OpenSDI: Spotting Diffusion-Generated Images in the Open World

Supplementary Material

In this supplementary material, we provide additional details and results to complement the main paper. Section 1 offers a step-by-step explanation of the OpenSDID creation pipeline, detailing the processes and tools used. Section 2 includes a comprehensive analysis of the dataset's diversity and quality, along with a thorough discussion of the ethical and bias considerations. Section 3 provides a more detailed comparison of the proposed MaskCLIP with related works, highlighting its unique contributions and advantages. To further validate the performance and robustness of MaskCLIP, we present additional experimental results in Section 4 to validate the performance and robustness of MaskCLIP, including cross-dataset evaluations on traditional image forgery datasets and more recent AI-generated image benchmarks, robustness analysis under image degradation, and multiple runs to ensure stability. Additionally, we provide visualizations of the results and datasets in Sections 5 and 6, offering intuitive insights into the data and the method's performance. We are committed to making our research reproducible and accessible to the broader community. Therefore, we have made both the OpenSDID dataset and the complete codebase publicly available at https://github.com/iamwangyabin/OpenSDI.

1. OpenSDID Creation Pipeline

To create our OpenSDID, we designed a comprehensive framework to automatically generate a large and diverse set of edited images using advanced text-to-image (T2I) diffusion generative models. As illustrated in Figure 1, the process consists of four key steps: (A) load real images from a database, (B) generate textual instructions for editing, (C) create visual masks for editing, and (D) produce AI-generated images with instructions and masks. For global image image generation, OpenSDID merely uses (B) and (D) without using masks to generate images.

We begin by randomly selecting an authentic image from the Megalith-10M database [3]. Next, we randomly choose a state-of-the-art large vision-language model, such as LLaMA3 Vision [1], LLaVA [24], InternVL2 [6], or Qwen2 VL [44], to generate detailed edit suggestions. These models are selected for their ability to produce high-quality and contextually relevant edit instructions, ensuring that the subsequent image manipulations are both creative and realistic. The prompts used to generate these edit suggestions are carefully crafted to guide the model in producing high-quality and relevant modifications. We emphasizes the importance of considering a wide range of potential modifications, including face transformations, hair modifications, body alterations, clothing and accessory changes, object replacements, background change, style modifications, and other adjustments. Below is an example of the prompts that are likely to be used:

Prompt

```
You are an image editor with decades
of experience in digital manipulation
and a reputation for innovation.
Your expertise spans across various
genres,including portrait retouching,
architectural visualization,
product photography, and surreal
composites. Your task is to analyze
image descriptions and propose
compelling, realistic edits that
could dramatically enhance or
transform the image in unexpected
yet believable ways.

When presented with an image,
consider a comprehensive range of
potential modifications.
These modifications could be:
• Face transformations
• Hair modifications
• Body alterations
• Clothing and accessory changes
• Object replacements or additions
• Background transformations
• Architectural style modifications
• Vehicle transformations
• Food alterations and adjustments

Your suggestions should be both
imaginative and feasible,
taking into account the original
image's context, composition,
and lighting. Strive for a balance
between creativity and photorealism,
ensuring that your proposed edits
could theoretically be executed
by a skilled retoucher.
```

The output text edit instructions specify the selected region in the given real images and the desired content to modify that area. Thus, based on the edit instructions for the previous step, we utilize the Florence-2 [49] and SAM [16]

CVPR
#7349

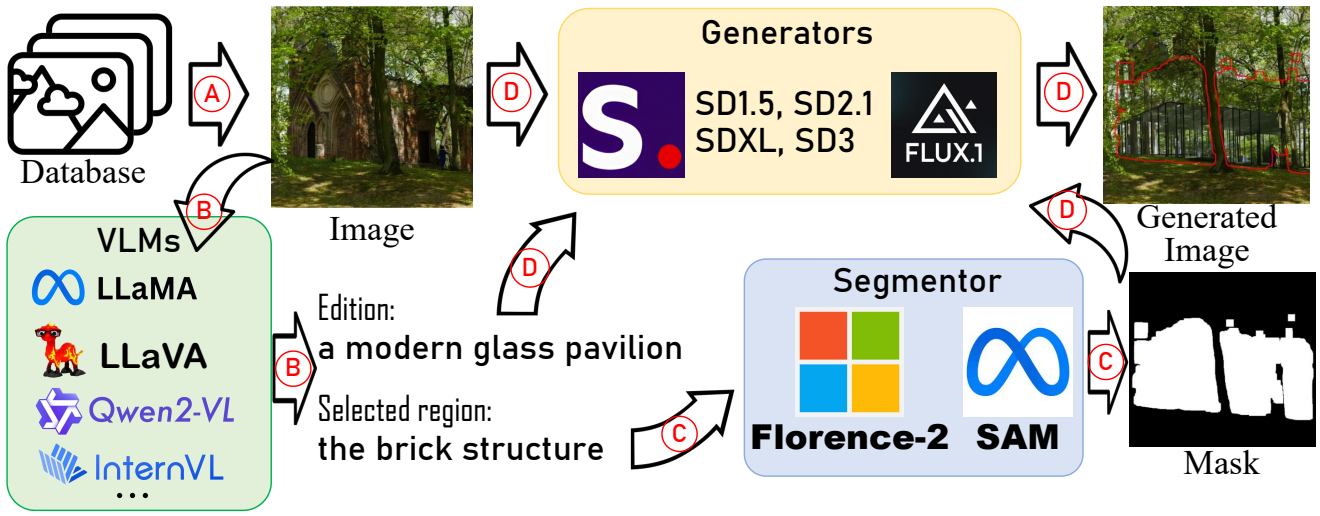

Figure 1. An OpenSDID pipeline for local modification on real image content: (A) Sampling real images from the Megalith-10M dataset, (B) Generating textual instructions for editing using Vision Language Models (VLMs), (C) Creating visual masks for modification through segmentation models, and (D) Producing AI-generated images with image generators based on the instructions and masks. For global image content generation, OpenSDID merely uses (B) and (D) without using real images to produce masks.

