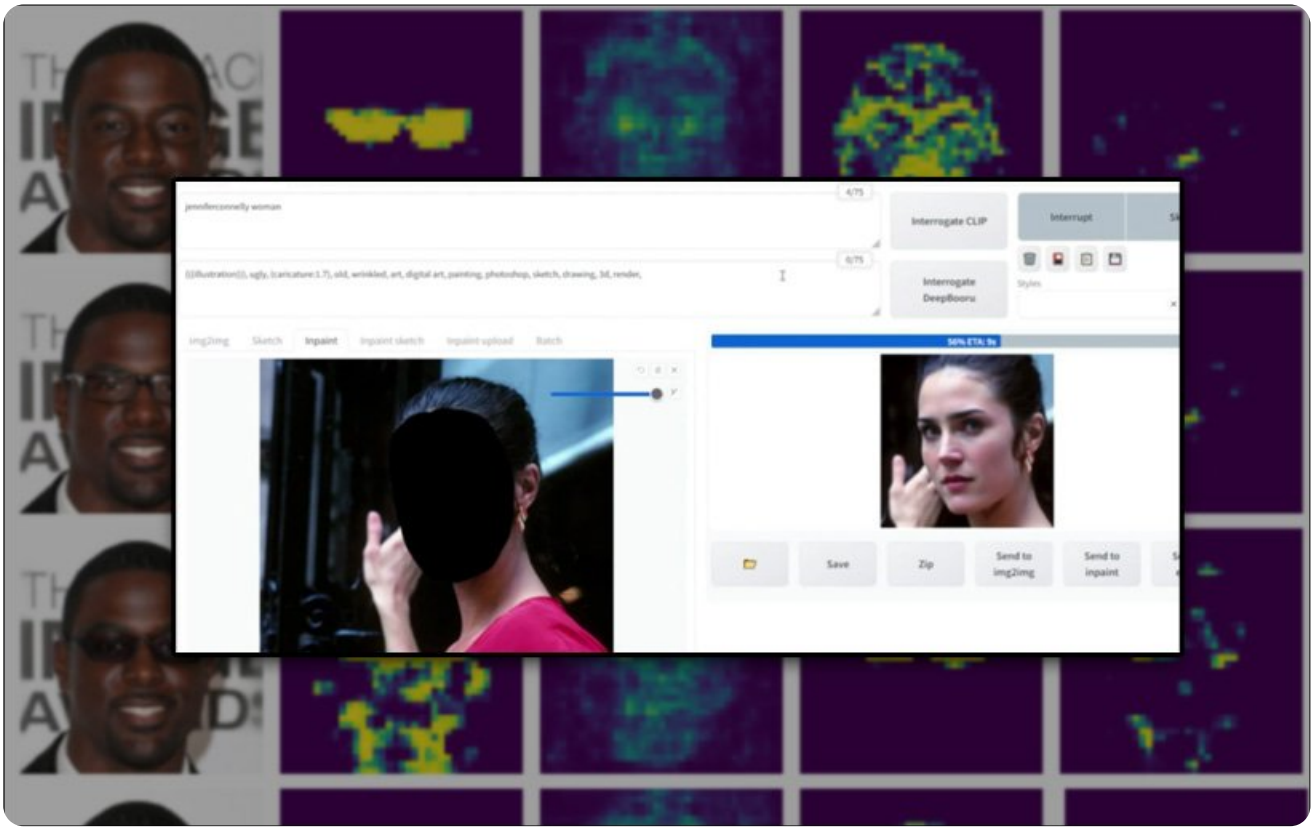
Detecting Stable Diffusion Deepfake Faces



November 13, 2023(https://blog.metaphysic.ai/2023/11/13/) 7:55 am

ABOUT THE AUTHOR

Martin Anderson

I'm Martin Anderson, a writer occupied exclusively with machine learning, artificial intelligence, big data, and closely-related topics, with an emphasis on image synthesis, computer vision, and NLP.

Author Website(https://martinanderson.ai) Author Archive(https://blog.metaphysic.ai/author/metaphysicai/)

Share This Post

Despite the vastly superior ability of Latent Diffusion Models (LDM) such as Stable Diffusion (https://blog.metaphysic.ai/stable-diffusion-is-video-coming-soon/) to create high-resolution representations of real people, in comparison to 2017-era autoencoder (https://blog.metaphysic.ai/future-autoencoder-deepfakes/) methods

(i.e., the methods used in deepfake (https://blog.metaphysic.ai/deepfakes/) videos over the past six years), the deepfake detection research sector has produced very few papers that address LDM's superior deepfaking capabilities.

