Local Diffusion Edits in Facial Images: Detection and Explainability with CNNs

Thomas Cook

November 2025

Abstract

This work investigates the detection of subtle, local facial edits created using diffusion models. These manipulations preserve the global structure and visual realism of the original image, making them difficult to detect using conventional forensic techniques. I use the Semi-Truths dataset, which provides real and manipulated image pairs along with edit metadata, to study this problem. A convolutional neural network classifier is trained to distinguish between real and edited faces, focusing on local alterations introduced by text-guided inpainting. Explainability methods are used to interpret model predictions and assess whether it attends to manipulated regions. The goal is to evaluate the model's ability to detect fine-grained diffusion-based manipulations and assess its interpretability.

1 Related Work

1.1 Subtle Local Edits in Diffusion-Based Facial Manipulation

Diffusion models have made image editing far more flexible by allowing small, targeted changes that still look natural. Latent diffusion models work in a compressed latent space and use cross-attention to decide where edits should appear and how strongly they should be applied [1]. Because only some regions are regenerated, the rest of the image stays mostly fixed. The examples in the original paper demonstrate that the model can fill in missing patches or update selected regions while maintaining the scene's layout and overall appearance. Even though the paper is not focused on faces, the same behaviour applies to facial editing, as the model maintains the structure while adjusting texture or appearance in the edited area.

Instruction-based editors build directly on this idea. InstructPix2Pix fine-tunes a diffusion model so it can follow natural language editing instructions, and the authors demonstrate a wide range of edits on faces and general scenes [2]. These include changing expressions, adjusting hair or colour tones, or adding small objects. Throughout their examples, most of the identity and layout stay in place. Although this is not formally measured, the results indicate that diffusion models can make clean, local changes without affecting the surrounding areas.

This ability to make realistic, small-area edits is exactly what makes diffusion models difficult for image forensics. Older deepfake pipelines usually modify large parts of a face or introduce strong blending artefacts. In contrast, diffusion-based edits can tweak just one detail, like the shape of a mouth or a highlight on the skin, without leaving obvious traces. Bazyleva et al. highlight this in the X-Edit benchmark, where standard detectors perform well on global manipulations but often fail to catch small text-guided diffusion edits [3]. Their results suggest that detectors trained on classic forgeries tend to pick up broad distortions rather than the fine-grained changes produced by diffusion inpainting.

Earlier datasets like FaceForensics++ mainly contain face swaps and reenactments that change almost the entire facial region and commonly introduce mismatches in geometry, lighting, or compression [4]. Detectors trained on such data learn to rely on these large-scale inconsistencies. When these same detectors face a diffusion edit that only touches a tiny region while leaving the rest of the face untouched, their assumptions no longer hold. Because of that, there is a clear need to study detectors in a setting where edits are subtle and local rather than global and obvious.

1.2 Datasets for Realistic Local Image Manipulations

More recent datasets have started to shift toward this idea of local editing. They focus on manipulations that adjust only a small part of the image, keep the wider scene intact, and provide masks or labels that mark exactly where the edit happened.

X-Edit is built specifically around text-guided diffusion edits, offering real–edited image pairs along with detailed masks that highlight the modified pixels [3]. The edits include changes to facial attributes, altered objects, and other minor adjustments. The dataset was created to show where current detectors fall short, and the authors report that local edits often go unnoticed, even when the change is meaningful.

CSI-IMD takes a broader view, including various editing types such as object removal, object insertion, and inpainting-style operations [5]. Each example comes with a spatial mask and a "semantic impact" score that reflects how much the manipulation alters the meaning of the image. Because the dataset includes both faces and general scenes, it allows researchers to compare detector behaviour across multiple contexts. It is especially useful for studying small edits that are still important for understanding the image.

RetouchingFFHQ focuses specifically on faces and models the small cosmetic adjustments that appear in everyday photo editing [6]. These include smoothing, whitening, enlarging the eyes, and similar operations. The authors apply each change at multiple intensity levels while keeping identity and pose fixed, which makes it easy to analyse how detection difficulty increases as the edit becomes more subtle.

Semi-Truths, the dataset used in this project, is built around real images and a large set of AI-augmented variants [7]. Each edited image comes with a manipulation mask and metadata describing how the modification was created. The dataset includes various augmentation types, but a relevant subset contains targeted diffusion-based edits that modify only a small region while preserving the global layout. This makes it a strong fit for studying subtle local facial manipulations, since it allows clean comparisons between real and locally altered versions of the same face.

1.3 CNN Architectures for Detecting Visual Forgeries

CNNs are still widely used in image forensics because they can learn both high-level and low-level patterns that separate real images from manipulated ones. Many approaches begin by modifying the early layers, allowing the network to focus more on noise patterns rather than semantic content. Bayar and Stamm introduce a constrained convolutional layer that forces the filters to behave like high-pass operators, removing most semantic information and highlighting manipulation traces [8]. Cozzolino et al. follow a similar idea by embedding steganalysis-inspired residual filters into the model and allowing them to be fine-tuned during training [9]. Both approaches aim to make the model more sensitive to small editing artefacts that appear in the residual domain.

