

# Supplementary Materials: Rethinking Image Editing Detection in the Era of Generative AI Revolution

Anonymous Authors

## A GRE CONSTRUCTION

### A.1 Regional Editing Pipeline

*A.1.1 ChatGPT instruction.* We utilize the GPT-3.5 model provided by ChatGPT to generate innovative editing ideas. To effectively inform ChatGPT about the content of the original image and the specific region selected for editing, we specifically employ the designed instruction:

Now you have an image described as image prompt: [Image Prompt], which has a region described as source prompt: [Source Prompt]. This area may be an object or an empty background area. Now you need to edit or tamper with this area. You can make creative edits out of fun, or you can change the original meaning of this image out of any intent. Please give your editing ideas and describe the expected target of this region as target prompt: [Target Prompt].

we replace the placeholders, including [Image Prompt] and [Source Prompt], in the instruction with specific prompts, derived from previous steps that describe the entire image and the original content of the selected region, respectively. We then extract [Target Prompt] from the results returned by ChatGPT as the editing idea for subsequent steps.

*A.1.2 Other ways to use ChatGPT.* Compared to GPT-3.5, GPT-4 can process multimodal information and complete multimodal tasks. Relying on GPT-4’s ability to understand images, we attempt to input the original image with a red boundary marking the region to be edited (as shown in our case demonstrations) into GPT-4, and provide the coordinates of the bounding box for further direct guidance. This approach is an effective solution and could further simplify our pipeline. However, access to GPT-4 is limited, preventing us from fully updating our pipeline to build a large-scale dataset. This limitation is worth considering for future enhancements.

Recently, ChatGPT introduced a region editing function based on DALL-E. As of now, this function only allows for the manual selection of regions in images, and the number of accesses is limited, so it cannot automatically generate large-scale data. From another perspective, a major advantage of the GRE dataset is its diversity, and our extensive experiments and analysis verify the significance of covering a variety of generative editing methods in the generalization evaluation of detection models. Therefore, the pipeline we designed in GRE remains meaningful and cannot be completely replaced.

### A.2 Various Generative Editing Methods

The diversity of editing methods used in the GRE dataset is a major advantage and characteristic. This diversity allows it to serve as an excellent training dataset for detection models, enabling them to detect generative regional editing better and generalize across various editing methods. Additionally, it functions as a reasonable

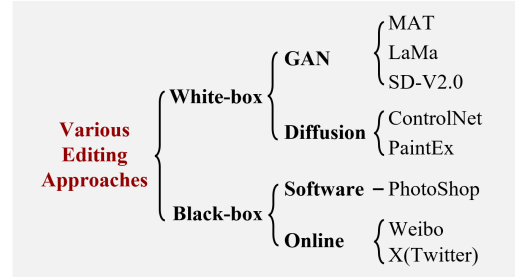


Figure 1: Various editing methods in the GRE dataset.

Table 1: The distinctive characteristics of representative generative regional editing methods.

Method	Architecture	Guidance
MAT	GAN	[Region Mask]
LaMa	GAN	[Region Mask]
SD-V2.0	Diffusion	[Region Mask],[Target Pro.]
ControlNet	Diffusion	[Region Mask],[Target Pro.]
PaintEx	Diffusion	[Region Mask],[Target Pro.],[Target Ex.]

benchmark, providing a more accurate assessment of the generalization performance. The GRE dataset includes many types of editing methods, mainly encompassing white-box editing methods used in our designed pipeline, as well as images edited with image editing software and collected from real-world online scenarios. All included editing methods are illustrated in Figure 1.

*A.2.1 Editing methods used in pipeline.* We choose five editing methods to complete the final stage in the pipeline, implementation. The emergence of diffusion models has significantly enhanced the capabilities of generative editing methods in terms of both convenience and effectiveness. Among diffusion-based editing methods, we select Stable Diffusion (SD-V2.0), ControlNet, and PaintByExample (PaintEx), each capable of performing removal, replacement, and creation editing types. Although diffusion models have demonstrated superior image generation capabilities compared to GANs, we must also acknowledge the significant performance improvements that have occurred recently in GAN-based image editing methods, particularly in removal tasks. MAT and LaMa are selected as representatives of these types of methods. Table 1 summarizes the architecture and required guidance for these methods.