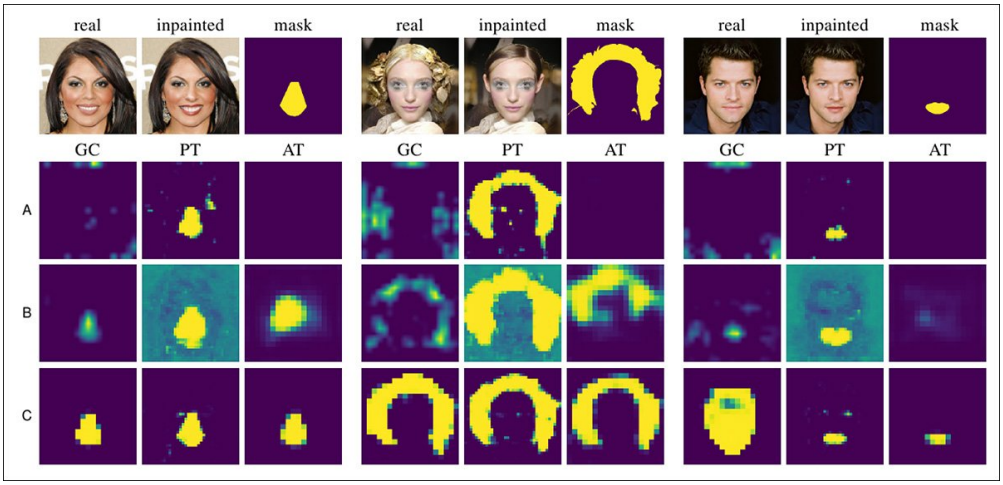
This may be due to the ongoing difficulty that LDMs have in creating temporally consistent (https://blog.metaphysic.ai/temporally-coherent-human-video-deepfakes-via-diffusion/) video. Since autoencoder methods such as DeepFaceLab (https://github.com/iperov/DeepFaceLab) and FaceSwap (https://github.com/deepfakes/faceswap/) (despite onerous training regimes, and results that are inferior to LDMs in quality) can produce *consistent* video fakes, and since video is considered the number one threat (https://www.reuters.com/article/bc-finreg-rising-threat-of-ai-deepfakes/tech-experts-see-rising-threat-of-genai-deepfakes-fbi-warns-of-generative-adversarial-networks-idUSKBN2YQ15Q) in the deepfake scene, security-based research into the facial deepfake properties of systems such as Stable Diffusion is quite nascent at the moment.

To date, this avenue of investigation has tended towards *fully-supervised* approaches. A fully-supervised approach assumes an unusual level of access to the technologies involved; in the security research sector, the equivalent field of study is 'white box' attacks, where new lines of research assume a level of access to the target technologies that is possible, but quite unlikely.

In the case of deepfake detection, a supervised approach will include knowledge of labels and other aspects of the data that are usually only available to the generating system, but cannot usually be known (or inferred) based on the output.

A *weakly-supervised* approach, by contrast, is equivalent to a 'black box' scenario in security research – where the methodology has access only to the superficial results of the system, and yet, hopefully, is able to infer some useful functionality only from this.

Considering these factors, a new paper from Bitdefender and the Polytechnic University of Bucharest offers such an approach, revising former methods so that they can be considered 'weakly-supervised', and therefore more generically applicable to a wider range of deepfake technologies – and especially diffusion-based approaches.



Soft localization maps visualized in the new paper, using revisions of three prior approaches to work in a weakly-supervised way. Source: https://arxiv.org/pdf/2311.04584.pdf

The methods offered in the paper show an improvement over prior offerings, and offer the possibility of increasing the generalization (https://blog.metaphysic.ai/what-is-generalization/) capabilities of deepfake detection models, so that they do not need to be constantly updated with the data from the most recent faking techniques.

The new approach concentrates on determining whether *individual sections* of an image have been altered, in contrast to the majority of work in recent years, which has primarily sought to determine whether an *entire image* has been generated.

The new approach is centered around alterations to original images, chiefly via diffusion-based methods, rather than GAN-style complete image generations, or generations where the background has been reproduced via image-to-image approaches, but nonetheless is not the original background.

The paper throws into sharp relief the laggard momentum of the anti-deepfake research sector, which continued to concentrate on Generative Adversarial Networks (https://blog.metaphysic.ai/the-future-of-generative-adversarial-networks-in-deepfakes/) (GANs) for some years after the almost-complete conquest of the deepfake scene by autoencoder methods, from late 2017 onward.

Likewise, with the sunk cost' of so much subsequent prior research into autoencoder deepfakes (a technology that has had no significant increase in quality in at least four years, and which can be considered to be 'stalled' at this point) finds the sector continuing to concentrate on autoencoders.

The new work focuses instead on the inpainting (https://huggingface.co/docs/diffusers/api/pipelines/stable_diffusion/inpaint) capabilities of LDMs, a functionality which converts source material into

latent embeddings and performs manipulations (such as face-swapping, for instance using LoRAs) directly in latent space (https://blog.metaphysic.ai/what-is-the-latent-space-of-an-image-synthesis-system/) rather than pixel space:

Inpainting a substituted likeness in the AUTOMATIC1111 Stable Diffusion distribution, using a free LoRA downloaded from civit.ai – the work of a minute.

