Journal on
Information Security

RESEARCH Open Access

TGIF2: extended text-guided inpainting forgery dataset and benchmark

Hannes Mareen[1*], Dimitrios Karageorgiou[2], Paschalis Giakoumoglou[2], Peter Lambert[1], Symeon Papadopoulos[2] and Glenn Van Wallendael[1]

Abstract

Generative AI has made text-guided inpainting a powerful image editing tool, but at the same time a growing challenge for media forensics. Existing benchmarks, including our text-guided inpainting forgery (TGIF) dataset, show that image forgery localization (IFL) methods can localize manipulations in spliced images but struggle in fully regenerated (FR) images, while synthetic image detection (SID) methods can detect fully regenerated images but cannot perform localization. With new generative inpainting models emerging and the open problem of localization in FR images remaining, updated datasets and benchmarks are needed. We introduce TGIF2, an extended version of TGIF, that captures recent advances in text-guided inpainting and enables a deeper analysis of forensic robustness. TGIF2 augments the original dataset with edits generated by FLUX.1 models, as well as with random non-semantic masks. Using the TGIF2 dataset, we conduct a forensic evaluation spanning IFL and SID, including fine-tuning IFL methods on FR images and generative super-resolution attacks. Our experiments show that both IFL and SID methods degrade on FLUX.1 manipulations, highlighting limited generalization. Additionally, while fine-tuning improves localization on FR images, evaluation with random non-semantic masks reveals object bias. Furthermore, generative super-resolution significantly weakens forensic traces, demonstrating that common image enhancement operations can undermine current forensic pipelines. In summary, TGIF2 provides an updated dataset and benchmark, which enables new insights into the challenges posed by modern inpainting and AI-based image enhancements. TGIF2 is available at https://github.com/IDLabMedia/tgif-dataset.

Keywords Image forensics, Forgery detection, Forgery localization, Synthetic image detection, AI-generated image detection, Super resolution

1 Introduction

Digital image manipulation has become increasingly accessible and efficient, producing more realistic forged images with recent advances in generative AI (GenAI) technology [1]. A few years ago, creating convincing image edits required expertise in software such as Adobe Photoshop and significant manual effort. More recently, deepfake technology [2] enabled realistic face swaps and modifications, but these were typically limited to faces and required technical skills. Today, modern open-source and commercial GenAI models, such as Stable Diffusion (SD) [1], Adobe Firefly [3], and FLUX [4], can perform high-resolution edits that are often indistinguishable from authentic images. Additionally, these GenAI models are accessible to users without technical expertise as they work by simple text prompting [1, 5]. Such capabilities have positive applications in creative industries, image restoration, and education. However, they also pose serious risks, including fraud, disinformation, and the creation of fake evidence [6].

Among GenAI approaches, text-guided inpainting [1, 5] is particularly powerful. It allows users to select a region

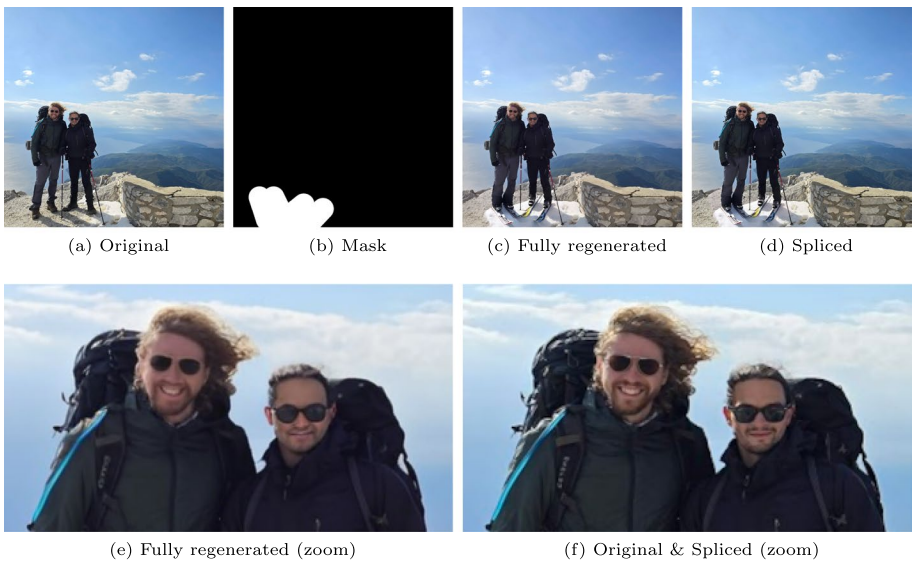
*Correspondence:
Hannes Mareen
hannes.mareen@ugent.be
[1] IDLab, Ghent University – imec, Gent, Belgium
[2] Information Technologies Institute, CERTH, Thessaloniki, Greece

of an (authentic) image and provide a textual prompt describing the desired edit. Diffusion-based models then regenerate the image so that the edited region matches the prompt, while leaving the rest of the image visually intact. Depending on the workflow, the edited region may be spliced into the original image (i.e., only changing pixels in the selected region) or the model may regenerate the entire image (i.e., potentially changing all pixels in the image, yet only semantically altering the selected region). This is demonstrated in Fig. 1. In previous work [7], we showed that when regenerating the entire image during editing, the forensic traces that traditional image forgery localization (IFL) methods use are destroyed. Moreover, our previous work has found that synthetic image detection (SID) methods can detect fully regenerated images as synthetic, but they are not designed to localize manipulated regions.

The Text-Guided Inpainting Forgery (TGIF) dataset proposed in our previous work [7] provided high-resolution images manipulated with three text-guided inpainting models: SD2 [1], SDXL [8], and Adobe Firefly [3]), including both spliced and fully regenerated images. While TGIF provided a valuable dataset, benchmark and insights, recent developments in generative models require an updated, more comprehensive dataset. First, new generative inpainting models were released since. Most notably, FLUX.1 [4] is a new generation of diffusion model that replaces the traditional U-net and diffusion pipeline of traditional diffusion models (such as SD) with a transformer and flow-matching framework. As such, FLUX.1 achieves higher-quality local and global generations [9]. Second, the previously acquired insight that generative edits significantly impact IFL models requires further exploration for potential solutions. This highlights the need for a more comprehensive benchmark and experiments that deepen our understanding of the impact of recent GenAI inpainting models on forensic methods.

To this end, we introduce TGIF2, an extended version of TGIF, with several new contributions:

- **FLUX inpainting:** TGIF was limited to using SD2, SDXL, and Adobe Firefly to inpaint images. TGIF2 extends the TGIF dataset by additionally including FLUX.1 models ([schnell], [dev], and Fill [dev]). The FLUX.1 models are a new generation of diffusion models, and are among the most capable open-source text-guided inpainting models available today [9]. In contrast to traditional SD models, they utilize flow-matching instead of denoising diffusion and rely on large multimodal transformers rather than a U-net for better contextual reasoning.
- **Random non-semantic masks:** in addition to semantic, object-centric masks, TGIF2 includes randomly positioned, non-semantic masks, allowing us to evaluate potential biases of IFL methods toward objects or other semantically meaningful regions.



(a) Original (b) Mask (c) Fully regenerated (d) Spliced

(e) Fully regenerated (zoom) (f) Original & Spliced (zoom)

Fig. 1 Examples of the two ways of inpainting. An **a** authentic image can be inpainted in a **b** selected region as mask, with *skis* as prompt. In the GenAI-based inpainting process (here using SDXL), **c** the full image is regenerated during editing. To minimize artifacts, **d** only the region corresponding to the mask can be spliced into the authentic image. **e** and **f** provide zoomed-in versions to more clearly show the differences between the original or spliced pixels, on the one hand, and the regenerated pixels outside of the masked area, on the other. Subtle differences due to the regenerative process can be observed in the teeth, sunglasses, beard, etc.

- **New experimental contributions:** we provide new results for IFL methods fine-tuned on TGIF2's fully regenerated images, the relation of the IFL performance with generative quality, and the effect of super resolution (SR) [10] as an attack against IFL and SID methods.

These additions make TGIF2 not only a larger dataset but enable new insights into forensic challenges. First, the inclusion of FLUX.1 inpainted images allows evaluation of how current IFL and SID methods generalize to newer generative models. Second, the provided random, non-semantic masks may reveal potential biases of IFL methods toward objects or semantically meaningful regions, highlighting limitations in their robustness. For example, we demonstrate that IFL models that are fine-tuned on non-random, semantic FR data enable localization in FR images. However, our evaluation on images inpainted with random, non-semantic masks reveals moderate semantic biases. Lastly, the super-resolution attack exposes the risk that routine GenAI-based image enhancement can erase traces used for forensic analysis. In summary, TGIF2 provides a more comprehensive dataset and evaluation framework that exposes current limitations and guides the development of more effective forensic methods for the detection and localization of realistic, text-guided inpainting manipulations.

The remainder of this paper is organized as follows. Section 2 reviews related work in generative media forensics and existing datasets. Section 3 presents the TGIF2 dataset, including the authentic data sources, inpainting models, workflow, and resulting subsets. Section 4 reports our forensic benchmark, covering the performance of existing IFL and SID methods, fine-tuning experiments, the generative quality, and the impact of super-resolution attacks. Finally, Section 5 discusses our findings and concludes the paper.

2 Related work

This section briefly discusses recent advances in IFL and SID methods, in Sects. 2.1 and 2.2, respectively. Then, Sect. 2.3 discusses related datasets.

2.1 Image forgery localization

Image forgery localization aims to identify the regions of an image that have been manipulated. Approaches in this area can be roughly grouped into two categories. Some methods focus on detecting subtle statistical inconsistencies, for example traces of compression [11] or sensor noise patterns [12]. Others exploit more visible cues, such as unnatural transitions along the boundary between pristine and altered content [13]. To improve robustness, several works combine multiple forensic signals within a unified framework [14–16]. Some existing research also considers forgeries created with inpainting techniques. Specialized IFL methods in this setting include detectors that leverage distortions in the Laplacian domain [17], or that learn discriminative features directly from data [18, 19]. These approaches, however, have largely been outperformed by more general methods designed to work across different manipulation types [20, 21].

Current state-of-the-art models [13, 20–25] are mostly based on deep neural networks and are trained on datasets covering a wide variety of manipulations, such as splicing, copy–move, or conventional inpainting. Related to inpainting, PSCC-Net [20] was partly trained on manipulated data using a conventional inpainting method [26], and MantraNet [24] was partly trained on images manipulated with OpenCV's inpainting method. More datasets that include inpainted images are discussed in Section 2.3. While training on a variety of manipulations makes models broadly applicable, these existing models have not been optimized for the recent case of text-guided image manipulation.

A complementary line of work has examined laundering attacks, where generative models erase forensic traces while leaving the semantic content unchanged [27]. These results demonstrate that generative compression, or diffusion-based regeneration in general, can serve as effective laundering mechanisms, substantially weakening IFL methods. The TGIF2 dataset proposed in Section 3 directly addresses these challenges by including fully regenerated manipulated content, thereby enabling systematic evaluation of IFL robustness in these challenging scenarios. Additionally, the impact of generative super resolution is evaluated in the TGIF2 benchmark, in Section 4.6.

2.2 Synthetic image detection

Synthetic image detection methods aim to distinguish between authentic and AI-generated visual content. The field initially developed alongside the rise of generative adversarial network (GAN)-based image synthesis, and many early detectors were trained specifically on GAN outputs [28–34]. With the rapid adoption of (latent) diffusion models (LDM or DMs), such as stable diffusion (SD) [1], more recent work has focused on this paradigm or explores detectors that can generalize across both GANs and DMs [35–40]. As these methods are trained and evaluated on academic datasets, in practice, their performance often decreases when evaluated on images found in the wild. However, their performance can be significantly improved by leveraging in-the-wild content for training [41].

Most approaches rely on the idea that generative models leave behind subtle traces or fingerprints, which can be captured in various feature domains. The type of

representation is critical for generalization. For instance, some detectors trained purely on GAN-generated data are still able to identify DM-generated images, highlighting the persistence of shared artifacts across generative families [30, 33, 34].

Despite this progress, most SID research assumes a binary scenario where an entire image is either authentic or synthetic. This setting overlooks more subtle manipulations, such as text-guided inpainting, where a forged region is blended with pristine content. In such cases, SID methods produce a global score but cannot localize the edited area. Moreover, authentic images that are post-processed using generative models without altering their semantics have been shown to trigger SID methods [27, 42].