*A.2.2 Edited images in real-life scenarios.* We attempted to collect real examples of image edits from real-world online scenarios, such as on platforms like Twitter and Weibo. However, most of these edited images lack corresponding original images, making it challenging to obtain precise annotations of the edited regions. Therefore, we selected edited images from X (Twitter) and Weibo

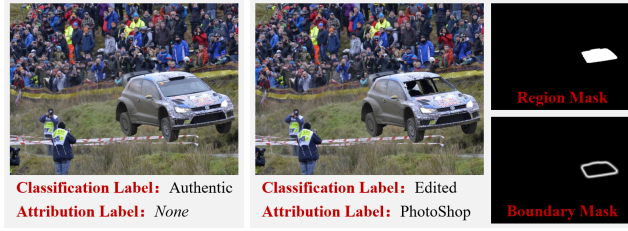


Figure 2: Annotation provided in the GRE dataset.

that have accurate corresponding original images, and we derived annotations for the edited regions through pixel-level difference calculations. Given the limited quantity of such accurately paired images, they have been incorporated as a minor component for evaluation. Additionally, we manually edited 200 images using Photoshop, which has recently introduced Generative AI functionality, to simulate real-world online scenarios. These examples are showcased in Figure 9 and Figure 10.

### A.3 Annotation

For each image in the GRE dataset, we provide multiple annotations, including a two-way classification label (authentic or edited), an n-way classification label for the image (authentic or specific editing methods), and a binary segmentation mask (authentic region or edited region). Furthermore, several methods in the field of image manipulation detection have validated that supervision on the boundaries of edited regions can enhance detection performance. Therefore, inspired by the boundary sliding mechanism in SAFL-Net [30], we derived soft boundary masks.

### A.4 Cases

In the manuscript, images edited using the designed pipeline are showcased, and both the edited images and their corresponding original images are cropped around the edited regions to save space and emphasize the effects of these edits. Now, to facilitate more direct observation of the diversity in the proportion of the edited region to the overall image and the diversity in the locations of these edited regions, the data without crop operations in the dataset are displayed in the sequence of (original image, edited region, edited image). It should be noted that for ease of layout and presentation, vertical images have been rotated by -90 degrees, but no such data augmentation is performed in the GRE dataset.

Figures 6, 7 and 8 showcase different types of edits implemented using various generative regional editing methods within the dataset: removal, replacement, and creation, respectively. These cases also clearly illustrate the types of region masks mentioned in the pipeline's area selection step, irregular object region masks and regular non-object region masks. Depending on the type of edit required, an appropriate type of region mask is selected. This mask defines the region where the edit occurs; in other words, only the pixels within this region are modified, while the pixels in the rest of the image remain unchanged.

Additionally, Figure 9 displays images that are manually edited using Photoshop software, which has recently incorporated Generative AI functionality. Figure 10 also includes images collected

Table 2: Results of cross-dataset evaluation under the pixel-level edited region localization task.

Method	Training Dataset	Testing Dataset (Pixel-level F1)					
		CASIA	DEFACTO	NIST16	IMD20	GRE	Avg.
Unet-Eb4	CASIA	51.8	19.6	21.4	19.5	11.0	24.7
	DEFACTO	5.3	63.2	4.8	3.7	2.4	15.9
	GRE	25.6	23.5	30.3	22.6	66.9	33.8
	Mixed#1	14.7	64.1	4.3	4.4	5.2	18.5
	Mixed#2	45.9	60.5	23.0	18.7	9.4	31.5
	Mixed#3	9.6	61.2	3.2	5.1	53.7	26.6
	Mixed#4	48.1	61.0	29.7	24.3	63.6	45.3
MVSS-Net	CASIA	44.7	25.1	26.3	22.2	16.5	27.0
	DEFACTO	7.9	54.9	4.3	4.1	1.7	14.6
	GRE	23.0	19.4	21.2	22.5	51.6	27.5
	Mixed#1	16.1	59.8	6.1	3.5	3.0	17.7
	Mixed#2	42.7	53.3	25.4	19.4	10.9	30.4
	Mixed#3	8.8	56.2	5.6	5.0	41.7	23.5
	Mixed#4	45.6	54.4	23.9	22.6	49.1	39.1
SAFL-Net	CASIA	48.2	15.2	24.0	21.6	9.8	23.8
	DEFACTO	6.1	60.5	4.9	3.0	2.7	15.4
	GRE	21.8	20.5	28.8	19.8	62.2	30.6
	Mixed#1	11.5	62.4	5.8	6.0	7.5	18.6
	Mixed#2	42.9	57.0	25.1	19.7	9.8	30.9
	Mixed#3	8.1	57.1	5.2	5.7	48.8	25.0
	Mixed#4	44.7	59.3	26.9	20.1	60.9	42.4

from real-world online scenarios on platforms such as X (Twitter) and Weibo. These examples serve to demonstrate the practical applications of these editing tools in producing visually compelling content while also highlighting the challenges of detecting such edits in images sourced from real-world, socially shared media.