Because the source image is manipulated into the target image very deep in the noise-decoding process of the latent diffusion model's latent space, inpainting conforms the target content better to the source content than most autoencoder and GAN projection techniques are capable of, resulting in trivially-easy static image fakes that comfortably pass standard tests:

The diffusion-based deepfake above passes every test for a standard fake detector routine. In fact, the algorithm identifies as 'suspicious' areas which are unchanged from the original, such as the hair, lower neck, and parts of the background near the head – perhaps because these are traditionally vulnerable spots in a standard, Photoshop-based face swap. Source: fotoforensics.com

Though there is a growing expectation (https://blog.metaphysic.ai/native-temporal-consistency-in-stable-diffusion-videos-with-tokenflow/) that LDMs will soon gain temporally consistent face manipulation, and though the inherent challenges may mean a longer wait for this functionality than many are presuming, such an event seems likely, at the current state-of-the-art in LDM-based deepfake detection, to take the research community by surprise, which makes forays such as the new project from Bitdefender and the Bucharest polytechnic a welcome addition to the literature.

Additionally, it's worth noting that if LDMs do ever become capable of generative consistency across frames, they are capable of deepfaking entire bodies (https://blog.metaphysic.ai/the-road-to-realistic-full-body-deepfakes/), and not just the central section of faces, as autoencoder methods currently do.

The new paper (https://arxiv.org/pdf/2311.04584.pdf) is titled *Weakly-supervised deepfake localization in diffusion-generated images*, and comes from two researchers at Bitdefender, and one from the University Politehnica of Bucharest.

Method

The new work revisits three former approaches to the task at hand: Gradient class activation maps (https://blog.metaphysic.ai/entanglement-in-image-synthesis/#gradcam) (Grad-CAM (https://arxiv.org/pdf/1610.02391.pdf)); the truncated image classification network Patch-Forensics (https://arxiv.org/pdf/2008.10588.pdf) (called *'patches'* in results) which obtains a patch (https://www.cs.toronto.edu/~mangas/teaching/320/slides/CSC320L03.pdf)-level score from feature (https://blog.metaphysic.ai/features-in-machine-learning/) activations; and the Facial Forgery Detection (http://cvlab.cse.msu.edu/project-ffd.html) (FFD, called *'attention'* in results) initiative from Michigan University's Computer Vision Lab, which uses an attention (https://vaclavkosar.com/ml/cross-attention-in-transformer-architecture) mechanism to create a mask of interest within a studied image.

For the new project, the Grad-CAM method was augmented by the researchers with the addition of an Xception network (https://openaccess.thecvf.com/content_cvpr_2017/papers/Chollet_Xception_Deep_Learn which brings localization capabilities lacking in the original version. The new paper, its authors state, represents the first version of this method to be tested quantitatively, rather than qualitatively (i.e., to be evaluated via metrics and functions, rather than just soliciting user opinion).

In turn, the Patch-Forensics architecture, which originally experimented with both ResNet (https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVF and Xception backbones, the authors of the new paper dispense with ResNet (since the entire thrust of the new paper centers on the superior performance and utility of Xception).

Finally for the FFD version produced for the paper, the L1 (https://cpatdowling.github.io/notebooks/regression_2) loss function (https://blog.metaphysic.ai/loss-functions-in-machine-learning/) on the masks used was replaced with binary cross-entropy loss (https://www.youtube.com/watch?v=DPSXVJF5jIs), and the final weight cross-validated.

In addition, to support a fully-supervised localization procedure (for testing and comparison purposes), the authors added a fully convolutional layer (https://docs.nvidia.com/deeplearning/performance/dl-performance-fully-connected/index.html#fullyconnected-layer) to the Grad-CAM network (as had previously been done in the project *Fully convolutional networks for semantic segmentation* (https://arxiv.org/pdf/1411.4038.pdf)).

Since dataset generation and curation is deeply bound into the new initiative, we'll move onto Data and Tests, and take a further look at the method there.

Data and Tests

The data generated for the system, and for testing the system, was produced via Stable Diffusion, generating both complete images and inpainted images (where the background component was unchanged across the generation). The new work uses the 2022 RePaint (https://arxiv.org/pdf/2201.09865.pdf) technique to perform inpainting.