The TGIF2 dataset introduces subsets where manipulated content is confined to local regions but fully regenerated through advanced diffusion-based editing. This allows us to examine not only the limits of SID methods, but also their vulnerability to laundering-style manipulations during generative editing. Moreover, the impact of generative super resolution on SID methods is evaluated in Section 4.6.

2.3 Text-guided inpainting datasets

A wide range of datasets has been developed for image forensics, covering both IFL and SID, as summarized in several recent surveys [6, 43]. Many of these resources focus on specific manipulation types such as splicing, copy-move, or deepfakes, while others aim to provide more diverse manipulations. For inpainting, earlier datasets such as NIST16 [44] and DEFACTO [45] include localized edits, but these were created with non-AI-based inpainting tools and do not involve text guidance.

Early text-guided inpainting datasets Recently, several datasets with text-guided inpainted images have been released. First, COCO-Glide [21] includes 512 images inpainted with GLIDE [5]. While valuable as a proof of concept, COCO-Glide is relatively small and limited to low-resolution crops (256×256 px) from MS-COCO [46]. Additionally, the Autosplice dataset [47] contains approximately 3 k inpainted spliced images using DALL-E 2. Most importantly, both datasets do not account for differences between spliced and fully regenerated edits.

Fully regenerated inpainted images TGIF [7] presents approximately 75 k images manipulated using three text-guided inpainting techniques (SD2, SDXL, and Adobe Firefly) at higher resolutions (up to 1024×1024 px), sourced from approximately 3 k images from MS-COCO. TGIF also differentiates between spliced and fully regenerated images, which was one of the key contributions of the work. After the release of TGIF, several new datasets on text-guided inpainting were proposed. Most notably, SAGI [48] provides approximately 95 k manipulated images sourced from 3 datasets (MS-COCO, RAISE [49], and OpenImages [50]), with resolutions up to 2048×2048 px. SAGI uses advanced pipelines (HD-Painter, BrushNet, PowerPaint, ControlNet [51], and Inpaint-Anything) to semantically alter images, based on SD2 (a DM) and LaMa [52] (a GAN-based approach). To the best of our knowledge, SAGI is the only other existing dataset (i.e., apart from TGIF) that differentiates between spliced and fully regenerated images.

Other AI-based inpainting datasets COinCO [53] provides approximately 97 k manipulated images from COCO, using SD2 and out-of-context prompts. Furthermore, the GRE dataset [54] presents approximately 228 k images inpainted using six generative editing methods and pipelines (LaMa [52], MAT [55], SD2, ControlNet [51], PaintByExample, and Adobe Photoshop). The UniAIDet [56] benchmark includes a dataset of 10k inpainted images, only considering splicing. None of these inpainting datasets consider local manipulation using new generative models such as FLUX.1. To the best of our knowledge, local manipulation with FLUX.1 was only included in the OpenSDI [57] dataset, which additionally includes SD1.5, SD2.1, SDXL, and SD3, to comprise approximately 150 k images that are either globally synthesized or locally manipulated. Note that they do not consider local manipulation with full regeneration.

Going beyond object replacement to limit bias All of the above datasets inpaint objects in images, which we hypothesize may introduce a potential bias when using them for training or evaluation. There are few datasets that go beyond object replacement, namely Beyond the Brush (BtB) [58], GIM [59], and BR-Gen [60]. In BtB [58], images are inpainted using Fooocus, which internally uses SDXL. Objects are automatically segmented from images, and a vision-language model is used to generate appropriate inpainting prompts (spanning from replacing the object with a new object type, to removing it altogether). Next, GIM [59] is a million-scale dataset of inpainted images, either by replacing objects using SD2 and GLIDE or removing them using the Denoising Diffusion Null-Space Model (DDNM) [61]. Finally, BR-Gen [60] presents approximately 150k locally forged images, including not only global and main object inpaintings, but also secondary object and background inpaintings, therefore accounting for salient object bias. As generative editing methods, they use LaMa [52] and MAT [55] as GAN-based approaches and SDXL [8], BrushNet [62], and PowerPaint [63] as diffusion-based

pipelines. Bias is an important challenge in dataset design. Recent studies [40, 64] have shown that forensic models can exploit shortcuts or artifacts that are specific to a dataset rather than to the underlying manipulation process. This not only inflates reported performance but also limits cross-dataset generalization. Designing datasets that minimize such biases is therefore critical for robust training and evaluation. Hence, in our work, we carefully inspect potential semantic biases, originating from training on object-only replacement.

Summary of text-guided inpainting datasets An overview of these datasets with generative inpainted images is given in Table 1, where we focus on spliced vs. fully regenerated images, the used generative models and pipelines, as well as the presence of a subset that goes beyond object replacement. In summary, no dataset exists that covers all three of the following aspects. First, only OpenSDI includes local manipulation with recent generative models such as FLUX [4], which has established itself as one of the strongest diffusion-based editors. Second, only TGIF and SAGI explicitly differentiate between spliced and fully regenerated images, whereas providing localization in the latter is still an open problem. Third, few datasets (i.e., only BtB, GIM, and BR-Gen) consider potential object-based biases and provide non-object-based generative inpainted images. The lack of a dataset that addresses all these gaps has motivated the creation and release of the TGIF2 dataset.

3 TGIF2 dataset

The TGIF2 dataset is an extension of the original TGIF dataset [7]. In general, compared to existing datasets in Table 1, the TGIF2 dataset is the only dataset that has all three of the following properties: (1) includes newer generative models (i.e., FLUX.1 models), (2), explicitly differentiates between spliced and fully regenerated images, and (3) goes beyond object replacement (by additionally using random, non-semantic masks). TGIF2 is available for download at https://github.com/IDLabMedia/tgif-dataset.

The remainder of this section is organized as follows. Section 3.1 describes the source of authentic images and their properties. Section 3.2 details the inpainting models used to generate manipulations. Section 3.3 explains the automated workflow for generating the variety of inpainted images, ending with an overview of the different subsets in Table 2. Finally, Section 3.4 quantifies the dataset size and splits.

3.1 Source of authentic images

TGIF2 uses the same source of authentic images as TGIF, namely the MS-COCO dataset [46] (*val2017*), which is licensed under a Creative Commons Attribution 4.0 License. COCO provides images with associated captions, segmentation masks, and object bounding boxes, covering 80 object categories. To comply with restrictions in Adobe Firefly, we excluded two categories (*knife* and, surprisingly, *frisbee*) that were not allowed as prompts in the app. From each object category, a maximum of 50 images were selected. Images were filtered to ensure both dimensions are at least 512 pixels, while the free Flickr service limited the dimensions to 1024 px each. We additionally required object bounding boxes to be smaller than 512×512 px and larger than 64^2 px in area. This ensures sufficient resolution for high-fidelity inpainting while avoiding excessively small objects that may produce ambiguous manipulations. These selection and filtering criteria ensure that TGIF2 builds on a diverse yet

Table 1 Overview of datasets with generative inpainted images. SP = spliced, FR = fully regenerated

Name	SP	FR	# fake images	Gen. models and pipelines	Beyond obj. repl.
Coco-Glide [21]	✓	–	512 k	GLIDE	–
AutoSplice [47]	✓	–	3 k	DALL-E2	–
TGIF [7]	✓	✓	75 k	SD2, SDXL, Adobe Firefly	–
SAGI [48]	✓	✓	95 k	SD2, LaMa, HD-Painter, BrushNet, PowerPaint, ControlNet, Inpaint-Anything	–
COinCO [53]	✓	–	97 k	SD2	–
GRE [54]	✓	–	228 k	LaMa, MAT, SD2, ControlNet, PaintByExample, Photoshop	–
UniAIDet [56]	✓	–	10 k	SD2, SDXL, SD3, DreamShaper, PaintByExample	–
OpenSDI [57]	✓	–	150 k	SD1.5, SD2.1, SDXL, SD3, FLUX.1	–
BtB [58]	✓	–	22 k	Fooocus (SDXL)	✓
GIM [59]	✓	–	1.1 M	SD, GLIDE, DDNM	✓
BR-Gen [60]	✓	–	150 k	LaMa, MAT, SDXL, BrushNet, PowerPaint	✓
TGIF2 (ours)	✓	✓	271 k	SD2, SDXL, Adobe Firefly, FLUX.1 schnell/dev/fill-dev	✓

well-controlled set of authentic images suitable for high-quality and consistent inpainting experiments.

3.2 Inpainting models

TGIF2 includes several inpainting models to cover a wide range of generative capabilities. Most notably, the three FLUX.1 model variants are new in TGIF2, compared to TGIF [7]. Note that the images inpainted with PS, SD2, and SDXL are exactly equal in TGIF and TGIF2, with the exception of those with random non-semantic masks (as explained further in Section 3.3).

- **Adobe firefly (PS)**: commercial model accessed through Photoshop 25.4.0, in early 2024. This produces spliced inpainted regions with an automatically added gradient border.
- **Stable diffusion 2 (SD2)**: a first-generation open-source diffusion model. We used the inpainting-fine-tuned model fine-tuned (*stabilityai/stable-diffusion-2-inpainting*).
- **Stable diffusion XL (SDXL)**: a second-generation open-source diffusion model that supports higher-resolution outputs (i.e., up to 1024×1024 px) and improved visual quality. We used the inpainting-fine-tuned model (*stabilityai/stable-diffusion-xl-1.0-inpainting-0.1*).
- **FLUX.1 models**: a third-generation open-source inpainting architecture, offering improved attention mechanisms and context-aware generation. Specifically, three FLUX.1 variants are included:
  - **FLUX.1 [schnell]**: model capable of generating high-quality images in only 1 to 4 diffusion steps. It is trained using latent adversarial diffusion distillation (based on the closed-source FLUX.1 [pro]), hence the optimized speed comes at the cost of a decrease in image quality.
  - **FLUX.1 [dev]**: model optimized to balance speed with image quality. Trained using guidance distillation (based on the closed-source FLUX.1 [pro]), making FLUX.1 [dev] more computationally efficient yet producing slightly lower quality images.
  - **FLUX.1 Fill [dev]**: same architecture as FLUX.1 [dev], but fine-tuned for the inpainting task.

Note that FLUX.1 [schnell] and FLUX.1 [dev] were trained for full synthetic image generation, in contrast to FLUX.1 Fill [dev] and the SD models which were fine-tuned for the inpainting use case. However, these full-image generation models can be plugged into the inpainting workflow without any adaptations, and perform (surprisingly) well for this use case. In fact, on average, they produce higher quality inpaintings than FLUX.1 Fill [dev] (see Section 4.4). Using the FLUX.1 [schnell] model is especially practical, as it requires much fewer diffusion steps than FLUX.1 Fill [dev] and hence has significantly lower computational complexity (as further discussed in Section 3.3).

It should additionally be noted that FLUX.1 [pro] is available as well, although only through a pay-per-use API. Closed-source models behind APIs do not allow easy reproducibility by the community, and they only provide limited control over their inputs, making them unsuitable for an in-depth forensic analysis. Therefore, we did not include this generative model in our dataset.