## B GRE ANALYSIS

### B.1 Necessity

Existing image tampering detection datasets primarily focus on traditional types of image manipulations. Only a few datasets pay attention to manipulations performed using generative models, and the range of included generative models is very limited. In our manuscript, to demonstrate that existing datasets fail to effectively encompass the types of generative regional editing, as well as to highlight the distinctions between traditional tampering types and generative tampering types, we organize cross-dataset experiments. Among the datasets commonly used for training image tampering detection methods, we select two representative datasets: CASIA (v1&v2), which contains only traditional tampering types, and DEFACTO, which includes traditional tampering types as well as generative tampering types implemented using GAN. In contrast, GRE encompasses tampered images edited through a variety of generative editing methods.

Observations from the experiments in our manuscript reveal that models trained on the GRE dataset significantly improved the detection capabilities for generative regional edited images, which aligns with our initial purpose for constructing the GRE dataset. However, since the GRE dataset does not include other types of tampering, it is challenging to provide generalization capabilities for detecting images with other tampering types. To further demonstrate the necessity of the GRE dataset, we evaluated the generalizability

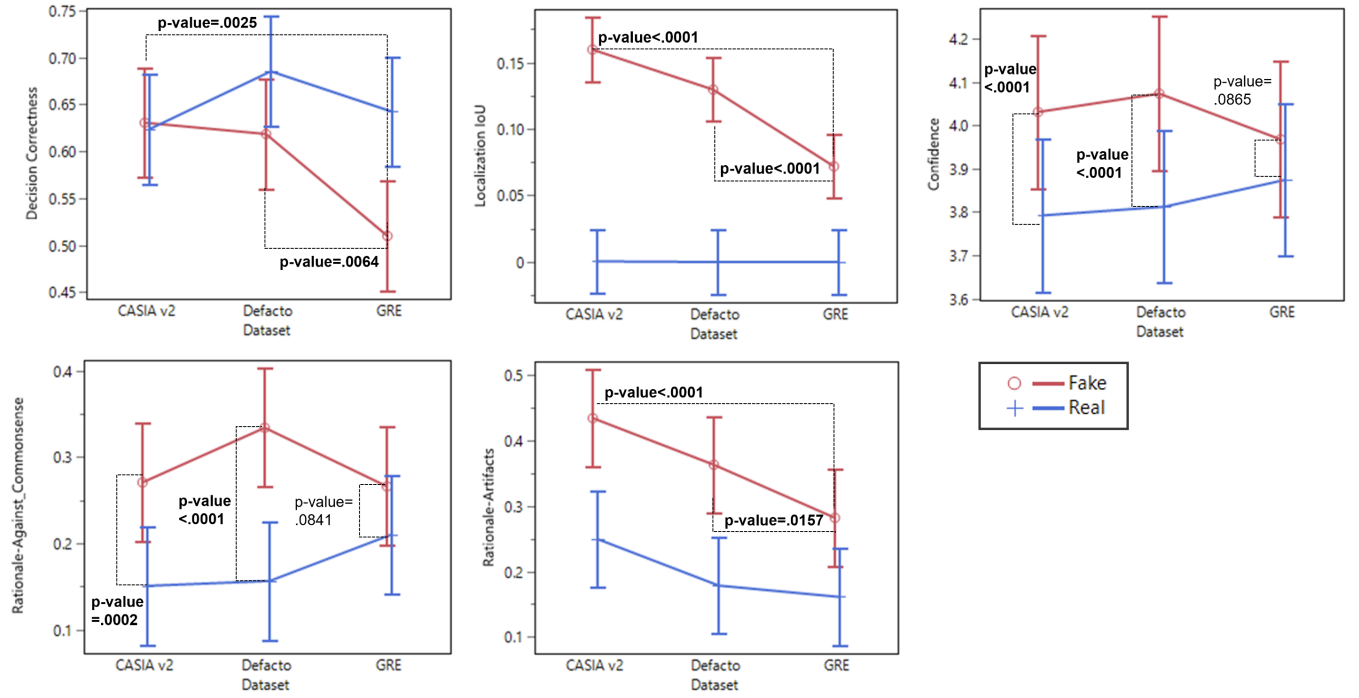


Figure 3: User-study.

of detection models trained under several different dataset combinations, as shown in Table 2. The configurations of the different combinations are as follows:

- Mixed#1: CASIA + DEFACTO
- Mixed#2: CASIA + DEFACTO3k
- Mixed#3: CASIA + DEFACTO + GRE
- Mixed#4: CASIA + DEFACTO3k + GRE

In Mixed#2 and Mixed#4, DEFACTO3k refers to a random sampling of 3,000 tampered images from the complete DEFACTO dataset. This approach was taken into account because the DEFACTO dataset is large but has a uniform operation of tampering, which can easily lead to model overfitting. The experiments have proven that random sampling is a reasonable approach that benefits model training. The Mixed#4 combination, which includes all three datasets, performed well across various test datasets. This robust performance strongly supports the necessity of the GRE dataset as it fills the current gap in detecting generative editing images.