Examples of inpainting from the 2022 RePaint project. Source: https://arxiv.org/pdf/2201.09865.pdf

The researchers devised a variation on this prior project called *Repaint-LDM*. They explain*:

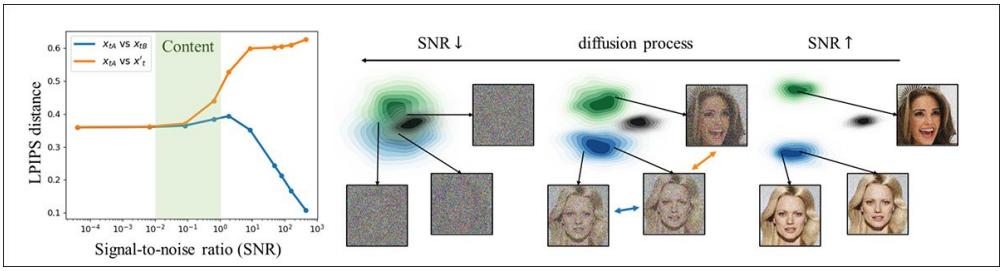
'Latent diffusion models (LDM (https://arxiv.org/pdf/2112.10752.pdf)) have been shown to offer a scalable approach to generating high-fidelity images. Their main idea consists of performing diffusion in the (low-dimensional) latent space of a variational autoencoder (VAE).

'We translate this idea to inpainting by running the Repaint [scheduler] in the latent space, x ← enc(x), of the variational autoencoder and using an appropriately downsized mask, m ← resize(m). This procedure generates an (inpainted) latent code, ˆx, which is then inverted to the original pixel space using the decoder of the VAE, dec(ˆx).

'Notably, this method allows us to inpaint an image using any existing pretrained LDM model. To the best of our knowledge, this approach to inpainting is novel.

Model training and evaluation was carried out using the popular CelebA-HQ (https://github.com/suvojit-0x55aa/celebA-HQ-dataset-download) and FFHQ (https://arxiv.org/pdf/1812.04948.pdf) datasets, largely because they were used in prior related projects, and allowed some continuity of testing criteria. From each of these, the researchers selected a subset of 9,000 training and 900 validation (https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data) images, to match the number of fake images that were generated for the project.

For the fake images, the authors used the perception-prioritized methods outlined in a prior 2022 paper (https://arxiv.org/pdf/2204.00227.pdf) from Korea.



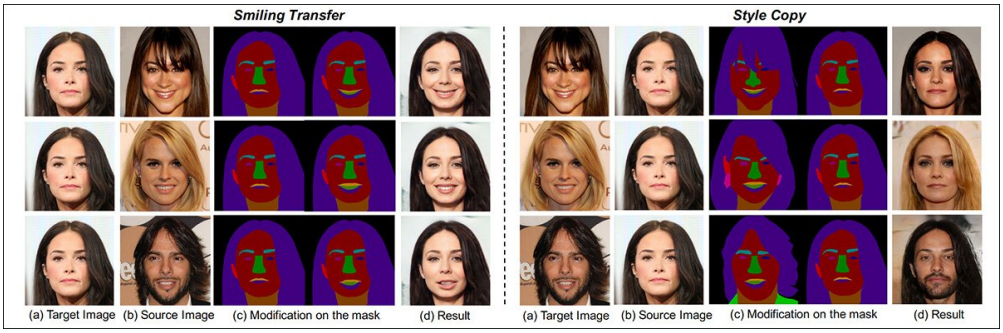
From the 2022 Korean paper, a visualization of the Perceptual distance of corrupted images as a function of signal-to-noise ratio (SNR). Source: https://arxiv.org/pdf/2204.00227.pdf

This particular project was chosen because it had leveraged CelebA-HQ and FFHQ.

The authors of the new work used this approach to generate a 90/10 training/validation corpus of fake images (called in results 'P2/CelebA-HQ' and 'P2/FFHQ', respectively).

The inpainted regions involved isolating the skin, hair, eyes, mouth, and nose of images, and addressing the addition or removal of glasses – all standard computer vision segmentation/synthesis tasks, many dating back (https://github.com/lecomte/glasses-removal-gan) to the earliest days of GANs

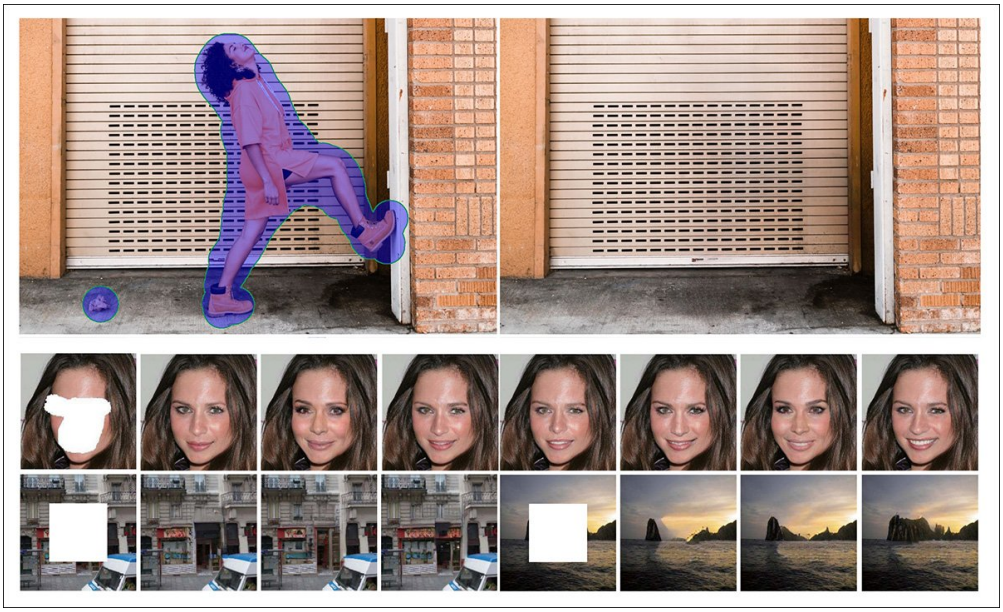
Two RePaint-based datasets were generated – one for CelebA-HQ, and one for FFHQ. In the case of CelebA-HQ, existing annotations could be used for the labeling needs of the project. Since FFHQ lacks such masks and ground truth, this was obtained by running the sub-set through MaskGAN (https://arxiv.org/pdf/1907.11922.pdf):