Figure 2 shows a side-by-side comparison of an example authentic image, the bounding box used as mask, and six inpainted images generated by the six GenAI models, respectively. Note that, in this example, Fig. 2c, e (inpainted using SD2 and PS, respectively) are spliced, whereas the other inpainted images are fully regenerated. Additionally, a bounding box was used as mask. Together, these models span commercial and open-source approaches across three generations of diffusion

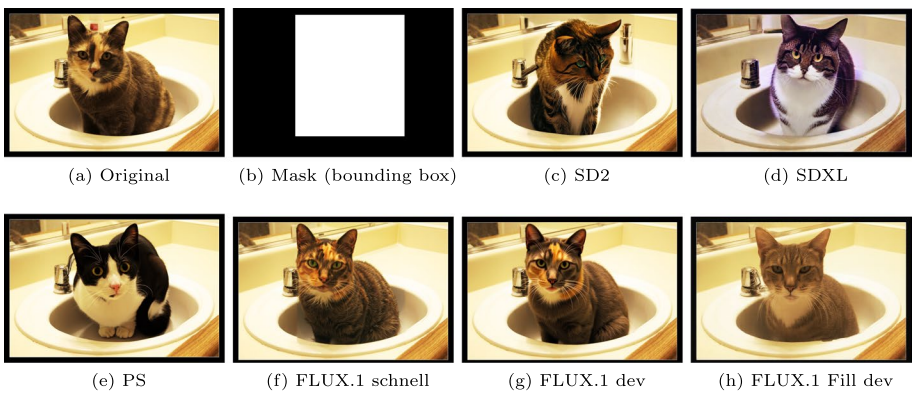


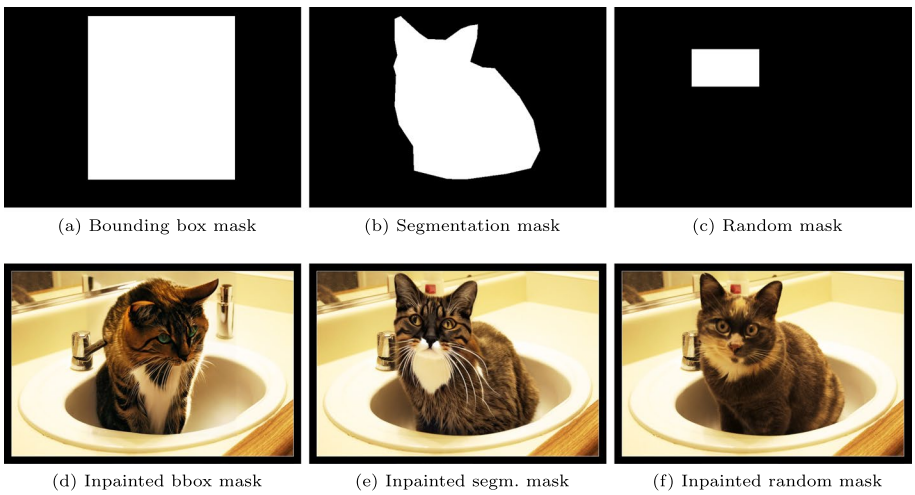
Fig. 2 Side-by-side comparison of **a** a real image of a cat with **b** the mask used for inpainting, and **c–h** 6 inpainted versions using 6 different inpainting models used in TGIF2, respectively. Note that other masks can be used as well (see Fig. 3)

architectures, enabling TGIF2 to capture a broad spectrum of inpainting characteristics.

3.3 Inpainting workflow

The inpainting workflow is exactly the same for all inpainting models, with the only difference being the generative model that is called. In particular, for each authentic image, inpainted variants are automatically generated following these steps:

- **Mask selection**: for each selected image, we use either the object's semantic segmentation mask, the object's bounding box (bbox) mask (both provided by COCO), or a random non-semantic mask. The random, non-semantic masks are at a random location in the (cropped) image, and with a random rectangular size (min. 64 px and max. 60% of the image's width and height). Examples of these three mask types and corresponding inpainted versions are shown in Fig. 3.
- **Prompt generation**: for SD2, SDXL and FLUX.1, the prompt combines the object category and the image caption (provided by COCO). During initial experiments, we noticed that including the caption resulted in better results, which was also observed in related work [65, 66]. For Adobe Firefly, only the object category is used, as we noticed that this was sufficient in our initial experiments. For the random, non-semantic mask, an empty prompt is used.
- **Generation parameters**: inference steps, guidance scale, and seed were randomly selected per image. For each model and each mask, three variations were generated with three different seeds. The number of inference steps is between 10 and 50 for the SD and FLUX models, and between 3 and 6 for FLUX.1 [schnell]. The guidance scale is between 1 and 10, except for FLUX.1 [schnell] where it is set to 0 (as it does not support this parameter).
- **Splicing vs. full regeneration**: we save up to two variants of each inpainted image, namely a fully regenerated (FR) and a spliced (SP) version. The differences between these two are illustrated in Fig. 1 and explained below.
  - *Full regeneration (FR)*: when inpainting using generative models, the entire input image is used to condition the diffusion process. While only the masked area is semantically changed, the non-masked area is subtly altered as well (as visualized in Fig. 1).
  - *Splicing (SP)*: after performing full regeneration, the inpainted region is inserted (or *spliced*) back into the original image using the provided mask. This approach preserves the rest of the original image and follows the strategy used in common inpainting pipelines, such as Adobe Firefly and HuggingFace diffusers [67]. This is also the most commonly used inpainting method in existing datasets (see Section 2.3).
- **Generative quality metrics**: each image also includes aesthetic quality scores (NIMA [68], GIQA [69]) and image-text-matching (ITM) scores [70], measured on the object's inpainted area. Additionally, we include the preservation fidelity, calculated using the peak signal-to-noise ratio (PSNR), Structural SIMilarity index measure (SSIM), and the LPIPS perceptual



(a) Bounding box mask (b) Segmentation mask (c) Random mask

(d) Inpainted bbox mask (e) Inpainted segm. mask (f) Inpainted random mask

Fig. 3 The **a–c** three types of masks used and **d–f** three corresponding inpainted examples, respectively. All inpainted images used Fig. 2a as input image

metric [71] between the real and FR image, in the real but fully regenerated areas.

Each of the steps above are applied on each inpainting model, with some exceptions and notes. First, SD2 is limited to processing images with resolution 512×512, which is the resolution of the fully regenerated images in the corresponding subsets. In contrast, for SDXL and the FLUX.1 variants, the resolution is up to 1024 px. This only affects the fully regenerated variant, and not the spliced variant. Second, Adobe Firefly (Photoshop, PS) only provides a spliced variant, but no full regeneration variant. Third, we do not provide the spliced variant for SDXL, as it discolors the entire image, which is extremely visible when splicing into the original (and a known issue [72]). This discoloration can also be observed when comparing Fig. 2a with 2d. Fourth, we do not generate random masks for Adobe Firefly/PS, since it only produces spliced edits, which is the least challenging use case (as discussed in Section 4.2).

Table 2 gives an overview of the 19 resulting TGIF2 subsets.

3.4 Dataset statistics and splits

Starting from 3124 authentic images, TGIF2 contains 18,744 manipulated images per SP subset or FR subset, and 9372 manipulated images per Random (SP or FR) subset (as only one mask type, i.e., a random rectangular one, is used as opposed to two, i.e., a segmentation mask and bounding box). This amounts to a total of 271,788 manipulated images in TGIF2, spread over the 19 subsets presented in Table 2 (of which 4 subsets containing 74,976 images are already included in the original TGIF dataset). Furthermore, the dataset is split approximately by 80/10/10% into training, validation, and test sets, carefully avoiding category overlap and image leakage between splits. This makes TGIF2 a reliable resource for training and evaluating forensic models.

4 Forensic benchmark

We provide a comprehensive benchmark of image forensic methods on the TGIF2 dataset. Compared to the original TGIF benchmark, TGIF2 enables a more diverse and challenging evaluation by introducing new subsets, new fine-tuned models, and new attack scenarios. Section 4.1 describes the evaluation setup employed in the experiments. Then, in Section 4.2, we evaluate existing IFL methods on the expanded TGIF2 subsets. Section 4.3 investigates whether fine-tuning IFL models on our fully regenerated training sets improves their ability to localize manipulations in FR inpainted images while accounting for potential object biases. Section 4.4 analyzes the generative quality and whether it impacts IFL performance.

Table 2 Overview of TGIF2 dataset subsets grouped by inpainting model. Columns indicate whether spliced (SP), fully regenerated (FR), and semantic or random mask variants are included

Inpainting model	Semantic		Random	
	SP	FR	SP	FR
SD2	✓*	✓*	✓	✓
SDXL	–	✓*	–	✓
Adobe Firefly (PS)	✓*	–	–	–
FLUX.1 [schnell]	✓	✓	✓	✓
FLUX.1 [dev]	✓	✓	✓	✓
FLUX.1 Fill [dev]	✓	✓	✓	✓

* These subsets are already included in the original TGIF dataset

Section 4.5 extends the benchmark of SID methods by including the new TGIF2 subsets and recent state-of-the-art SID methods. Section 4.6 evaluates the generative super-resolution attack scenario as a post-processing step.

4.1 Evaluation setup

We perform evaluations on our TGIF2 test set, separately considering each of the subsets in Table 2. This setup allows us to systematically compare model behavior across different manipulation types and mask strategies.

For IFL methods, as performance metric, we report the mean pixel-level F1 score over manipulated images, using a threshold of 0.5 to decide whether pixels are real or fake. We additionally provide the mean intersection over union (IoU) results in Appendix 1. This provides a calibrated measure of localization accuracy, which is important for practical deployment and human interpretability. We select several recent high-performing AI-based IFL methods, namely the PSCC-Net [20], SPAN [22], ImageForensicsOSN [23], MVSS-Net++ [13], MantraNet [24], CAT-Net (v2) [25], TruFor [21], and MMFusion (MMF) [14]. During evaluation, we specifically distinguish among the semantic (Sem) and the random (Rand) subsets.

For SID methods, we report the area under the ROC curve (AUC) as performance metric. We opt for AUC rather than the threshold-specific F1, since we noticed that some methods tend to overpredict the fake class, artificially inflating F1, whereas the AUC better reflects the trade-off between false positives and false negatives across thresholds. We select several AI-based SID methods available in the SIDBench framework [73], namely CNNDetect [28], DIMD [35], Dire [36], FreqDetect [29], UnivFD [30], Fusing [74], GramNet [31], LGrad [32], NPR [33], RINE [37], DeFake [38], and PatchCraft [34]. Additionally, new in TGIF2, we incorporate the recent

models SPAI [39], SPAI-ITW [41], RINE-ITW [41], and B-Free [40]. By combining a diverse set of recent SID approaches with a systematic evaluation protocol, we provide a comprehensive benchmark of current forensic capabilities on generative image inpainting.

4.2 Image forgery localization

We benchmark recent IFL methods on the TGIF2 dataset. Recall that, compared to the original TGIF benchmark, TGIF2 includes several new subsets (recent FLUX.1 models and random non-semantic masks). Table 3 and the top part of Table 4 report the F1 values for all evaluated IFL methods on the spliced subsets and on the fully regenerated subsets, respectively. We highlight F1 values above 0.7 in bold font, chosen empirically as a threshold to allow for quickly grasping which methods demonstrate good detection performance.

For the non-random Sem spliced subsets, some IFL methods (i.e., CAT-Net, TruFor, and MMFusion) show decent performance. This was already reported in the original TGIF work [7], and is in line with our expectations, as the splicing creates significant differences in forensic traces between forged and pristine regions. When evaluating the Sem FLUX.1 subsets, we notice that CAT-Net, TruFor, and MMFusion still perform well, although there is a performance drop on the latter.