Other work combines residual cues with more standard RGB information. Zhou et al. propose a two-stream CNN where one branch processes the RGB image and the other processes SRM residual maps [10]. These two representations are fused before prediction, allowing the network to utilise both semantic and noise-based signals. Rao and Ni demonstrate that patch-level CNNs can effectively detect local inconsistencies, such as splicing boundaries or copy-move traces, at the patch level [11]. ManTra-Net extends this line of work by utilising a fully convolutional model trained on a diverse range of synthetic manipulations, enabling it to detect anomalous regions without requiring explicit masks during training [12].

The limitation shared across most of these architectures is that they were developed for traditional manipulation types, where edits tend to be large or produce clear artefacts. Their performance on subtle, diffusion-based local edits remains unclear. This gap motivates evaluating how a modern CNN behaves when the only difference between two face images is a very small, high-quality diffusion edit.

1.4 Visual Explanations for CNN-Based Forensics

Interpretability tools are useful in forensics because they help verify that a detector is focusing on the manipulated region rather than unrelated parts of the image. Class Activation Mapping (CAM) demonstrates that CNNs with global average pooling can generate heatmaps that highlight regions associated with a prediction [13]. Grad-CAM builds on this idea and generalises it to many different architectures by using gradients to create class-specific localisation maps [14]. These methods are common in forensic research because they provide a quick way to verify whether a model is focusing on the correct area.

Other explanation methods provide different views of the model's behaviour. Zeiler and Fergus use a deconvolutional approach to visualise which input patterns activate specific feature maps at different layers [15]. Layer-wise Relevance Propagation takes the final prediction and distributes relevance scores back to individual pixels, producing an explanation of what contributed to the model's decision [16]. Together, these approaches help show whether a detector's reasoning aligns with the actual manipulated region.

In this project, Grad-CAM is used to inspect the trained detector and verify whether its predictions are indeed driven by the subtle diffusion edits present in the manipulated facial regions.

2 Dataset Selection and Collection

2.1 Overview of the Dataset

This project uses the Semi-Truths dataset, a large-scale benchmark designed for evaluating the robustness of detectors against AI-generated edits. Semi-Truths contains 27,600 real images and more than 1.47 million manipulated variants created through two editing pipelines: diffusion inpainting and text-guided prompt-based editing. Each manipulated image is accompanied by rich, structured metadata describing how the edit was produced, including the edited region, the mask used, the diffusion model, and several similarity and quality-control metrics.

Unlike classic deepfake datasets, which typically involve global face swaps or full-region manipulations, Semi-Truths emphasises small, targeted changes. These edits preserve the global structure of the image and often modify only a tiny local region. This makes the dataset particularly suited to my thesis objective: investigating whether convolutional neural networks can detect subtle, local diffusion-based manipulations on human faces.

2.2 Why I Chose This Dataset

The goal of this project is to detect very small, localised edits on faces. Semi-Truths aligns with this aim for three main reasons.

Local, region-specific edits. Semi-Truths provides masks that isolate individual facial regions such as the eyes, nose, ears, hair, and mouth. This is a major advantage over datasets like FaceForensics++, where the entire face is replaced or reenacted. Because my work focuses on subtle local edits rather than global identity swaps, this property of Semi-Truths is essential.

Diverse diffusion models. The subset of Semi-Truths used in this project contains edits produced by five different diffusion models:

- Stable Diffusion XL (35,639 edited images),
- Stable Diffusion v4 (32,453),
- Stable Diffusion v5 (30,428),
- OpenJourney (28,714),
- Kandinsky 2.2 (14,730).

This diversity prevents the CNN from overfitting to the artefacts of a single model and ensures a more realistic and challenging evaluation setting.

Paired real and edited images. Each manipulated image is paired with its corresponding unedited original. This pairing allows me to train the detector as a binary classifier while also enabling spatial analysis using the edit masks. Because the real and edited images share identical global structure, the detection task becomes sensitive to only the altered region.

Together, these features make Semi-Truths the most suitable publicly available dataset for studying subtle diffusion-based edits on facial images.

2.3 Why I Selected the CelebAHQ Subset

Semi-Truths aggregates several source datasets, including ADE20K, SUN RGBD, CityScapes, OpenImages, HumanParsing, and CelebAHQ. Only CelebAHQ provides high-resolution, clean, front-facing human faces, which is essential for a focused investigation of facial manipulation detection. For this reason, I restricted my work to the CelebAHQ subset.

The final working subset contains:

- 5,000 real CelebAHQ images, and
- 141,964 edited CelebAHQ images.

Despite being a subset, it still includes a large variety of edit types. The distribution of edited facial regions is shown in Table 1. These masks confirm that the dataset covers many meaningful facial landmarks and features, giving the model ample variation for learning subtle cues.

This broad coverage across facial regions ensures that the CNN does not become biased toward a single type of manipulation but instead learns generalisable features for subtle edit detection.

Table 1: Distribution of edited facial regions in the CelebAHQ subset.