## B.2 Rationality

A key objective in designing the entire editing pipeline is to ensure the edited images are reasonable and realistic. We conducted a controlled user study to compare the GRE dataset with existing datasets on whether the regional edited images are easily noticeable by humans.

**B.2.1 User study task.** We designed a user decision task that classifies regional edited images from authentic ones. The procedure of the user study is as follows. After giving consent and answering demographic questions, firstly, we presented the participants with

a tutorial on examples of different types of regional edited images including object creation, replacement, copy-move, and removal. There were also examples of methods to distinguish edited images such as image editing artifacts and image semantics against commonsense. Then, the participants were asked to decide whether an image was edited by making a binary choice. The participants also reported their confidence in the decision with a 5-point Likert scale question. To understand how humans distinguish edited images from real ones, we collected the participants' rationales with a multiple-choice question. They could select from editing artifacts, image content against logic or commonsense, both of the above, neither of the above, or specifying other reasons. Finally, if participants considered an image as an edited fake image, they would draw on the image to localize the regions where they believed edited.

**B.2.2 Experiment treatment.** The experiments focus on one primary independent variable, Dataset, which has three conditions: CASIA v2, Defacto, and GRE. We select CASIA v2 and Defacto to compare with GRE because they are widely used large-scale regional editing image training sets whose editing pipelines are logical by replacing objects with similar semantics. Defacto has the most generative edited image samples among all existing datasets. We treat Dataset as a within-subjects design, thus each participant sees images from all different datasets. We randomly sampled 100 regional edited images and their corresponding original authentic images from each dataset adding up to 600 images tested in the user study. The study has 50 trials. In each trial, a participant sees one image randomly selected from the 600 images. Any pair of an edited image and its corresponding real image will not be

shown to the same participant. We checked the participants' attention and asked if they would take a break every ten trials. As for the dependent variables, we measured the Decision Correctness, Decision Confidence, Rationale, and Localization Correctness. We calculate the Decision Correctness with the image-level binary user decision and the ground truth label. The Decision Confidence is the self-reported 5-level confidence score. The Rationale includes Against Common Sense and Artifacts, which are binary indicators. The Localization Correctness is calculated by the Intersection over Union (IoU) between the pixel-level user-annotated mask and the ground truth image editing mask. Note that for real images or the images that users misclassified as real, the IoU is zero. The study takes about 25 minutes and the compensation is at least 30 CNY (4.2 USD) and at most 20 CNY (2.8 USD) bonus incentives for better classification performance. The experiment was implemented as an online Qualtrics survey and approved by Institutional Review Boards. The participants are recruited through online social media groups of a university.

**B.2.3 User study results.** We collected 35 complete responses after filtering out a few participants failing the attention checks. There are 16 women and 19 men aged from 18 to 52. Their education backgrounds are one high school, 11 undergraduates, 18 master's degrees, and five doctorate degrees. As for their expertise in image editing, four of them are experts very familiar with image editing, 16 of them are familiar with image editing and have used it before, 13 of them know about image editing and have seen some edited images before, and two of them do not know about image editing before. Each image is tested at least twice.

For the image-level Decision Correctness, when the image is edited (Fake), the human decision correctness is significantly lower on the GRE dataset (50.99%) than the CASIA v2 (63.05%) and the Defacto (61.85%) dataset, with p-values 0.0025 and 0.0064, respectively. It is also harder for humans to classify edited images than real images. Although the decision correctness is similar to a random guess for the GRE dataset, participants' confidence in their decision has no significant differences for all three datasets with confidence levels of 3.9 out of 5 points. An interesting finding is that when the participant's decision on the image is Fake, their confidence is significantly higher than their confidence when the decision is Real with a p-value smaller than 0.0001. The self-report confidence suggests that identifying a fake image is subjectively easier than a real one for the participants, but the decision correctness objectively suggests converse decision difficulties. The confidence gap between Fake and Real decisions is only significant for CASIA v2 (p-value<0.0001) and Defacto (p-value<0.0001), but not for the GRE dataset (p-value=0.0865). This indicates that identifying fake images in the GRE dataset may not be as simple as the other two. The results of the Rationale further prove this point. For the Rationale Against Common Sense, the gap between Fake and Real images is significant for both CASIA v2 (p-value=0.0002) and Defacto (p-value<0.0001), but no significant difference for GRE (p-value=0.0841). CASIA and Defacto both introduced noticeable abnormal image semantics onto their original images. For the Rationale Artifacts, the GRE dataset has significantly fewer artifacts in the fake images than the CASIA v2 (p-value<0.0001) and is marginally significantly fewer than Defacto (p-value=0.0157). The pixel-level