We also consider inpainted images with random masks, for which the results can be observed in the Rand subsets of Table 3. Compared to the corresponding Sem subsets, we see a significant drop in performance for MMFusion and TruFor, especially on the FLUX.1 [schnell] and [dev] subsets. In contrast, CAT-Net remains highly effective in detecting the spliced region in the Rand subsets, as well as TruFor and MMFusion on the SD2 subset. This

Table 3 Evaluation of image forgery localization methods on the spliced (SP) subsets of our TGIF2 dataset, for both the semantic (Sem) and random (Rand) versions (F1 score). F1 scores above 0.7 are highlighted in bold

IFL method	SD2		PS	FLUX.1						Average		
				[schnell]		[dev]		Fill [dev]				
	Sem	Rand	Sem	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	All
PSCC-Net [20]	0.15	0.17	0.39	0.03	0.09	0.02	0.05	0.08	0.21	0.13	0.13	0.13
SPAN [22]	0.00	0.00	0.00	0.00	0.00	0.0	0.00	0.00	0.01	0.00	0.00	0.00
ImageForensicsOSN [23]	0.23	0.09	0.36	0.31	0.11	0.23	0.08	0.12	0.10	0.25	0.09	0.18
MVSS-Net++ [13]	0.07	0.03	0.08	0.15	0.06	0.11	0.03	0.09	0.09	0.10	0.05	0.08
Mantranet [24]	0.15	0.15	0.56	0.06	0.02	0.04	0.02	0.06	0.04	0.17	0.06	0.12
CAT-Net [25]	**0.89**	**0.92**	**0.86**	**0.88**	**0.93**	**0.87**	**0.92**	**0.86**	**0.93**	**0.87**	**0.92**	**0.89**
TruFor [21]	**0.83**	**0.87**	**0.79**	**0.79**	0.69	**0.73**	0.58	**0.74**	**0.72**	**0.77**	**0.72**	**0.75**
MMFusion [14]	**0.73**	**0.74**	**0.72**	**0.70**	0.59	0.59	0.45	0.63	0.59	0.68	0.59	0.64

Table 4 Evaluation of image forgery localization methods on the fully regenerated (FR) subsets of our TGIF2 dataset, for both the semantic (sem) and random (rand) versions (F1 score). The last rows show the IFL methods fine-tuned on the FR subsets. F1 scores above 0.7 are highlighted in bold

IFL method	SD2		SDXL		FLUX.1						Average		
					[schnell]		[dev]		Fill [dev]				
	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	All
PSCC-Net [20]	0.05	0.04	0.05	0.09	0.02	0.02	0.02	0.02	0.03	0.05	0.03	0.04	0.04
SPAN [22]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ImageForensicsOSN [23]	0.20	0.09	0.18	0.09	0.30	0.08	0.23	0.07	0.09	0.05	0.20	0.08	0.13
MVSS-Net++ [13]	0.06	0.03	0.09	0.05	0.15	0.03	0.12	0.03	0.07	0.04	0.10	0.04	0.07
Mantranet [24]	0.03	0.01	0.05	0.06	0.05	0.02	0.04	0.01	0.02	0.01	0.04	0.02	0.03
CAT-Net [25]	0.04	0.01	0.03	0.00	0.05	0.01	0.04	0.01	0.03	0.02	0.04	0.01	0.02
TruFor [21]	0.19	0.07	0.18	0.09	0.33	0.09	0.25	0.06	0.16	0.08	0.22	0.08	0.14
MMFusion [14]	0.15	0.05	0.19	0.08	0.34	0.08	0.20	0.04	0.12	0.06	0.20	0.06	0.13
TruFor – fine-tuned	0.49	0.39	0.68	**0.83**	**0.87**	**0.94**	**0.82**	**0.88**	**0.76**	**0.85**	**0.72**	**0.78**	**0.75**
MMFusion – fine-tuned	0.48	0.39	0.66	**0.83**	**0.86**	**0.94**	**0.81**	**0.87**	**0.75**	**0.84**	**0.71**	**0.77**	**0.74**

suggests that MMFusion and TruFor likely suffer from a detection bias towards object boundaries or semantic cues. The limitations exposed in this subsection should be taken into account when deploying these methods in real-world detection systems.

In contrast to spliced images, the FR setting is much more challenging, as was already observed in the original TGIF benchmark [7]. These insights are confirmed in the new FR FLUX.1 subsets, for which the average F1 scores are still low (i.e., well below 0.5). This is the case for both the Sem and Rand subsets. In FR images, the clear boundary between forensic traces in manipulated and authentic regions disappears due to the regenerative process of the entire image during manipulation.

In summary, while some recent state-of-the-art IFL methods perform reliably on spliced manipulations across generative editing models, they struggle on fully regenerated images. The random-mask experiments further reveal that some methods, such as MMFusion and TruFor, may rely on semantic or object-related biases rather than imperceptible forensic traces. Together, these findings highlight the need for developing IFL approaches that generalize beyond semantic object boundaries, and that remain effective in fully regenerated images. To further delve into this issue, Section 4.3 evaluates whether we can fine-tune IFL methods in order to detect FR images.

4.3 Fine-tuning IFL approaches on fully regenerated images

To address the poor performance of existing IFL methods on FR images (see Section 4.2 and Table 4), we fine-tune selected IFL models on the TGIF2 FR training subsets (i.e., SD2, SDXL, Flux.1 [schnell], Flux.1 [dev], and Flux.1 Fill [dev], both with semantic and random masks). We do this for each subset individually (but combining the three FLUX.1 subsets into one main FLUX.1 subset), as well as for all subsets collectively. In the collective fine-tuning, we sample an equal number of images from SD2, SDXL and FLUX.1 to avoid a bias to FLUX.1, since it has three times more images than SD2 or SDXL. As base models, we select TruFor and MMFusion, i.e., two of the IFL models with the best performance on spliced and fully regenerated images. CAT-Net was excluded from fine-tuning experiments due to its significantly more complex training process. We followed the training setups specified in the original papers for both models. That is, both models follow a two-stage training procedure. In the first stage, the model is trained on the forgery localization task for 100 epochs while in the second stage, part of the model is frozen and is trained on the forgery detection task for another 100 epochs. Then, the checkpoint with the best validation loss is chosen as the final checkpoint. We maintain the hyperparameters and other training configurations specified in the original papers. This fine-tuning approach allows us to adapt high-performing IFL models to better localize manipulations in fully regenerated images, while remaining consistent with the original training protocols.

Table 5 reports the F1 scores for the fine-tuned models. In short, we observe that fine-tuning significantly improves the performance. However, performance gains are substantially larger when training and testing on semantically aligned masks than when evaluated on random-mask FR images. This is discussed more elaborately in the remainder of this section.

First, we consider the finetuned models on all semantic datasets. The performance on the Sem datasets is relatively high, but is significantly lower for the random datasets (with the exception of SDXL). We demonstrate this in Fig. 4, which shows an image inpainted (and fully regenerated) using a semantic mask as well as with a random mask, along with the corresponding ground-truth masks and localization results from TruFor finetuned on the SD2-FR-Sem subset. The fine-tuned model is able to localize the semantic object (i.e., the tennis racket), but not the random mask in the background. Note that the inpainted random box, in fact, made notable changes to the sign in the background, and hence should be detectable by a better detection model Additionally, note that the object categories do not overlap between the training and test set, i.e., no images of inpainted tennis rackets are seen in the fine-tuned training set. This indicates that training exclusively on semantic datasets can induce a bias toward salient or semantically meaningful regions.

Second, including the Rand subsets during fine-tuning mitigates the semantic bias observed in the previous setting. In fact, additionally including random masks during finetuning improves performance on the random subsets without degrading—and in several cases even improving—performance on the semantic subsets. We select the fine-tuned models on All Sem+Rand subsets as the main one, and also include these results in the bottom rows of Table 4.

Third, beyond mask-related effects, we also observe a discrepancy in IFL performance between SD2, SDXL, and the FLUX.1 family. That is, the IFL performance of FLUX.1 is significantly higher than SD2. This is especially notable, since we notice contrasting behavior when evaluating the IFL performance on spliced images in Table 3. Note that we do not only observe this when training on All subsets, but also when training on each subset individually. We hypothesized that this discrepancy could be explained by their difference in generative quality, although find no supporting evidence of this, in Section 4.4. It remains to be investigated, in future work, why

Table 5 Evaluation of individually fine-tuned image forgery localization methods on the fully regenerated (FR) subsets of our TGIF2 dataset, for both the semantic (Sem) and random (Rand) versions (F1 score). F1 scores above 0.7 are highlighted in a bold font. We notice that in-domain (i.e., same generative method) performance is relatively good, but generalization to out-of-domain (i.e., different generative method) performance is bad. Interestingly, when fine-tuning on only Sem masks, the evaluation on Rand subsets is lower than the corresponding Sem subsets, suggesting a potentially introduced bias towards semantics

IFL method	Fine-tune set		SD2		SDXL		FLUX.1						Average		
							[schnell]		[dev]		Fill [dev]				
			Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	All
TruFor	All	Sem+Rand	0.49	0.39	0.68	**0.83**	**0.87**	**0.94**	**0.82**	**0.88**	**0.76**	**0.85**	**0.72**	**0.78**	**0.75**
		Sem	0.50	0.23	0.61	0.62	**0.80**	0.62	0.69	0.46	0.65	0.59	0.55	**0.76**	0.65
		Rand	0.31	0.40	0.50	**0.80**	0.69	**0.92**	0.62	**0.85**	0.63	**0.82**	0.55	**0.76**	0.58
	SD2	Sem	0.49	0.22	0.29	0.17	0.40	0.16	0.31	0.08	0.21	0.10	0.34	0.14	0.24
		Rand	0.31	0.40	0.17	0.22	0.12	0.10	0.09	0.06	0.12	0.12	0.16	0.18	0.17
	SDXL	Sem	0.09	0.01	0.66	**0.71**	0.19	0.02	0.08	0.01	0.06	0.02	0.22	0.15	0.19
		Rand	0.01	0.00	0.50	**0.82**	0.02	0.00	0.01	0.01	0.03	0.03	0.11	0.17	0.14
	Flux	Sem	0.10	0.02	0.20	0.05	**0.90**	**0.90**	**0.86**	**0.81**	**0.78**	**0.79**	0.57	0.51	0.54
		Rand	0.02	0.02	0.05	0.06	**0.73**	**0.95**	0.68	**0.91**	0.68	**0.84**	0.43	0.56	0.49
MMFusion	All	Sem+Rand	0.48	0.39	0.66	**0.83**	**0.86**	**0.94**	**0.81**	**0.87**	**0.75**	**0.84**	**0.71**	**0.77**	**0.74**
		Sem	0.47	0.17	0.58	0.61	**0.80**	0.58	**0.72**	0.50	0.69	0.67	0.65	0.50	0.63
		Rand	0.26	0.40	0.48	**0.80**	0.62	**0.91**	0.58	**0.83**	0.62	**0.81**	0.51	**0.75**	0.58
	SD2	Sem	0.52	0.22	0.28	0.19	0.35	0.15	0.27	0.09	0.18	0.09	0.32	0.15	0.23
		Rand	0.30	0.34	0.18	0.19	0.08	0.05	0.07	0.03	0.09	0.08	0.14	0.14	0.14
	SDXL	Sem	0.08	0.01	0.66	**0.70**	0.12	0.01	0.06	0.01	0.05	0.02	0.19	0.15	0.17
		Rand	0.01	0.01	0.50	**0.80**	0.03	0.00	0.02	0.01	0.04	0.04	0.12	0.17	0.15
	Flux	Sem	0.10	0.02	0.20	0.05	**0.89**	**0.87**	**0.85**	**0.76**	**0.78**	**0.80**	0.56	0.50	0.53
		Rand	0.03	0.02	0.06	0.07	**0.73**	**0.95**	0.69	**0.90**	0.69	**0.84**	0.44	0.56	0.50

FLUX.1 fully regenerated images are easier to detect than SD, after fine-tuning.

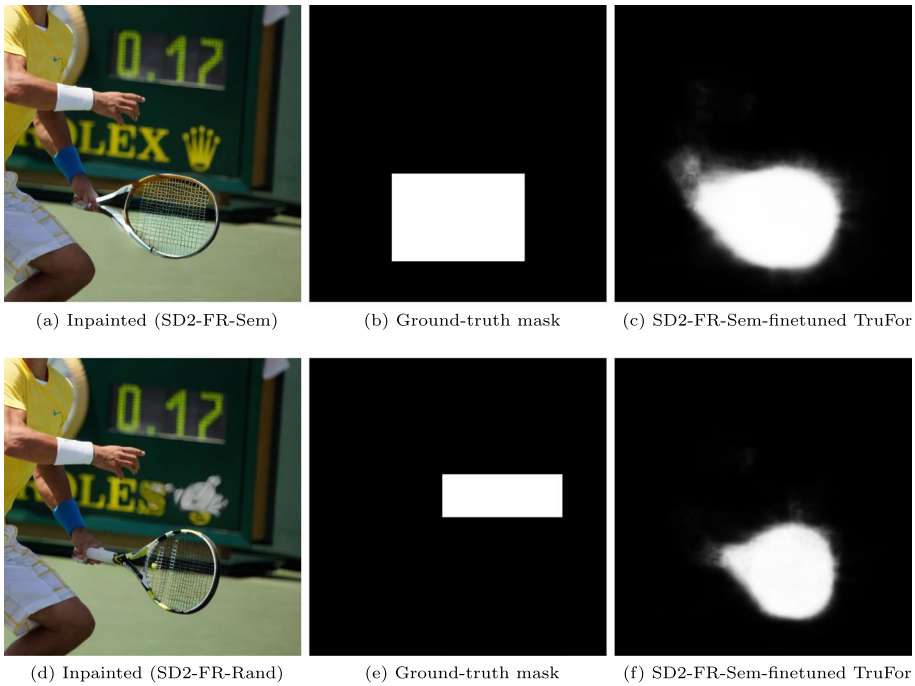
Fourth and finally, the models fine-tuned on individual subsets expose that performance on other, out-of-domain generative models is substantially lower than on the in-domain subset that they were trained on. Note that these methods were originally designed for splicing and copy-move detection. Hence, an improved training strategy alone may not suffice. Instead, more robust localization mechanisms may be required for localization in fully regenerated images with better generalization.