MaskGAN uses semantic segmentation to isolate sections of the face, which can be separately treated in subsequent facial synthesis. Source: https://arxiv.org/pdf/1907.11922.pdf

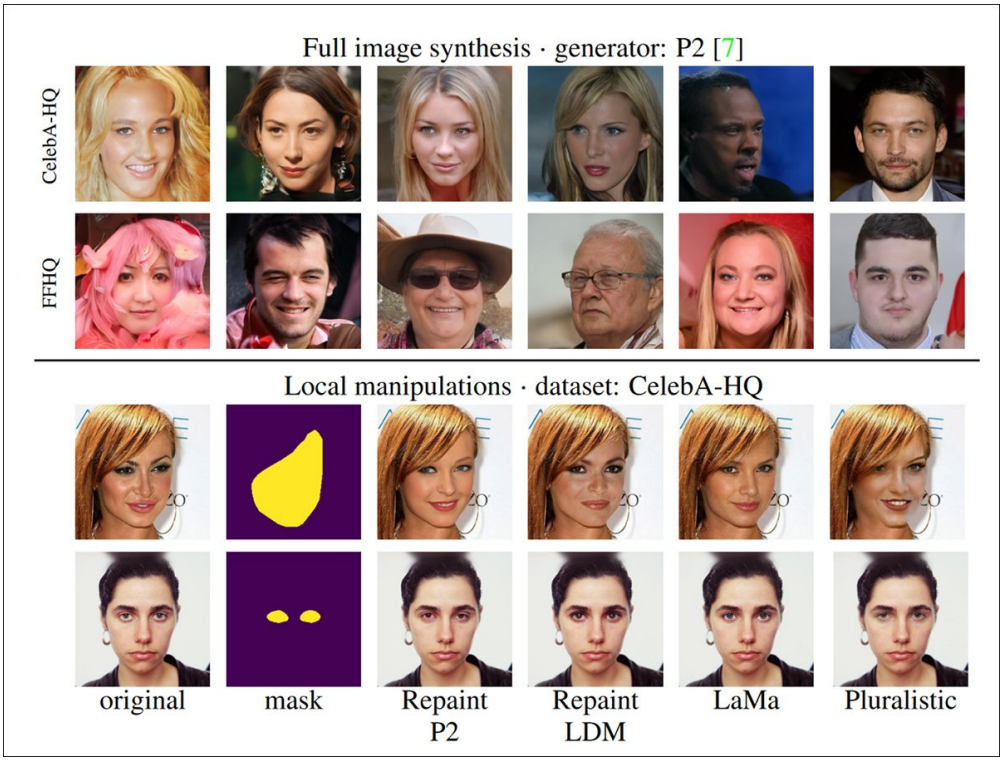
Mask facets, from those listed, were selected randomly, and the resulting sets are called 'Repaint-P2/CelebA-HQ' and 'Repaint-P2/FFHQ' in results. Only the first of these was used extensively in testing, while the second was primarily used for training.

For comparison, two additional inpainting methods were tested: LaMa (https://arxiv.org/pdf/2109.07161.pdf) (which uses Fourier convolutions (https://proceedings.neurips.cc/paper/2020/file/2fd5d41ec6cfab47e32164d5624269b1-Paper.pdf)) and Pluralistic (https://arxiv.org/pdf/1903.04227.pdf).



Above, LaMa removes a person using inpainting; below, examples of transformations achieved via Pluralistic. Sources (respectively): https://arxiv.org/pdf/2109.07161.pdf and https://arxiv.org/pdf/1903.04227.pdf

LaMA's Fourier convolutions are part of an autoencoder framework, while Pluralistic is a conditional variational autoencoder (https://arxiv.org/pdf/1606.05908.pdf) (VAE). Again, both these projects were trained on CelebA-HQ and FFHQ, which match them well to the new initiative (even though it arguably contributes to the difficulty in getting better and improved datasets embedded into the research sector – a 'dataset entropy' that we have discussed before (https://blog.metaphysic.ai/repairing-demographic-imbalance-in-face-datasets-with-stylegan3/)). Especially, this parity permitted the researchers to obtain consistent and comparable results in regards to the use of the same masks across examples, and helped to individuate the differences across generators.