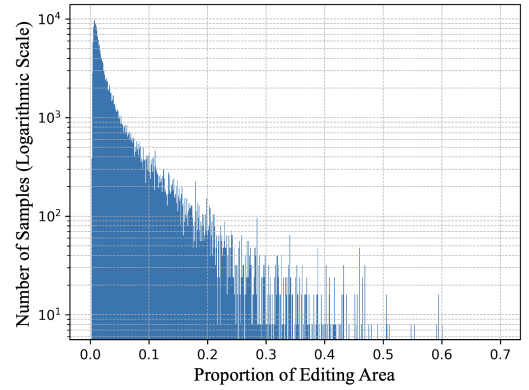


Figure 4: Statistic of editing area proportion.

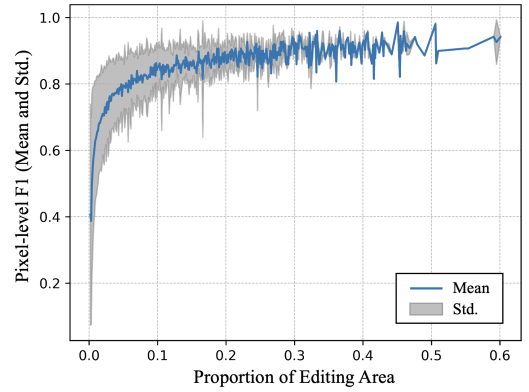


Figure 5: Correlation between detection model performance and editing area proportion.

Localization IoU also shows that our GRE is much harder for humans to identify the editing region. The IoU of the GRE dataset is significantly lower than CASIA v2 and Defacto (p-value=0.0001). In summary, the user study results demonstrate that our GRE dataset makes it harder for humans to distinguish regional edited images on both image- and pixel-level. Our pipeline introduced fewer artifacts and irrational image content than existing popular datasets CASIA v2 and Defacto.

### B.3 Diversity

The proportion of the editing area to the total image area is another factor that influences the performance of detection models. Figure 4 presents the statistic from the GRE dataset on the proportion of the editing area in the total image area, highlighting the diversity of our dataset in this aspect.

Additionally, we utilize a baseline model, Unet with EfficientNet-B4 to analyze the correlation between the proportion of the editing region and the detection performance. Figure 5 visually demonstrates this correlation and suggests that smaller areas of edits indeed increase the difficulty of detection. However, due to the diversity of the GRE dataset in this regard, the models trained on this dataset did not show a strong correlation between the proportion



of the editing region to the total image area and detection performance. This indicates that the GRE dataset can effectively provide the model with the ability to detect edited regions of different sizes, thereby enhancing its robustness and generalization.

C LIMITATION AND FUTURE RESEARCH

In the editing pipeline designed for this study, large models such as ChatGPT, which are equipped with safety limitations, cannot generate editing ideas with malicious intent. As a result, while the final dataset does not exhibit biases associated with the edited entities, this restriction might impact the detection of malicious editing scenarios. Considering that malicious edits occur in real-world contexts and can have more significant consequences and impacts, future work should focus on how to label the intent behind edited images and distinguish between malicious tampering and non-malicious image edits. Additionally, exploring the analysis of editing intent and its impact on enhancing model detection performance will be crucial. This approach could lead to more robust detection methods capable of identifying not only the fact of manipulation but also the underlying intent, which is vital for media security.

The pipeline designed in the paper is extendible, and future enhancements could consider using more advanced language models and multimodal models to generate editing ideas. This would support generative regional editing methods in producing images that are not only visually realistic but also context-coherent. By integrating these sophisticated models, the pipeline could further improve the quality and diversity of the edits, thereby increasing the performance of detection models on more complex and subtle image manipulations in real-world scenarios.

Additionally, with the emergence of various generative regional editing methods, there is potential to enhance the diversity of the data further. Manually edited images, including those modified using image editing software and those collected from real-world online scenarios, are relatively scarce. However, such images are crucial for validating model detection capabilities in real-world contexts. It is essential to gather more data involving unknown generative methods to ensure that detection models are robust and effective across a broad range of editing techniques encountered in actual applications. This effort would help create a more comprehensive dataset.