In summary, this section demonstrated that inpainted areas can be detected in fully regenerated images, which is in line with similar observations in related work that trained or fine-tuned models perform well on FR image data [48, 75]. However, fine-tuning exclusively on semantic subsets can induce object-level bias, underscoring the importance of incorporating non-semantic FR data during both training and evaluation.

4.4 Generative quality vs. IFL performance

This subsection investigates whether generative quality is predictive of forensic localization performance. Intuitively, higher-quality or more realistic generations might be expected to suppress forensic traces and therefore reduce IFL performance. Conversely, systematic artifacts in lower-quality generations may facilitate detection. We therefore analyze whether aesthetic quality, image–text alignment, and preservation fidelity correlate with IFL performance across subsets and on a per-image level. We additionally examine whether such quality differences help explain the performance discrepancy between SD and FLUX in fine-tuned FR settings (as observed in Section 4.3).

As mentioned at the end of Section 3.3, we measure the generative quality metrics in three ways. First, we measure the aesthetic quality using the NIMA [68] and GIQA [69] metrics on the inpainted area. Second, we measure the image-text matching performance [70] on the inpainted area. Third, we measure the preservation fidelity in real but fully regenerated areas using PSNR, SSIM, and LPIPS. Finally, we relate these generative quality metrics to the IFL performance in two ways: (1) analyzing trends across the subsets, and (2) we calculate the Spearman's correlation coefficients between per-image F1 scores and each generative quality metric. We have



(a) Inpainted (SD2-FR-Sem) (b) Ground-truth mask (c) SD2-FR-Sem-finetuned TruFor

(d) Inpainted (SD2-FR-Rand) (e) Ground-truth mask (f) SD2-FR-Sem-finetuned TruFor

Fig. 4 An example of a fully regenerated inpainted image with SD2, using a bounding box of the tennis racket as mask (**a**, SD2 Sem), or a random mask in the image (**d**, SD2 Rand), along with the corresponding ground-truth masks (**b**, **e**) and the fine-tuned TruFor detection results (**c**, **f**). The fine-tuned TruFor was trained on the SD2-FR-Sem training set. The inpainted tennis racket in the SD2-FR-Sem test image is detected relatively well by the fine-tuned model. However, the inpainted random box of the SD2-FR-Rand subset is *not* detected; instead, the tennis racket is wrongly detected—suggesting that the fine-tuned model is be biased towards semantics or salient objects. Note that the inpainted random box, in fact, made notable changes to the sign in the background, and hence should be detectable by a better detection model

done this analysis for the TruFor and MMFusion IFL models, namely for the original variants on the SP subsets, and the fine-tuned variants (on All Sem+Rand) on the FR subsets. Note that we have not done this for the original IFL models evaluated on the FR subsets, as their detection capability is too low in this case (as discussed in Section 4.2).

Table 6 shows the average generative quality metrics for each semantic subset of TGIF2. For the reader's convenience, we again include the IFL performance of TruFor and MMFusion (as presented before in Tables 3 and 4). Additionally, Table 7 shows the correlation values between the generative metrics and the IFL performance, calculated on all semantic subsets simultaneously. Recall, as also observed previously in Sections 4.2 and 4.3, that SD2 has the best and FLUX.1 [dev] has the worst IFL (SP) performance. In contrast, the IFL (FR) performance is best for FLUX.1 [schnell], and significantly worse for SD2. Note that, for the preservation fidelity, it only makes sense to relate it with IFL (FR) but not to IFL (SP) performance. In the following, we inspect whether these performance discrepancies can be explained by any of the generative quality metrics.

Concerning aesthetic quality, in Table 7, on a per-subset level, there appears to be no relation with IFL (SP) nor IFL (FR) performance. In contrast, on a per-image level, in Table 7, we notice there is a weak to moderate correlation (i.e., correlation values around 0.3 for NIMA and 0.4 for GIQA) between the aesthetic quality and IFL performance (both for SP and FR). Interestingly, higher aesthetic scores are weakly associated with better IFL performance at the per-image level, contradicting the intuition that visually appealing generations should be harder to detect. However, since we do not observe this relation on a per-subset level, and hence may not serve as an explanation as to why FLUX.1 performs significantly better than SD2 in IFL (FR).

For ITM, in Table 6, on a per-subset level, we do not observe a correlation with the IFL (SP) nor IFL (FR) performance. On a per-image level, in Table 7, we see a negligible to weak negative correlation for ITM (i.e., approx. -0.3 for IFL (SP) and 0 for IFL (FR)), and a negligible to weak positive correlation for the ITM cosine similarity (i.e., approx. 0.1 for IFL (SP) and 0.2 for IFL (FR)). Hence, we conclude that there is no clear relationship between ITM and IFL performance.

Table 6 Average generative quality per (semantic) subset. In addition, the average detection performance is given as well (same as in Table 4). For each metric, we highlight the best performing subset in a bold font, and the worst performing subset in an italic font

Metric class		Metric	SD2	PS	SDXL	FLUX.1		
						[schnell]	[dev]	Fill [dev]
Aesthetic Quality		NIMA mean↑	5.05	5.16	*4.88*	5.16	**5.28**	5.21
		GIQA↑	−399	−373	−423	−383	**−371**	*−424*
Image-Text Matching		ITM↑	0.28	**0.32**	0.26	0.31	0.30	*0.25*
		cos↑	0.37	**0.38**	*0.34*	**0.38**	**0.38**	0.36
Preservation Fidelity (FR only)		PSNR↑	27.1	N/A	*18.1*	**32.5**	**32.5**	28.5
		SSIM↑	0.79	N/A	*0.69*	**0.92**	**0.92**	0.86
		LPIPS↓	0.04	N/A	*0.17*	0.02	**0.01**	0.06
IFL (SP)	TruFor – original	F1 score↑	**0.83**	0.79	N/A	0.79	*0.73*	0.74
	MMFusion – original	F1 score↑	**0.73**	0.72	N/A	0.70	*0.59*	0.63
IFL (FR)	TruFor – fine-tuned	F1 score↑	*0.49*	N/A	0.68	**0.87**	0.82	0.76
	MMFusion – fine-tuned	F1 score↑	*0.48*	N/A	0.66	**0.86**	0.81	0.75

Table 7 Spearman's correlation and corresponding *p*-value between the IFL performance (F1 score), on the one hand, and the generative quality metrics, on the other. All correlation values are relatively low, with *p*-values equal to 0.000

IFL method	Version	Subsets	Aesthetic quality		ITM		Preservation fidelity		
			NIMA	GIQA	ITM	cos	PSNR	SSIM	LPIPS
TruFor	Original	SP	0.36	0.42	−0.31	0.07	N/A	N/A	N/A
MMFusion	Original	SP	0.30	0.39	−0.29	0.11	N/A	N/A	N/A
TruFor	Fine-tuned	FR	0.28	0.34	−0.03	0.23	0.24	0.28	−0.32
MMFusion	Fine-tuned	FR	0.28	0.35	−0.02	0.24	0.24	0.28	−0.32

For the preservation fidelity, in Table 6, on a per-subset level, we observe no correlation with IFL (FR). On a per-image level, in Table 7, we observe only a weak positive correlation (i.e., approx. 0.2–0.3) with IFL (FR). Hence, we conclude that there is no notable relationship between preservation fidelity and IFL performance.

Overall, generative quality metrics do not meaningfully predict forensic localization performance. The detectability differences between SD and FLUX in FR settings therefore cannot be explained by the evaluated quality measures. However, note that we were limited by the current generative quality metrics. That is, on the one hand, the aesthetic quality metrics used could only be calculated on the inpainted area, but not on the real-but-regenerated area (as they do not allow masked calculation). On the other hand, it only makes sense to calculate the preservation fidelity metrics on the real-but-regenerated area (and not on the inpainted area). A more informative direction for future work may be to measure intra-image quality contrast, i.e., discrepancies between inpainted regions and real-but-regenerated regions. Such contrast-based metrics could better capture subtle inconsistencies exploited by forensic localization models, and hence could potentially explain the performance discrepancy between subsets in the fine-tuned FR setting.

4.5 Synthetic image detection

We benchmark state-of-the-art SID methods on the TGIF2 FR dataset, which now includes FR subsets generated with FLUX.1 models (in addition to only SD2 and SDXL in the original TGIF dataset), as well as both semantic and random-mask variants. Note that we do not report results on spliced subsets, since SID methods are not designed to detect local manipulations and hence consistently fail to detect these cases. Instead, in this section, we focus only on the fully regenerated subsets. We also expanded the benchmark with four recently introduced methods: SPAI [39], SPAI-ITW [41], RINE-ITW [41], and B-Free [40]. Table 8 reports AUC values for all evaluated methods, where we highlight scores above 0.8 in a bold font. This threshold was chosen empirically to allow for quickly grasping which methods demonstrate decent detection performance.

From the 17 evaluated SID models, 11 demonstrate relatively good performance (AUC > 0.8) on either SD2,

SDXL, or both. Specifically, the newly included methods (SPAI, SPAI-ITW, RINE-ITW, and B-Free) all have relatively strong performance on these subsets, with B-Free even scoring 1.00 AUC on both sets.

The evaluation of the new FLUX subsets provides a more challenging picture. Out of the 11 SID methods that perform reliably well on SD2 and SDXL, only 4 maintain an AUC above 0.8 across all FLUX subsets. In general, all methods experience a performance drop on FLUX, suggesting that its generation process produces fewer or less detectable forensic traces. For example, B-Free results in AUC values between 0.85 and 0.88 for the FLUX subsets, while it resulted in an AUC value of 1.00 for the SD2 and SDXL subsets. This finding indicates that FLUX represents a significantly harder detection setting, and highlights the importance of developing SID models that generalize across different families of generative models.

We notice no significant difference in SID performance between semantic and random-mask subsets.

In summary, TGIF2 confirms that modern SID methods can successfully detect fully regenerated images produced by earlier diffusion models such as SD2 and SDXL, but new generative models such as FLUX introduce new challenges and degrade detection performance. However, it should be stressed that binary classification in SID models are not able to localize generative edits.

4.6 Generative super-resolution attack

This subsection studies the robustness of forensic methods against generative super resolution. Specifically, we apply the open-source Real-ESRGAN [10], which is one of the most popular super-resolution methods to date (with over 34,000 stars on GitHub [72]). Additionally, we apply bicubic interpolation as baseline comparison. For the experiments, we apply an increase in resolution with a factor of either 2 or 4, followed by a decrease in resolution with the same factor, in order to obtain an attacked image with the same resolution as the original one. On average, the PSNR between the bicubic up-and-down-sampled image and the original image is 19.2 (regardless of the factor), and the LPIPS is 0.21 and 0.22, for the factor of 2 and 4, respectively. For the super-resolution method, the PSNR is 18.8 (regardless of the factor), and the LPIPS [71] is 0.25 and 0.26, for the factor of 2 and 4, respectively. These attacked images are evaluated using both IFL and SID methods, but restricted to the models that demonstrate strong performance in Sections 4.2 and 4.5. Tables 9 and 10 present the results for IFL methods applied on all non-random SP subsets and SID methods applied on all non-random FR subsets, respectively. For each evaluated IFL or SID model, the tables show the average performance on all subsets in addition to the performance and corresponding delta in performance when applying the bicubic and super-resolution attacks.