Samples from the generated datasets – 'fake' images. Above are samples of entirely-synthesized pictures, and below, inpainted pictures.

In testing the system, the authors followed the procedures outlined in Patch–Forensics, which ensured that both real and fake images were subject to identical preprocessing steps before being passed through to the detection approaches. Therefore the images in both the real-world datasets were resized to $256px^2$.

For the fake detection stage, average precision (https://www.v7labs.com/blog/mean-average-precision) (AP) was used, with each image obtaining a per-image 'fakeness' score.

Three setups were arranged for the tests. The first of these is 'Setup A (label & full)', in which the researchers have access to fully-generated images with only image-level labels, consisting of 9,000 fake images fully synthesized by P2, and 9,000 related images from the dataset on which P2 was originally trained.

The second setup is called 'Set B (label & partial)'. This is a weakly-supervised configuration where image-level labels are available, but no localization information (i.e., which parts of the image have been changed). Thus an image labeled 'fake' by the detection process may not be *entirely* fake. This uses inpainted images from Repaint-P2 and 9,000 real images from the corresponding real-world training dataset.

The third and final setup is called 'Setup C (mask & partial)'. This is a fully-supervised setup with access to ground truth localization masks, and consists of 30,000 inpainted images from repaint-P2. No real images are used here.

Initially these setups were evaluated for localization.

(Note: the results section of this paper is characterized by complex codification, which makes the results unusually opaque and difficult to understand – not least because excessive concurrent and consecutive tests have been parsed into single table results; please bear with us as we attempt to decode the terminology and labyrinthine nature of the results)

In the localization results table below, Grad-CAM (*'GC'*), patches (*'PT'*) and Attention (*'AT'*) are all tested on the Repaint-P2/CelebA-GHQ dataset under the three levels of supervision described in the three setups outlined above. Localization is evaluated using Intersection over Union (https://giou.stanford.edu/) (IoU) and Pixel-wise binary classification Accuracy (http://cvlab.cse.msu.edu/pdfs/dang_liu_stehouwer_liu_jain_cvpr2020.pdf) (PCBA). 'AP', as mentioned earlier, means 'average precision'.

			IoU (%)			PBCA (%)			AP (%)		
setup	sup.	generator	GC	PT	AT	GC	PT	AT	GC	PT	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	**98.6**	97.6	–	–	–

Results for the test for localization. See paragraph above for a breakdown of the terms used here.

Of these results, the authors comment:

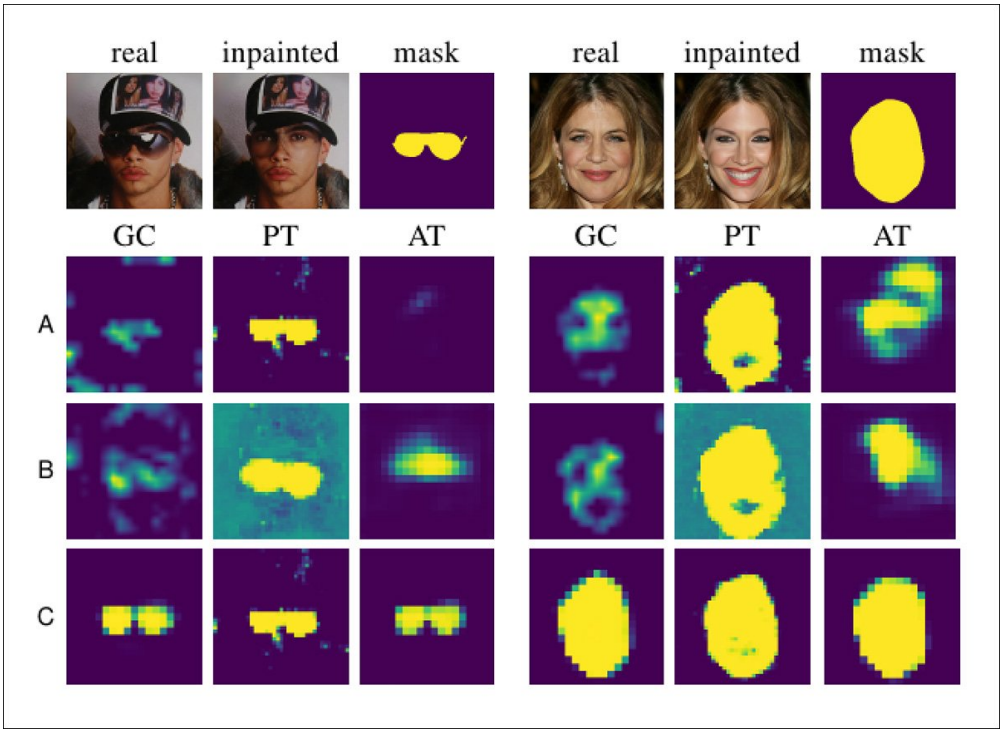
'[We] see that Patches generally outperforms the other two approaches across multiple setups and [metrics]. We see that localization performance is strong for all methods when training in the fully supervised scenario (setup C) and performance drops as we move to the two weakly supervised setups (setups A and B). Interestingly, GradCAM and Attention perform better in setup B than in setup A, while for Patches we observe the reverse trend.