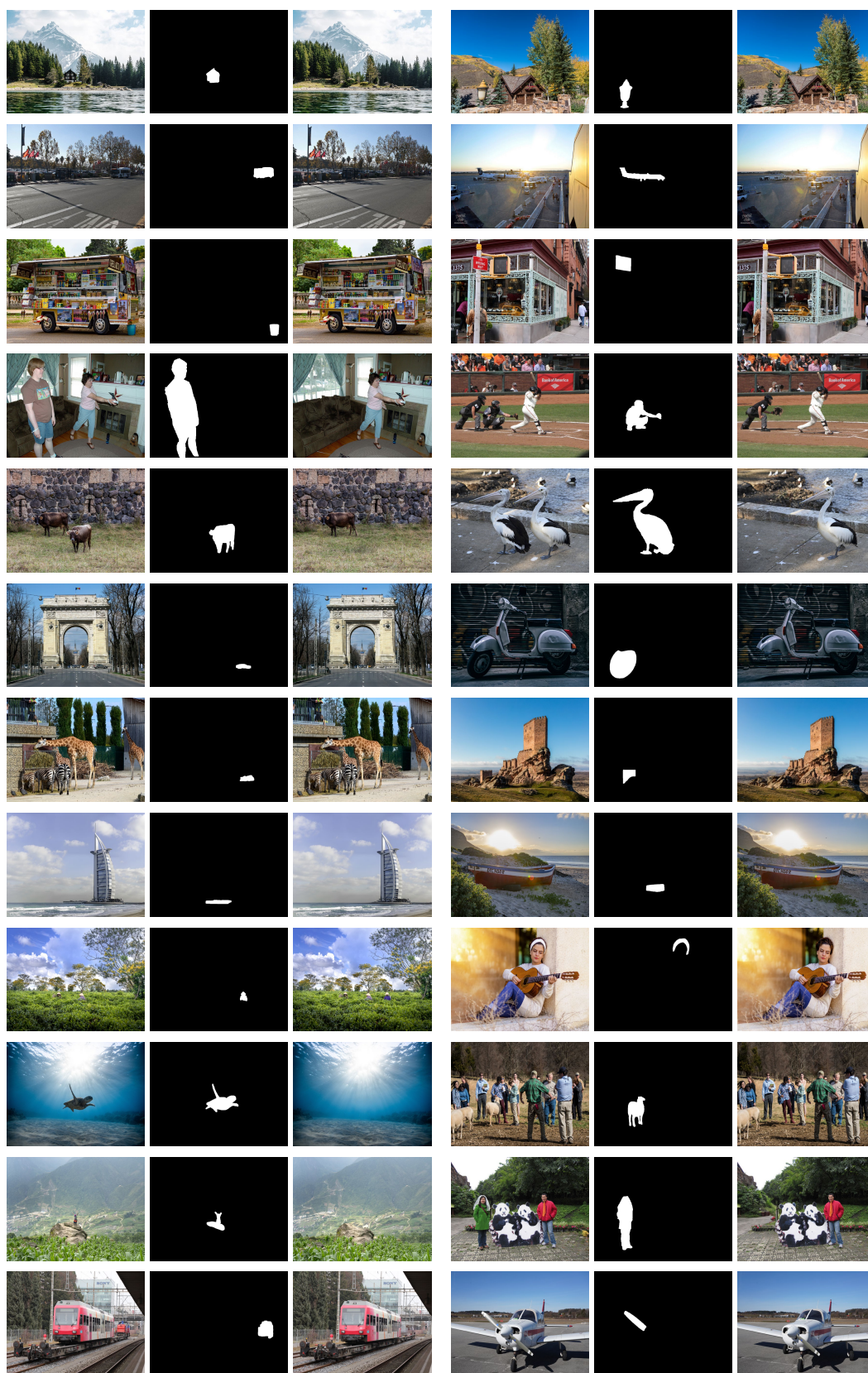


Figure 6: Removal cases in the GRE dataset.