Table 8 Evaluation of synthetic image detection methods on the semantic subsets of our TGIF2 Dataset (AUC Score). AUC scores above 0.8 are highlighted in a bold font

SID method	Trained on	SD2		SDXL		FLUX.1						Average		
						[schnell]		[dev]		Fill [dev]				
		Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	All
CNNDetect [28]	ProGAN	0.57	0.48	0.61	0.62	0.50	0.47	0.49	0.50	0.62	0.62	0.56	0.54	0.55
Dire [36]	ADM	0.49	0.52	0.62	0.64	0.49	0.54	0.48	0.52	0.48	0.49	0.51	0.54	0.53
FreqDetect [29]	GANs	0.50	0.42	0.40	0.41	0.39	0.36	0.37	0.38	0.39	0.34	0.41	0.38	0.40
Fusing [74]	ProGAN	0.55	0.48	0.47	0.47	0.55	0.53	0.55	0.56	0.62	0.62	0.55	0.53	0.54
NPR [33]	ProGAN	0.51	0.49	0.67	0.31	0.49	0.50	0.49	0.48	0.44	0.43	0.52	0.44	0.48
DeFake [38]	Diffusion	0.62	0.60	0.61	0.59	0.67	0.66	0.70	0.67	0.63	0.61	0.65	0.63	0.64
DIMD [35]	LDM	0.62	0.60	0.61	0.59	0.67	0.66	0.70	0.67	0.63	0.61	0.65	0.63	0.64
UnivFD [30]	ProGAN	**0.82**	**0.80**	**0.80**	**0.81**	0.63	0.60	0.58	0.59	0.69	0.71	0.70	0.70	0.70
GramNet [31]	GANs	**0.82**	0.79	0.55	0.54	**0.81**	**0.81**	**0.82**	**0.82**	**0.84**	**0.84**	0.77	0.76	0.76
LGrad [32]	ProGAN	**0.85**	**0.85**	**0.83**	**0.80**	**0.84**	**0.84**	**0.85**	**0.86**	**0.86**	**0.84**	**0.85**	**0.84**	**0.84**
RINE [37]	ProGAN	**0.89**	**0.86**	**0.89**	**0.90**	0.67	0.65	0.62	0.62	0.77	0.77	0.77	0.76	0.76
RINE [37]	LDM	**0.97**	**0.97**	**0.93**	**0.98**	0.69	0.75	0.67	0.74	0.78	**0.87**	**0.81**	**0.86**	**0.84**
PatchCraft [34]	ProGAN	**0.98**	**0.98**	**0.95**	**0.96**	**0.94**	**0.94**	**0.95**	**0.95**	**0.92**	**0.90**	**0.95**	**0.95**	**0.95**
SPAI [39]	LDM	**0.97**	**0.96**	0.77	0.78	**0.85**	**0.84**	**0.86**	**0.86**	**0.83**	**0.81**	**0.86**	**0.85**	**0.85**
SPAI-ITW [41]	ITW	**0.81**	0.75	**0.83**	**0.82**	0.53	0.52	0.53	0.56	0.55	0.52	0.65	0.63	0.64
RINE-ITW [41]	ITW	**0.98**	**0.97**	**0.98**	**0.98**	0.73	0.73	0.72	0.77	0.74	0.71	**0.83**	**0.83**	**0.83**
B-Free [40]	LDM	**1.00**	**1.00**	**1.00**	**0.99**	**0.86**	0.78	**0.85**	0.77	**0.88**	**0.82**	**0.92**	**0.87**	**0.90**

Note that a delta with a negative value means a decrease in performance, and a positive value means an increase in performance. This allows us to systematically evaluate how super-resolution attacks impact the performance of strong forensic models.

For IFL methods, Table 9 shows that super resolution has a pronounced negative impact. While bicubic interpolation produces only a modest decrease in F1 scores (i.e., average F1 score decrease between 0.02 and 0.11), applying Real-ESRGAN leads to a substantial drop across all evaluated subsets (i.e., average F1 score decrease between 0.51 and 0.80). This indicates that generative SR effectively removes the fine-grained pixel-level traces that IFL methods rely on, making localization of manipulated regions far more difficult. This is a similar observation as we made for the impact of full regeneration during generative inpainting (in Section 4.2). Moreover, a similar observation was made in related work, which demonstrated that IFL methods performance drops when applying generative compression using JPEG AI [27]. This highlights that generative image processing constitutes a significant challenge for current IFL methods, requiring new defenses or adaptations.

Table 10 shows that the effect of generative super resolution on SID methods is less pronounced than on IFL methods. For several models, Real-ESRGAN causes a clear but moderate reduction in AUC, suggesting that their detection cues are also weakened by the generative upscaling process. However, other SID models demonstrate much smaller performance drops. In some cases, the performance remains relatively stable (e.g., for UnivFD and RINE-ITW). Notably, for SPAI-ITW, the performance even increases when applying the Real-ESRGAN super-resolution attack. In comparison, applying the bicubic interpolation (as baseline) does not affect detection performance (approximately 0.0 difference in AUC). This suggests that some SID approaches rely on more global or semantic inconsistencies that are less affected by generative SR, while others depend on low-level statistical traces that Real-ESRGAN can affect.

In summary, our experiments show that generative super resolution represents a serious threat to current forensic methods. It can erase the traces needed for IFL, and in many cases also moderately harm SID performance, while bicubic interpolation alone does not have such strong effects. This highlights the importance of accounting for post-processing operations such as SR in the development of robust forensic methods.

Table 9 Evaluation of super-resolution attack on image forgery localization methods on spliced images of our TGIF2 Dataset, alongside a cubic interpolation attack as baseline (F1 score, along with the Δ F1 Score)

IFL method	Original	Super-resolution method			
		Cubic interpolation		Real-ESRGAN	
		×2	×4	×2	×4
CAT-Net [25]	0.87	$0.76^{-0.11}$	$0.73^{-0.14}$	$0.06^{-0.81}$	$0.07^{-0.80}$
TruFor [21]	0.78	$0.76^{-0.02}$	$0.76^{-0.02}$	$0.21^{-0.57}$	$0.22^{-0.56}$
MMFusion [14]	0.69	$0.65^{-0.04}$	$0.64^{-0.05}$	$0.18^{-0.51}$	$0.51^{-0.51}$

Table 10 Evaluation of super-resolution attack on synthetic image detection methods on fully regenerated images in our TGIF2 dataset, alongside a cubic interpolation attack as baseline (AUC Score, along with the Δ AUC score)

SID method	Trained on	Original	Super-resolution method			
			Cubic interpolation		Real-ESRGAN	
			×2	×4	×2	×4
DIMD [35]	LDM	0.88	$0.87^{-0.01}$	$0.87^{-0.01}$	$0.75^{-0.13}$	$0.80^{-0.08}$
UnivFD [30]	ProGAN	0.70	$0.70^{+0.00}$	$0.69^{-0.01}$	$0.68^{-0.02}$	$0.68^{-0.02}$
GramNet [31]	GANs	0.76	$0.76^{+0.00}$	$0.75^{-0.01}$	$0.50^{-0.26}$	$0.50^{-0.26}$
LGrad [32]	ProGAN	0.85	$0.85^{+0.00}$	$0.84^{-0.01}$	$0.49^{-0.36}$	$0.48^{-0.37}$
RINE [37]	ProGAN	0.77	$0.77^{+0.00}$	$0.76^{-0.01}$	$0.72^{-0.05}$	$0.72^{-0.05}$
	LDM	0.81	$0.83^{+0.02}$	$0.83^{+0.02}$	$0.70^{-0.11}$	$0.71^{-0.10}$
PatchCraft [34]	ProGAN	0.95	$0.98^{+0.03}$	$0.98^{+0.03}$	$0.53^{-0.42}$	$0.48^{-0.47}$
SPAI [39]	LDM	0.86	$0.84^{-0.02}$	$0.84^{-0.02}$	$0.73^{-0.13}$	$0.73^{-0.13}$
SPAI–ITW [41]	ITW	0.65	$0.65^{+0.00}$	$0.66^{+0.01}$	$0.78^{+0.13}$	$0.83^{+0.18}$
RINE–ITW [41]	ITW	0.93	$0.83^{+0.00}$	$0.83^{+0.00}$	$0.81^{-0.02}$	$0.83^{+0.00}$
B–Free [40]	LDM	0.92	$0.91^{-0.01}$	$0.91^{-0.01}$	$0.86^{-0.06}$	$0.84^{-0.08}$

5 Discussion and conclusion

Text-guided inpainting forgeries represent a rapidly evolving and increasingly realistic form of AI-based image editing. With TGIF2, we extend our previous dataset (TGIF) by incorporating recent inpainting models, namely the FLUX.1 family, and by introducing random-mask subsets to probe biases in detection methods. Additionally, we extend our previous benchmark by performing fine-tuning to tackle the problem of localization in manipulated but fully regenerated images, as well as evaluating the impact of generative super-resolution methods on forensic methods.

Our experiments demonstrate that existing IFL methods perform well on traditional splicing scenarios, yet some showcase a notable difference in performance between edits with first-generation inpainting methods (i.e., SD and Adobe Firefly) and new-generation methods (i.e., FLUX.1), revealing limitations in their ability to generalize to new inpainting models that significantly increase visual fidelity. Additionally, we noticed that some IFL methods drop in performance when evaluated on the random, non-semantic subsets, which suggests that these existing methods exhibit a bias towards objects and semantics.

Moreover, existing IFL models are unable to localize manipulated areas in FR images. We demonstrated that fine-tuning on FR subsets improves their ability to localize these manipulations, but that it also may introduce biases, which became evident when evaluating on the random-mask subsets. This highlights the importance of training and evaluating methods beyond object-centric manipulations.

We also analyzed the generative quality of the inpainted images, demonstrating that FLUX images are of higher quality. However, we could not establish conclusive relations between generative quality and IFL performance.

For SID, our experiments show that several methods perform strongly on SD2 and SDXL, but performance decreases significantly for FLUX subsets. Hence, FLUX generations are harder to detect for existing IFL and SID methods. This emphasizes the need to keep investing in generalizable SID methods, as well as including new generative methods in training sets.

Furthermore, we show that generative super resolution significantly weakens forensic traces and thereby reduces the effectiveness of forensic methods, similar to the fully regenerated manipulations and generative compression [27]. This is especially disruptive in the case of IFL methods, but also moderately affects some SID methods. In fact, the results suggest that semantics-based SID methods show increased robustness to SR attacks.

It is interesting to contrast the increased SR-attack robustness of semantics-based SID methods (in Section 4.6) to the revealed semantic-bias vulnerability in (fine-tuned) IFL methods that were exposed in random-mask evaluations (in Sections 4.2 and 4.3). This may seem conflicting, since we demonstrate that semantics is beneficial in certain use cases (e.g., robustness to SR attacks), while an over-reliance on semantic cues can introduce vulnerabilities (e.g., that become apparent under random-mask evaluations). It should be noted that the use of random, non-semantic masks is not intended to replace object-centric evaluations, but to serve as a stress test that decouples semantic content from manipulation boundaries, enabling a clearer analysis of model bias and cue reliance. As such, TGIF2 aims to provide complementary evaluation conditions that expose different failure modes. We see semantic reasoning and low-level forensic cues as orthogonal and necessary components, and our benchmark is designed to analyze their trade-offs rather than promote one over the other.

Future work could focus on expanding the TGIF2 dataset with next-generation AI-based inpainting tools, e.g., those that do not require a user-defined mask, such as FLUX Kontext [76], FLUX.2 [77], and the GPT-4o (ChatGPT) Image Generator [78]. Additionally, a key open challenge is improving the robustness of IFL methods against manipulated images that undergo full regeneration (either during manipulation or after, such as by generative super resolution). Regeneration significantly weakens or destroys forensic traces used by current IFL and SID methods. In this context, it is especially important to closely watch out for potential biases.

In summary, TGIF2 provides an updated benchmark that captures the latest advances in text-guided inpainting, highlights strengths and weaknesses of state-of-the-art forensic methods and their fine-tuned versions, and contributes new challenging subsets that will help the community to track progress against this continuously evolving threat.

Appendix 1 Image forgery localization–IoU

The IoU results for IFL on the spliced subsets are given in Table 11 (i.e., corresponding to the F1 results in Table 3). The IoU results for the fully regenerated subsets are given in Table 12 (i.e., corresponding to the F1 results in Table 4) for the original IFL methods, and in Table 13 (i.e., corresponding to the F1 results in Table 5), for the fine-tuned IFL methods.