Region	**Count**
neck	19,680
hair	19,124
nose	18,940
skin	18,393
lower lip	15,824
cloth	10,180
left ear	9,574
right ear	8,429
mouth	6,021
upper lip	5,344

2.4 Dataset Source and Format

Semi-Truths is publicly hosted on Hugging Face at:

`https://huggingface.co/datasets/semi-truths/Semi-Truths`

It is distributed under a Creative Commons licence and was originally released as part of the NeurIPS 2024 Datasets & Benchmarks Track.

The dataset is provided in three main components:

- WebDataset image shards (`.tar.bz2`) containing real and edited images,
- Segmentation masks for the inpainting regions,
- Metadata CSV files describing the edits.

For my project, I accessed the dataset through the Hugging Face Python API and downloaded the metadata files corresponding to the CelebAHQ inpainting subset.

2.5 Format of the Raw Data

The dataset is organised into several top-level folders:

- `original/` — real images and segmentation masks,
- `inpainting/` — diffusion-generated local edits,
- `prompt-based-editing/` — text-driven edits,
- `metadata/edited/` — detailed CSV metadata files.

For this project, the following files were most important:

- `metadata/edited/raw_values/inpainting.csv`,
- `original/images/*.tar.bz2`,
- `original/masks/*.tar.bz2`.

The metadata includes fields such as `img_id`, `perturbed_img_id`, `mask_name`, `area_ratio`, `diffusion_model`, and `pass_qc`. These structured labels made it straightforward to isolate CelebAHQ images and filter for facial edits.

2.6 Dataset Collection and Assembly Process

Although I did not collect images manually, assembling the working dataset required several processing and filtering steps.

Downloading and inspecting metadata. I downloaded the raw inpainting metadata from Hugging Face and loaded it into a pandas DataFrame to explore its structure and available fields.

Filtering for CelebAHQ facial edits. To restrict the data to human faces, I filtered the metadata using:

```
df = df[df["dataset"] == "CelebAHQ"]
df = df[df["mask_name"].isin(facial_regions)]
```

This produced 5,000 real images and 141,964 edited images.

Ensuring model diversity. I calculated the number of edits contributed by each diffusion model to confirm that the subset remains diverse and challenging.

Extracting relevant fields. For downstream training, I selected only the essential metadata fields:

```
["img_id", "perturbed_img_id", "dataset", "diffusion_model",
 "mask_name", "area_ratio", "pass_qc"]
```

The resulting file, `celebhq_faces_subset_metadata.csv`, serves as the index for loading paired real and edited images efficiently during training.

2.7 Ethical Considerations

All CelebAHQ images are anonymised and widely used in academic research. Semi-Truths adds only synthetic edits and does not introduce any personally identifiable information. For this reason, no additional ethical approval or anonymisation procedures were required. All data is used solely for research purposes and stored securely.

2.8 Summary

The Semi-Truths dataset provides exactly the kind of subtle, local facial manipulations required for this project. Restricting the dataset to the CelebAHQ subset produced a focused working set of 5,000 real images and nearly 142,000 manipulated ones. The filtering and assembly steps created a clean, well-structured dataset suitable for training and evaluating a CNN designed to detect fine-grained diffusion-based facial edits.

References

[1] Robin Rombach et al. “High-Resolution Image Synthesis With Latent Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022.

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. "InstructPix2Pix: Learning To Follow Image Editing Instructions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023.

[3] Valentina Bazyleva, Nicolo Bonettini, and Gaurav Bharaj. *X-Edit: Detecting and Localizing Edits in Images Altered by Text-Guided Diffusion Models*. 2025. arXiv: `2505.11753 [cs.CV]`. URL: `https://arxiv.org/abs/2505.11753`.

[4] Andreas Rossler et al. "FaceForensics++: Learning to Detect Manipulated Facial Images". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[5] Yuwei Chen et al. "A Semantically Impactful Image Manipulation Dataset: Characterizing Image Manipulations using Semantic Significance". In: *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*. Feb. 2025.

[6] Qichao Ying et al. "Retouchingffhq: A large-scale dataset for fine-grained face retouching detection". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023.

[7] Anisha Pal et al. *Semi-Truths: A Large-Scale Dataset of AI-Augmented Images for Evaluating Robustness of AI-Generated Image detectors*. 2024. arXiv: `2411.07472 [cs.CV]`. URL: `https://arxiv.org/abs/2411.07472`.

[8] Belhassen Bayar and Matthew C Stamm. "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection". In: *IEEE Transactions on Information Forensics and Security* 13.11 (2018).

[9] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. "Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection". In: *Proceedings of the 5th ACM workshop on information hiding and multimedia security*. 2017.

[10] Peng Zhou et al. "Learning rich features for image manipulation detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[11] Yuan Rao and Jiangqun Ni. "A deep learning approach to detection of splicing and copy-move forgeries in images". In: *2016 IEEE international workshop on information forensics and security (WIFS)*. IEEE. 2016.

[12] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. "Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[13] Bolei Zhou et al. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[14] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017.

[15] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer. 2014.

[16] Sebastian Bach et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140.