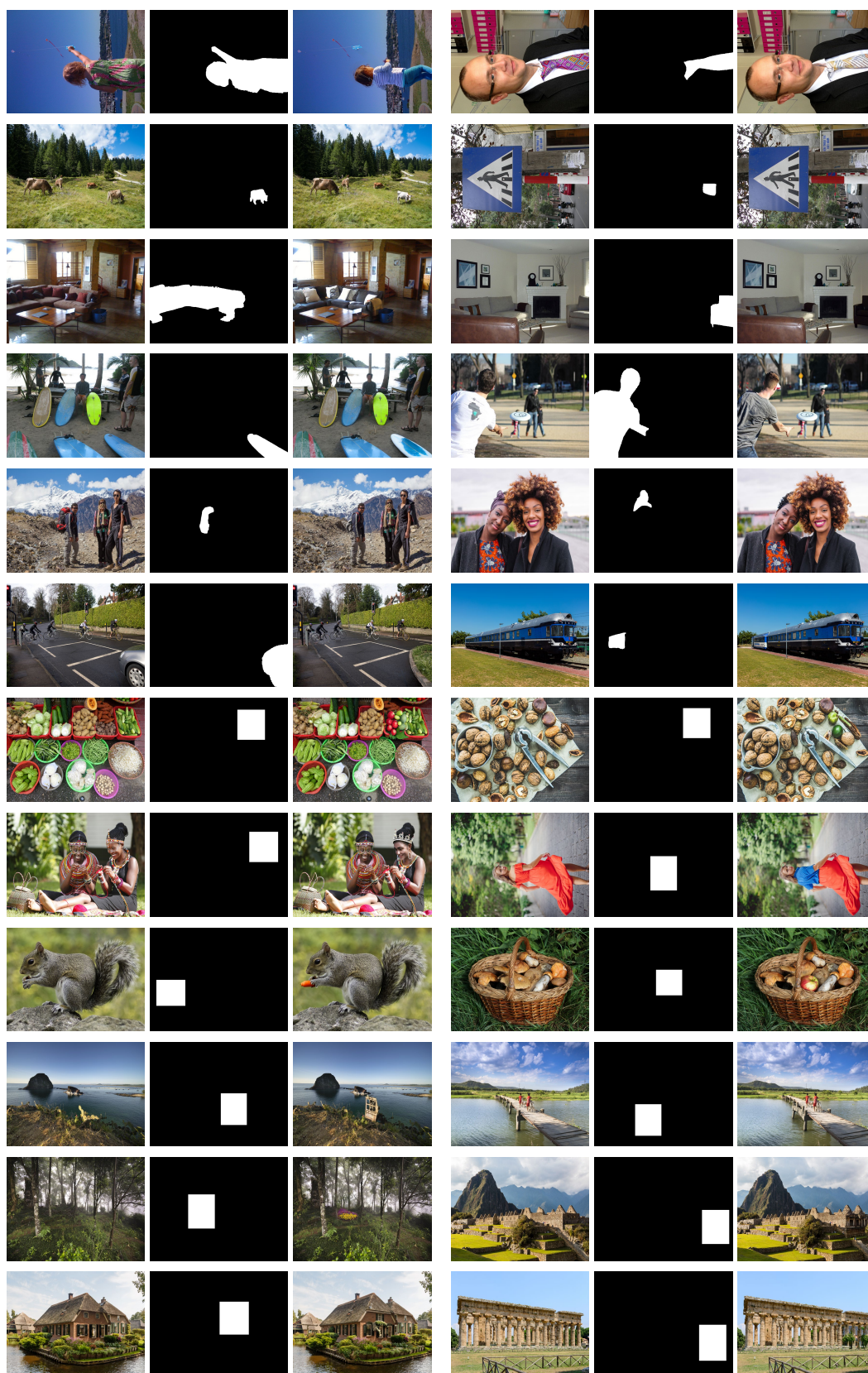


Figure 7: Replacement cases in the GRE dataset.