Table 11 Evaluation of image forgery localization methods on the spliced (SP) subsets of our TGIF2 dataset, for both the semantic (sem) and random (rand) versions (IoU). IoU values above 0.6 are highlighted in a bold font. Just as in Table 3, some IFL methods (i.e., CAT-Net, TruFor, and MMFusion) perform well on SP subsets

IFL method	SD2		PS	FLUX.1						Average		
				[schnell]		[dev]		Fill [dev]				
	Sem	Rand	Sem	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	All
PSCC-Net [20]	0.12	0.13	0.31	0.02	0.06	0.01	0.03	0.06	0.17	0.10	0.10	0.10
SPAN [22]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
ImageForensicsOSN [23]	0.16	0.05	0.25	0.22	0.07	0.16	0.04	0.08	0.07	0.18	0.06	0.12
MVSS-Net++ [13]	0.05	0.02	0.05	0.11	0.04	0.08	0.02	0.06	0.07	0.07	0.04	0.06
Mantranet [24]	0.10	0.09	0.44	0.04	0.01	0.03	0.01	0.03	0.02	0.13	0.03	0.08
CAT-Net [25]	**0.81**	**0.87**	**0.77**	**0.81**	**0.88**	**0.78**	**0.87**	**0.79**	**0.89**	**0.79**	**0.88**	**0.83**
TruFor [21]	**0.74**	**0.79**	**0.69**	**0.69**	**0.60**	**0.61**	0.48	**0.64**	**0.63**	0.67	0.63	0.65
MMFusion [14]	**0.62**	**0.63**	**0.62**	**0.60**	0.50	0.49	0.37	0.54	0.50	0.57	0.50	0.54

Table 12 Evaluation of image forgery localization methods on the fully regenerated (FR) subsets of our TGIF2 dataset, for both the semantic (sem) and random (rand) versions (IoU). The last rows show the IFL methods fine-tuned on All FR Sem subsets

IFL method	SD2		SDXL		FLUX.1						Average		
					[schnell]		[dev]		Fill [dev]				
	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	All
PSCC-Net [20]	0.03	0.02	0.03	0.05	0.01	0.01	0.01	0.01	0.02	0.03	0.02	0.03	0.02
SPAN [22]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ImageForensicsOSN [23]	0.14	0.05	0.12	0.05	0.22	0.05	0.16	0.04	0.06	0.03	0.14	0.04	0.09
MVSS-Net++ [13]	0.04	0.02	0.06	0.03	0.11	0.02	0.09	0.02	0.05	0.03	0.07	0.02	0.05
Mantranet [24]	0.02	0.00	0.03	0.03	0.03	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.02
CAT-Net [25]	0.03	0.01	0.02	0.00	0.04	0.01	0.03	0.00	0.02	0.01	0.03	0.01	0.02
TruFor [21]	0.13	0.04	0.12	0.06	0.25	0.05	0.18	0.04	0.10	0.05	0.16	0.05	0.10
MMFusion [14]	0.11	0.03	0.13	0.06	0.26	0.05	0.15	0.02	0.08	0.04	0.15	0.04	0.09
TruFor – fine-tuned	0.37	0.28	0.58	**0.75**	**0.80**	**0.90**	**0.74**	**0.82**	**0.67**	**0.78**	**0.63**	**0.71**	**0.67**
MMFusion – fine-tuned	0.37	0.29	0.55	**0.75**	**0.78**	**0.89**	**0.73**	**0.80**	**0.66**	**0.77**	**0.62**	**0.70**	**0.66**

Table 13 Evaluation of individually fine-tuned image forgery localization methods on the fully regenerated (FR) subsets of our TGIF2 dataset, for both the semantic (sem) and random (rand) versions (IoU). IoU values above 0.6 are highlighted in a bold font. Just as in Table 5, we notice good in-domain performance but bad out-of-domain generalization, as well as a potential bias towards semantics

IFL method	Fine-tune set		SD2		SDXL		FLUX.1						Average		
							[schnell]		[dev]		Fill [dev]				
			Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	All
TruFor	All	Sem+Rand	0.37	0.28	0.58	**0.75**	**0.80**	**0.90**	**0.74**	**0.82**	**0.67**	**0.78**	**0.63**	**0.71**	**0.67**
		Sem	0.21	0.29	0.41	**0.71**	**0.60**	**0.87**	0.53	**0.78**	0.53	**0.74**	0.55	0.41	0.48
		Rand	0.38	0.15	0.50	0.52	**0.71**	0.52	0.59	0.37	0.55	0.50	0.46	**0.68**	0.57
	SD2	Sem	0.38	0.14	0.20	0.11	0.29	0.11	0.22	0.05	0.15	0.06	0.25	0.09	0.17
		Rand	0.21	0.29	0.11	0.14	0.08	0.06	0.05	0.03	0.08	0.08	0.11	0.12	0.11
	SDXL	Sem	0.07	0.01	0.55	**0.61**	0.14	0.01	0.06	0.01	0.04	0.01	0.17	0.13	0.15
		Rand	0.01	0.00	0.41	**0.73**	0.01	0.00	0.00	0.01	0.02	0.02	0.09	0.15	0.12
	Flux	Sem	0.07	0.01	0.14	0.03	**0.83**	**0.85**	**0.79**	**0.74**	**0.70**	**0.72**	0.51	0.47	0.49
		Rand	0.01	0.01	0.03	0.04	**0.65**	**0.91**	**0.61**	**0.86**	0.58	**0.77**	0.38	0.52	0.45

IFL method	Fine-tune set		SD2		SDXL		FLUX.1 [schnell]		FLUX.1 [dev]		FLUX.1 Fill [dev]		Average		
			Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	Sem	Rand	All
MMFusion	All	Sem+Rand	0.37	0.29	0.55	**0.75**	**0.78**	**0.89**	**0.73**	**0.80**	**0.66**	**0.77**	**0.62**	**0.70**	0.66
		Sem	0.36	0.11	0.47	0.50	**0.70**	0.48	**0.62**	0.41	0.59	0.57	0.55	0.42	0.48
		Rand	0.18	0.30	0.39	**0.71**	0.54	**0.85**	0.49	**0.75**	0.52	**0.72**	0.42	**0.67**	0.55
	SD2	Sem	0.40	0.15	0.19	0.12	0.26	0.10	0.19	0.05	0.12	0.06	0.23	0.10	0.16
		Rand	0.20	0.24	0.11	0.12	0.05	0.03	0.04	0.02	0.06	0.05	0.09	0.09	0.09
	SDXL	Sem	0.06	0.01	0.55	**0.60**	0.10	0.01	0.04	0.01	0.03	0.01	0.16	0.13	0.14
		Rand	0.01	0.01	0.41	**0.71**	0.02	0.00	0.01	0.01	0.03	0.03	0.10	0.15	0.12
	Flux	Sem	0.07	0.01	0.14	0.03	**0.83**	**0.80**	**0.77**	**0.68**	**0.69**	**0.71**	0.50	0.45	0.47
		Rand	0.02	0.01	0.04	0.05	**0.65**	**0.90**	**0.61**	**0.85**	**0.60**	**0.77**	0.38	0.52	0.45

Acknowledgements
Not applicable.

Materials availability
The TGIF2 materials are available at https://github.com/IDLabMedia/tgif-dataset.

Code availability
The TGIF2 code is available at https://github.com/IDLabMedia/tgif-dataset.

Authors' contributions
HM: dataset preparation, initial experiments, interpretation of experiments, main manuscript text writing. DK: large-scale benchmark experiments, interpretation of experiments. PG: large-scale benchmark experiments, interpretation of experiments. SP: guidance, funding acquisition. PL: guidance, funding acquisition. GVW: guidance, funding acquisition. All authors revised and approved the manuscript.

Funding
This work was funded by the Flemish government's Department of Culture, Youth and Media (under the COM-PRESS project), by IDLab (Ghent University–imec), by Flanders Innovation and Entrepreneurship (VLAIO), by Research Foundation–Flanders (FWO) (V419524N and G0A2523N), and by the European Union under the Horizon Europe projects AI4Trust (grant number 101070190) and AI-CODE (grant number 101135437).

Data availability
The dataset is available at https://github.com/IDLabMedia/tgif-dataset.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
The authors give consent for publication.

Competing interests
The authors declare that they have no competing interests.

Received: 30 September 2025 Accepted: 24 March 2026
Published online: 10 April 2026

References

1. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, in *Proc. IEEE/CVF Conf. Computer Vission Pattern Recogn.*, High-resolution image synthesis with latent diffusion models (IEEE, New York, 2022)
2. J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Face2face: Real-time face capture and reenactment of rgb videos (IEEE, New York, 2016), pp. 2387–2395
3. Adobe. Bringing generative AI into creative cloud with Adobe Firefly (2023). https://blog.adobe.com/en/publish/2023/03/21/bringing-gen-ai-to-creative-cloud-adobe-firefly. Accessed 26 Sep 2025
4. Black Forest Labs. Flux (2024). https://github.com/black-forest-labs/flux. Accessed 26 Sep 2025
5. A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, M. Chen, in *Proceedings of the 39th International Conference on Machine Learning*, GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models (PMLR, 2022), pp. 16784–16804
6. L. Verdoliva, Media forensics and deepfakes: an overview. IEEE J. Sel. Top. Signal Process. **14**(5), 910–932 (2020)
7. H. Mareen, D. Karageorgiou, G.V. Wallendael, P. Lambert, S. Papadopoulos, in *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, TGIF: Text-guided inpainting forgery dataset (IEEE, New York, 2024)
8. D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, SDXL: Improving latent diffusion models for high-resolution image synthesis (2023). arXiv preprint arXiv:2307.01952
9. fal.ai. imgsys.org - a generative image model arena (2025). https://imgsys.org/rankings. Accessed 19 Sep 2025
10. X. Wang, L. Xie, C. Dong, Y. Shan, in *International Conference on Computer Vision Workshops (ICCVW)*, Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data (IEEE, New York)
11. H. Mareen, D. Vanden Bussche, F. Guillaro, D. Cozzolino, G. Van Wallendael, P. Lambert, L. Verdoliva, in *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*. ed. by J.J. Rousseau, B. Kapralos. Comprint: Image forgery detection and localization using compression fingerprints (Springer Nature Switzerland, Cham, 2023), pp. 281–299. https://doi.org/10.1007/978-3-031-37742-6_23
12. D. Cozzolino, L. Verdoliva, Noiseprint: a CNN-based camera model fingerprint. IEEE Trans Inf. Forensics Security **15**, 144–159 (2020)
13. C. Dong, X. Chen, R. Hu, J. Cao, X. Li, MVSS-Net: multi-view multi-scale supervised networks for image manipulation detection. IEEE Trans. Pattern Analysis and Machine Intel. **45**(3), 3539–3553 (2022)
14. K. Triaridis, V. Mezaris, in *Int. Conf. Multimedia Model.*, Exploring multi-modal fusion for image manipulation detection and localization (Springer-Verlag, Berlin, 2024), pp. 198–211
15. D. Karageorgiou, G. Kordopatis-Zilos, S. Papadopoulos, in *Proc. IEEE/CVF Conf. Computer Vission Pattern Recogn.*, Fusion transformer with object mask guidance for image forgery analysis (IEEE, New York, 2024), pp. 4345–4355
16. H. Mareen, L. De Neve, P. Lambert, G. Van Wallendael, Harmonizing image forgery detection & localization: fusion of complementary approaches. J. Imaging **10**(1), 4 (2023)
17. H. Li, W. Luo, J. Huang, Localization of diffusion-based inpainting in digital images. IEEE Trans Inf. Forensics Security **12**(12), 3050–3064 (2017)