'We believe that Patches is worse in setup B because the loss is set at patch-level, and the patch labels are inherently noisy as we use partially-manipulated images at input.

'In terms of detection (the 'AP' columns in Table 2), we observe strong performance of Patches in both weakly supervised setups, A and B. Interestingly, the detection performance is good for all models in setup B.

'In retrospect, this is expected since for the detection task in setup B the train data matches the test data.'

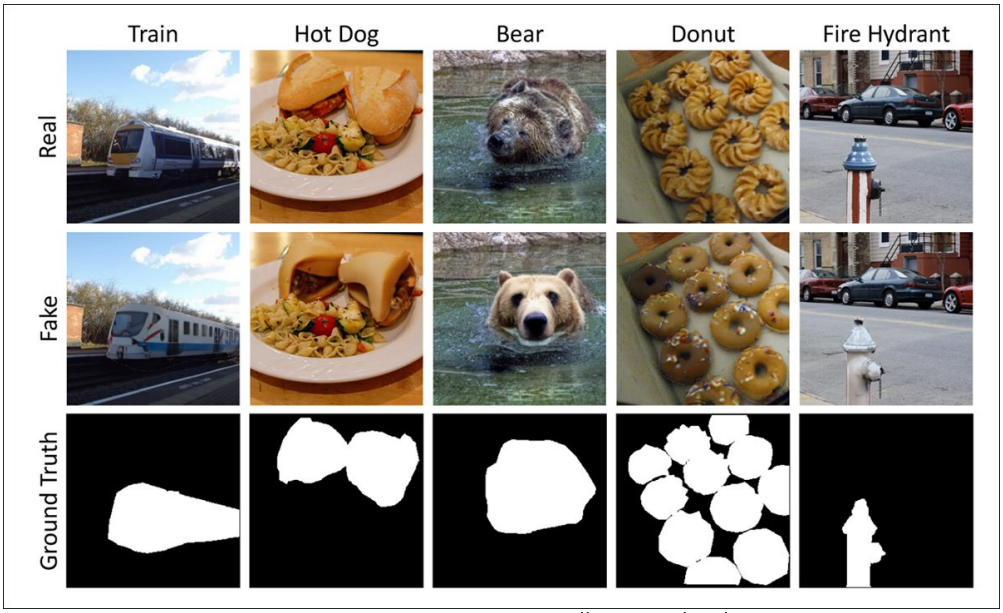
The paper provides some examples of the localization maps obtained by the detection methods in all three setups, which effectively equate to a deepfake detection result:



The authors suggest that while all methods are able to recover the manipulated region in the fully-supervised scenario, patch-level approaches may be superior either to Grad-CAM or attention.

Time and space don't permit us to cover the all of the exhaustive (and complexly-conveyed) results featured in this paper, some of which arguably constitute ablation studies (the paper has no ablation studies section) rather than core results; but we should mention one more test conducted that is particularly salient: *performance on unseen datasets*, which is the apposite context for a potential in-the-wild deepfake detector.

For this, an entirely 'alien' dataset – COCO Glide (https://arxiv.org/pdf/2212.10957.pdf) – was introduced, consisting of 512 images inpainted using a diffusion-based model.



Samples from the COCO Glide dataset. Source: https://arxiv.org/pdf/2212.10957.pdf

The authors cross-tested this data with five other prior methods, all trained on their own respective datasets: MantraNet (https://openaccess.thecvf.com/content_CVPR_2019/papers/Wu_ManTra-Net_Manipulation_Tracing_Network_for_Detection_and_Localization_of_Image_CVPR_2 Noiseprint (https://arxiv.org/pdf/1808.08396.pdf); PSCC-Net (https://arxiv.org/pdf/2103.10596.pdf); TruFor (https://arxiv.org/pdf/2212.10957.pdf); and HiFi-Net (https://arxiv.org/pdf/2303.17111.pdf). These were compared to the authors' own Repaint-P2/CelebA-HQ sets, as well as COCO Glide.

In order to compare with patches, the PSCC method was fine-tuned (https://blog.metaphysic.ai/fine-tuning-in-machine-learning/) in setup C on the Repaint-P2/CelebA-HQ dataset.