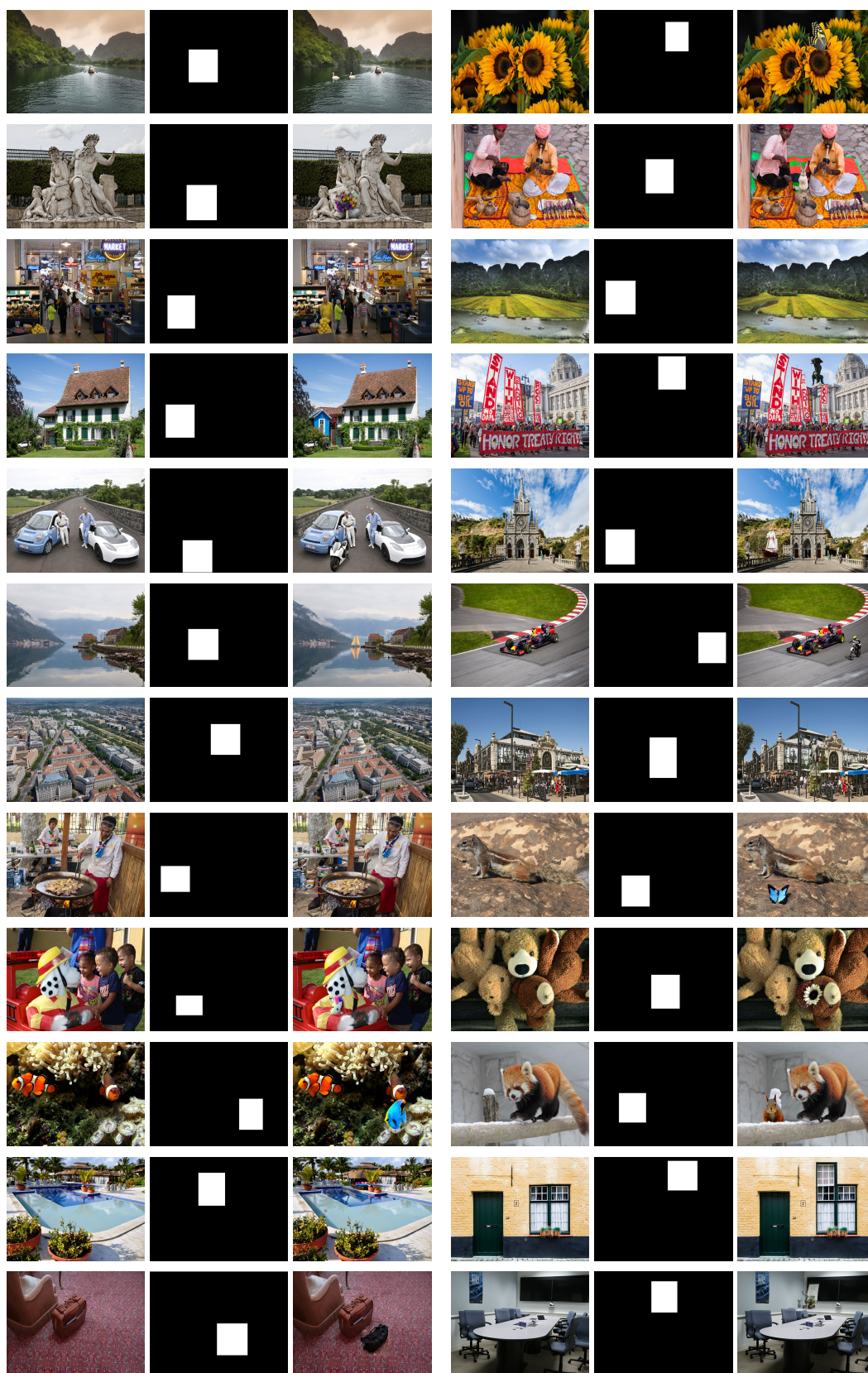


Figure 8: Creation cases in the GRE dataset.





Figure 9: Cases edited by PhotoShop in the GRE dataset.

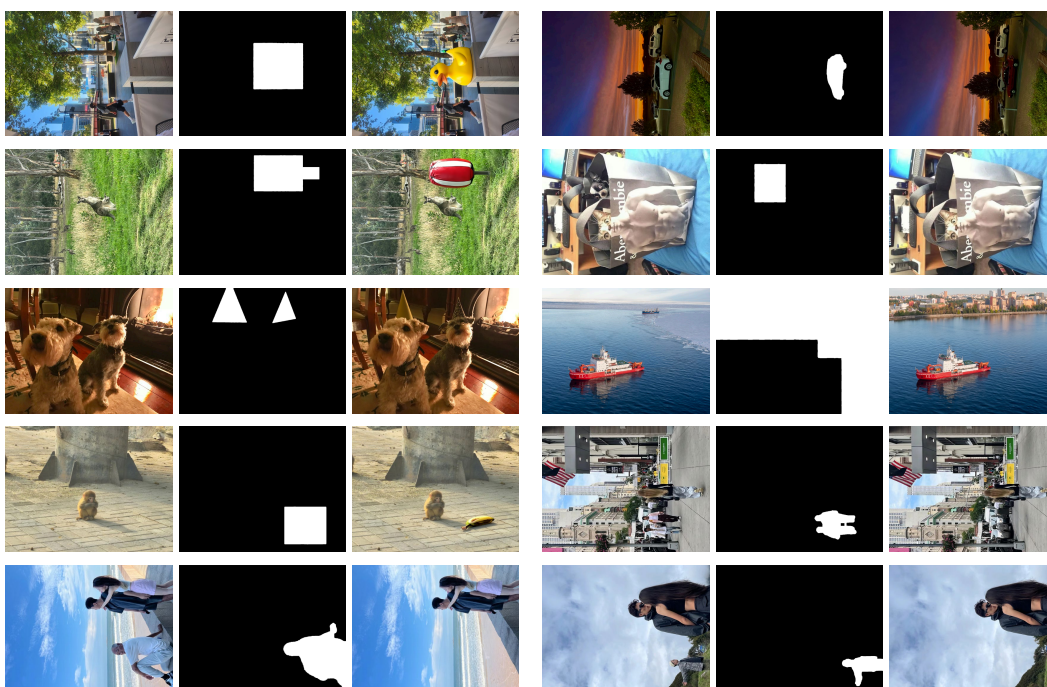


Figure 10: Cases collected online in the GRE dataset.