18. H. Wu, J. Zhou, Iid-net: image inpainting detection network via neural architecture search and attention. IEEE Trans. Circuits Syst. Video Technol. **32**(3), 1172–1185 (2021)
19. A. Li, Q. Ke, X. Ma, H. Weng, Z. Zong, F. Xue, R. Zhang, in *Proc. Int. Conf. Artificial Intel. IJCAI*, ed. by Z.H. Zhou, Noise doesn't lie: Towards universal detection of deep inpainting (International Joint Conferences on Artificial Intelligence Organization, 2021), pp. 786–792
20. X. Liu, Y. Liu, J. Chen, X. Liu, PSCC-Net: progressive spatio-channel correlation network for image manipulation detection and localization. IEEE Trans. Circuits Syst. Video Technol. **32**(11), 7505–7517 (2022)
21. F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, L. Verdoliva, in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recogn. (CVPR)*, TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization (IEEE, New York, 2023), pp. 20606–20615
22. X. Hu, Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang, R. Nevatia, in *Proc. Europ. Computer Vision Conf.*, SPAN: Spatial pyramid attention network for image manipulation localization (Springer, 2020), pp. 312–328
23. H. Wu, J. Zhou, J. Tian, J. Liu, Y. Qiao, Robust image forgery detection against transmission over online social networks. IEEE Trans. Inf. Forensics Secur. **17**, 443–456 (2022)
24. Y. Wu, W. AbdAlmageed, P. Natarajan, in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recogn.*, ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features (IEEE, New York, 2019), pp. 9543–9552
25. M.J. Kwon, S.H. Nam, I.J. Yu, H.K. Lee, C. Kim, Learning JPEG compression artifacts for image manipulation detection and localization. Int. J. Comput. Vis. **130**(8), 1875–1895 (2022)
26. J. Li, N. Wang, L. Zhang, B. Du, D. Tao, in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recogn.*, Recurrent feature reasoning for image inpainting (IEEE, New York, 2020), pp. 7760–7768
27. E.D. Cannas, S. Mandelli, N. Popović, A. Alkhateeb, A. Gnutti, P. Bestagini, S. Tubaro, Is jpeg ai going to change image forensics? (2024). arXiv preprint arXiv:2412.03261
28. S.Y. Wang, O. Wang, R. Zhang, A. Owens, A.A. Efros, in *Proc. Int. Conf. Pattern Recogn. (CVPR)*, Cnn-generated images are surprisingly easy to spot...for now (IEEE, New York, 2020)
29. J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, T. Holz, in *Int. Conf. Machine Learning*, Leveraging frequency analysis for deep fake image recognition (PMLR, 2020), pp. 3247–3258
30. U. Ojha, Y. Li, Y.J. Lee, in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recogn.*, Towards universal fake image detectors that generalize across generative models (IEEE, New York, 2023), pp. 24480–24489
31. Z. Liu, X. Qi, P.H. Torr, in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recogn.*, Global texture enhancement for fake face detection in the wild (IEEE, New York, 2020), pp. 8060–8069
32. C. Tan, Y. Zhao, S. Wei, G. Gu, Y. Wei, in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recogn.*, Learning on gradients: Generalized artifacts representation for gan-generated images detection (IEEE, New York, 2023), pp. 12105–12114
33. C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, Y. Wei, in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recogn.*, Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection (IEEE, New York, 2024), pp. 28130–28139
34. N. Zhong, Y. Xu, Z. Qian, X. Zhang, PatchCraft: Exploring texture patch for efficient ai-generated image detection (2023). arXiv preprint arXiv:2311.12397
35. R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, L. Verdoliva, in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, On the detection of synthetic images generated by diffusion models (IEEE, New York, 2023), pp. 1–5
36. Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, H. Li, in *Proc. IEEE/CVF Int. Conf Computer Vision (ICCV)*, DIRE for diffusion-generated image detection (IEEE, New York, 2023), pp. 22445–22455
37. C. Koutlis, S. Papadopoulos, in *Proc. Europ. Computer Vision Conf. (ECCV)*, Leveraging representations from intermediate encoder-blocks for synthetic image detection (Springer Nature Switzerland, Cham, 2024)
38. Z. Sha, Z. Li, N. Yu, Y. Zhang, in *Proc. ACM SIGSAC Conf. Computer Comm. Security*, DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models (Association for Computing Machinery, New York, 2023), pp. 3418–3432
39. D. Karageorgiou, S. Papadopoulos, I. Kompatsiaris, E. Gavves, in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, Any-resolution AI-generated image detection by spectral learning (IEEE, New York, 2025), pp. 18706–18717
40. F. Guillaro, G. Zingarini, B. Usman, A. Sud, D. Cozzolino, L. Verdoliva, in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, A bias-free training paradigm for more general ai-generated image detection (IEEE, New York, 2025), pp. 18685–18694
41. D. Konstantinidou, D. Karageorgiou, C. Koutlis, O. Papadopoulou, E. Schinas, S. Papadopoulos, Navigating the challenges of AI-generated image detection in the wild: What truly matters? (2025). arXiv preprint arXiv:2507.10236
42. S. Mandelli, P. Bestagini, S. Tubaro, in *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, When synthetic traces hide real content: Analysis of stable diffusion image laundering (IEEE, 2024)
43. L. Lin, N. Gupta, Y. Zhang, H. Ren, C.H. Liu, F. Ding, X. Wang, X. Li, L. Verdoliva, S. Hu, Detecting multimedia generated by large AI models: A survey (2024). arXiv preprint arXiv:2402.00045
44. H. Guan, M. Kozak, E. Robertson, Y. Lee, A.N. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith, J. Fiscus, in *Proc. IEEE Winter App. Computer Vision Workshops (WACVW)*, MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation (IEEE, 2019), pp. 63–72
45. G. Mahfoudi, B. Tajini, F. Retraint, F. Morain-Nicolier, J.L. Dugelay, P. Marc, in *Europ. Signal Process. Conf. (EUSIPCO)*, Defacto: Image and face manipulation dataset (IEEE, 2019), pp. 1–5
46. T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, in *Proc. Europ. Computer Vision Conf.*, Microsoft COCO: Common objects in context (Springer, 2014), pp. 740–755
47. S. Jia, M. Huang, Z. Zhou, Y. Ju, J. Cai, S. Lyu, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Autosplice: A text-prompt manipulated image dataset for media forensics (IEEE, New York, 2023), pp. 893–903
48. P. Giakoumoglou, D. Karageorgiou, S. Papadopoulos, P.C. Petrantonakis, in *Proc. IEEE/CVF Int. Conf Computer Vision (ICCV)*, SAGI: Semantically aligned and uncertainty guided ai image inpainting (IEEE, New York, 2025)
49. D.T. Dang-Nguyen, C. Pasquini, V. Conotter, G. Boato, in *Proceedings of the 6th ACM multimedia systems conference*, RAISE: A raw images dataset for digital image forensics (Association for Computing Machinery, New York, 2015), pp. 219–224
50. A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, V. Ferrari, The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale. Int. J. Comput. Vis. **128**(7), 1956–1981 (2020)
51. L. Zhang, A. Rao, M. Agrawala, in *Proceedings of the IEEE/CVF international conference on computer vision*, Adding conditional control to text-to-image diffusion models (IEEE, New York, 2023), pp. 3836–3847
52. R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, V. Lempitsky, Resolution-robust large mask inpainting with fourier convolutions (2021). arXiv preprint arXiv:2109.07161
53. T. Yang, T. Jordan, N. Liu, J. Sun, Common inpainted objects in-n-out of context (2025). arXiv preprint arXiv:2506.00721
54. Z. Sun, H. Fang, J. Cao, X. Zhao, D. Wang, in *Proceedings of the 32nd ACM International Conference on Multimedia*, Rethinking image editing detection in the era of generative AI revolution (Association for Computing Machinery, New York, 2024), pp. 3538–3547
55. W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, J. Jia, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, MAT: Mask-aware transformer for large hole image inpainting (IEEE, New York, 2022), pp. 10758–10768
56. H. Zhang, X. Wan, UniAIDet: A unified and universal benchmark for ai-generated image content detection and localization (2025). arXiv preprint arXiv:2510.23023
57. Y. Wang, Z. Huang, X. Hong, in *Proceedings of the Computer Vision and Pattern Recognition Conference*, OpenSDI: Spotting diffusion-generated images in the open world (IEEE, New York, 2025), pp. 4291–4301
58. G. Bertazzini, C. Albisani, D. Baracchi, D. Shullani, A. Piva, in *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, Beyond the brush: Fully-automated crafting of realistic inpainted images (IEEE, 2024), pp. 1–6

59. Y. Chen, X. Huang, Q. Zhang, W. Li, M. Zhu, Q. Yan, S. Li, H. Chen, H. Hu, J. Yang, W. Liu, J. Hu, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, GIM: A million-scale benchmark for generative image manipulation detection and localization (Association for the Advancement of Artificial Intelligence Press, Washington, DC, 2025), pp. 2311–2319
60. L. Cai, H. Wang, J. Ji, Y. ZhouMen, Y. Ma, X. Sun, L. Cao, R. Ji, Zooming in on fakes: A novel dataset for localized ai-generated image detection with forgery amplification approach (2025). arXiv preprint arXiv:2504.11922
61. Y. Wang, J. Yu, J. Zhang, Zero-shot image restoration using denoising diffusion null-space model (2022). arXiv preprint arXiv:2212.00490
62. X. Ju, X. Liu, X. Wang, Y. Bian, Y. Shan, Q. Xu, in *European Conference on Computer Vision*, BrushNet: A plug-and-play image inpainting model with decomposed dual-branch diffusion (Springer, 2024), pp. 150–168
63. J. Zhuang, Y. Zeng, W. Liu, C. Yuan, K. Chen, in *European Conference on Computer Vision*, A task is worth one word: Learning with task prompts for high-quality versatile image inpainting (Springer, 2024), pp. 195–211
64. S. Smeu, D.A. Boldisor, D. Oneata, E. Oneata, in *Proceedings of the Computer Vision and Pattern Recognition Conference*, Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning (IEEE, New York, 2025), pp. 18815–18825
65. S. Xie, Z. Zhang, Z. Lin, T. Hinz, K. Zhang, in *Proc. IEEE/CVF Conf. Computer Vission Pattern Recogn.*, SmartBrush: Text and shape guided object inpainting with diffusion model (IEEE, New York, 2023), pp. 22428–22437
66. S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D.J. Fleet, R. Soricut, J. Baldridge, M. Norouzi, P. Anderson, W. Chan, in *Proc. IEEE/CVF Conf. Computer Vission Pattern Recogn.*, Imagen editor and EditBench: Advancing and evaluating text-guided image inpainting (IEEE, New York, 2023), pp. 18359–18369
67. Diffusers: State-of-the-art diffusion models for image and audio generation in pytorch and flax. https://github.com/huggingface/diffusers. Accessed 5 July 2024
68. H. Talebi, P. Milanfar, NIMA: neural image assessment. IEEE Trans. Image Process. **27**(8), 3998–4011 (2018)
69. S. Gu, J. Bao, D. Chen, F. Wen, in *Proc. Europ. Computer Vision Conf. (ECCV)*, GIQA: Generated image quality assessment (Springer, 2020), pp. 369–385
70. J. Li, D. Li, C. Xiong, S. Hoi, in *Int. Conf. Machine Learning*, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation (PMLR, 2022), pp. 12888–12900
71. R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recogn. (CVPR)*, The unreasonable effectiveness of deep features as a perceptual metric (IEEE, New York, 2018), pp. 586–595
72. Diffusers github issue - inpainting produces results that are uneven with input image. https://github.com/huggingface/diffusers/issues/5808. Accessed 5 July 2024
73. M. Schinas, S. Papadopoulos, in *MAD '24: Proc. ACM Int. W. Multimedia AI ag. Disinform.*, SIDBench: A python framework for reliably assessing synthetic image detection methods (Association for Computing Machinery, New York, 2024), pp. 55–64
74. Y. Ju, S. Jia, L. Ke, H. Xue, K. Nagano, S. Lyu, in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Fusing global and local features for generalized ai-synthesized image detection (IEEE, 2022), pp. 3465–3469
75. S. Smeu, E. Oneata, D. Oneata, in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, DeCLIP: Decoding CLIP representations for deepfake localization (IEEE, 2025), pp. 149–159
76. B.F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English, P. Esser, S. Kulal, K. Lacey, Y. Levi, C. Li, D. Lorenz, J. Müller, D. Podell, R. Rombach, H. Saini, A. Sauer, L. Smith, Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025). arXiv:2506.15742
77. B.F. Labs. FLUX.2: Frontier Visual Intelligence (2025). https://bfl.ai/blog/flux-2. Accessed 2 Feb 2025
78. S. Chen, J. Bai, Z. Zhao, T. Ye, Q. Shi, D. Zhou, W. Chai, X. Lin, J. Wu, C. Tang, S. Xu, T. Zhang, H. Yuan, Y. Zhou, W. Chow, L. Li, X. Li, L. Zhu, L. Qi, An empirical study of GPT-4o image generation capabilities (2025). arXiv preprint arXiv:2504.05979

Publisher's Note