	R-P2/CelebA		COCO Glide	
Method	IoU	PBCA	IoU	PBCA
MantraNet [58]	4.8	**81.9**	25.1	79.8
Noiseprint [11]	18.2	23.8	23.9	29.0
PSCC [35]	14.3	66.5	**33.3**	80.6
TruFor [21]	**23.1**	81.3	. 29.2	**81.4**
HiFi-Net [22]	0.0	81.0	2.6	3.2
Methods trained on Repaint–P2/CelebA-HQ in setup C				
PSCC [35]	89.0	98.8	13.3	18.4
Patches	84.5	98.7	30.8	64.8

Evaluation of diverse systems on novel, unseen data.

Here the authors comment:

'We observe that the generalization performance is modest on either of the two datasets: the best out-of-domain performance on Repaint–P2/CelebA-HQ is 23.1%, obtained by TruFor, while on COCO Glide is 33.3%, obtained by PSCC.

'Even methods that have shown to generalize (TruFor [21]) or that have been trained specifically on diffusion images (HiFi-Net [22]) have difficulties on out-of-domain datasets. Patches shows competitive results (second best in terms of IoU on COCO Glide), even if it was trained solely on faces. Interestingly, this is not the case for PSCC. While PSCC obtains top performance in-domain, on Repaint–P2/CelebA-HQ, it struggles to [generalize] to COCO Glide.

'This behaviour suggests that overfitting is [occurring], which is not surprising given that the model capacity of PSCC (3.6M parameters) is an order of magnitude larger than the one of Patches (200k parameters).'

In concluding, the authors reiterate that the patch-based method outperforms the other two approaches tested, and that detection performance in in the image label & partial manipulations scenario performs well in a number of possible configurations.

This suggest, the authors contend, that inpainted images are a strong contender for the training of deepfake classifiers. However, they concede that localization of diffusion-inpainted images is 'very challenging even in the most optimistic scenario'.

Conclusion

Unfortunately, the opacity and organizational compression of this paper makes it one of the most inaccessible that we have ever covered – which is a shame, as it has a couple of interesting takeaways, in a sector which seems poised to explode in the next 12-18 months.

The success of patch-based approaches indicates that this may be a fruitful line of research, and the fact that almost the entirety of the new paper constitutes an ablation study that could arguably have preceded more focused follow-on research, means that this was a hard-won revelation.

The second encouraging facet of the paper is that it was able to indicate a road forward at all in the area of weakly-supervised fake detection. In a research line that's currently setting out on a potentially futile watermark war (https://arstechnica.com/ai/2023/10/researchers-show-how-easy-it-is-to-defeat-ai-watermarks/?comments=1), this represents a refreshing and even promising direction.

** My conversion of the researchers' inline citations to hyperlinks.*

(https://blog.metaphysic.ai/cgi-style-object-control-with-stable-ıttps://blog.metaphysic.ai/restoring-facial-expressions-with-cyclegan/)

More To Explore

Controllable Deepfakes With Gaussian Avatars (Https://Blog.Metaphysic.Ai/Controllable-Deepfakes-With-Gaussian-Avatars/)

Could Gaussian Splatting become the hottest new deepfake technology since 2017? The massive surge of interest from the research sector suggests it might – and the latest innovation not only brings full controllability to neural or deepfaked faces, but also lets you become someone else at an unprecedented level of photorealism and efficiency.

Martin Anderson · December 5, 2023

Badly-Compressed Images Affect CLIP's Performance, New Research Contends (Https://Blog.Metaphysic.Ai/Badly-Compressed-Images-Affect-Clips-Performance-New-Research-Contends/)

CLIP is the new darling of the computer vision research, and of image-based generative AI, with wide uptake of the image/text analysis framework across the sector. However, new research indicates that CLIP's efficiency and usefulness is negatively affected by badly-compressed images. Though this should not be a problem in the modern high-speed broadband age, it is – because so much essential data and methodologies still in use data back several decades.

Martin Anderson · November 28, 2023

> It is the mark of an educated mind to be able to entertain a thought without accepting it.
>
> ARISTOTLE

METAPHYSIC

QUICK LINKS

Home(https://metaphysic.ai/)

Every Anyone(https://everyany.one)

Synthetic Futures(https://syntheticfutures.org)

CONNECT WITH US

Discord(https://discord.gg/5vshCNWTuw)

Tiktok(https://www.tiktok.com/@deeptomcruise)

Twitter(https://twitter.com/Metaphysic_ai)

Youtube(https://www.youtube.com/channel/UClbSYyDnUCa6NzLjLqPdMoA)

Instagram(https://www.instagram.com/metaphysic.ai/)

Github(https://github.com/Metaphysic-ai)

Linkedin(http://www.linkedin.com/company/metaphysic-ai/)

CONTACT INFO

info@metaphysic.ai

press@metaphysic.ai