models to obtain precise mask regions through open vocabulary detection and segmentation. The Florence-2 model identifies specific regions within the image using text input, while the SAM model refines these regions by generating more precise masks. These masks are further processed by removing small disconnected components and expanded to ensure comprehensive coverage of the areas designated for inpainting.

One of the state-of-the-art T2I diffusion models (SD1.5 [36], SD2.1 [36], SDXL [33], SD3 [10]) and Flux.1 [21]) is then employed to generate new images based on the masked regions and the corresponding prompts. The employed diffusion model takes three inputs: the original image, the generated mask, and the prompt describing the desired new content. To enhance the generation diversity, we randomly adjust key parameters, including the number of inference steps, guidance scale, and strength.

Finally, we use CLIP [34] to compute similarity scores of the generated images and the given edit instructions. Only images that achieve high similarity scores are included in the final dataset, ensuring the maintenance of high-quality standards.

2. OpenSDID Details and Analysis

2.1. Comparison with Existing Datasets

We give a more comprehensive comparison of the existing datasets in Table 1. In terms of user-like diversity, existing datasets have often suffered from limited variability in user input due to their reliance on standardized generation pipelines. This strategy results in reduced diversity across generated images. While DiffusionDB made progress by collecting generated images from public Discord channels with varied generator parameters, it remained confined to Stable Diffusion 1.5 and globally generated images. Our dataset addresses these limitations by introducing substantial user-like diversity through the integration of multiple state-of-the-art Visual Language Models (VLMs), including LLaMA3 Vision and LLaVA. These VLMs simulate a broad spectrum of human-like editing behaviors, generating diverse text prompts that authentically reflect real-world manipulation intentions. Additionally, we randomize the generation hyperparameters during the diffusion process, including sampling steps, guidance scale, and seed values, to further enhance the diversity of our generated images.

Regarding Model Innovation, many existing datasets, such as DFFD [7] and HiFi-Net [13], primarily focus on images generated by GANs and early diffusion models, lacking representation of recent technological advances. OpenSDID addresses this limitation by incorporating multiple cutting-edge T2I diffusion models, including various versions of Stable Diffusion [10, 33, 36] and Flux.1 [21]. This comprehensive inclusion provides a more robust platform for evaluating detection methods against the rapidly evolving landscape of image synthesis technologies.

With respect to Manipulation Scope, existing datasets often limit themselves to either global or local manipulations. For example, DiffusionDB [46] and GenImage [52] contain only globally synthesized images without local edits, whereas OpenSDID uniquely combines both global and local manipulations. This comprehensive manipulation spectrum more accurately reflects real-world scenarios,

where forgeries frequently involve complex combinations of global and localized alterations.

2.2. Ethical and Bias Issues

The proposed OpenSDID research has been approved by the University's Ethics and Research Governance Online team. Furthermore, we provide a more detailed analysis of potential ethical and bias issues in the dataset.

Copyright Considerations. To ensure the dataset is publicly accessible, we have made every effort to ensure that all locally manipulated images based on [3] are free from copyright restrictions. All images used in this research are freely available under one or more of the following licenses: No Known Copyright Restrictions, United States Government Work, Public Domain Dedication (CC0), or Public Domain Mark. These licenses permit unrestricted use, modification, and distribution of the visual materials. Specifically, the images fall into the following categories: (a) materials free from known copyright restrictions due to their age or provenance, (b) content created by U.S. federal government employees as part of their official duties, (c) works explicitly dedicated to the public domain through Creative Commons Zero designation, or (d) materials marked as public domain due to copyright expiration or other legal provisions. Although these licenses do not require attribution, we have included source citations where applicable to maintain academic integrity and adhere to best practices. The distribution of these copyright statuses is shown in Figure 2.

However, it is important to acknowledge that, given the nature of the Internet, there remains a possibility that a small number of images may have been uploaded by users without proper copyright clearance. While we reserve all rights to the AI-generated images produced in this research, we do not claim ownership of the original source images. Researchers can access these source images independently through the image URLs we have provided in our dataset.

SFW Ensurance. Our dataset consists of images retrieved through the Flickr API, implementing rigorous safety protocols with maximum safety settings (safety_level=1) to ensure content appropriateness [3]. For more comprehensive safety verification, particularly for artificially generated images, we employed multiple state-of-the-art NSFW detectors including:

- GantMAN NSFW Detector [20]
- LAION's CLIP-based NSFW Detector [38]
- Stable Diffusion Safety Checker [35]

Our multi-layered verification process confirmed that 100% of the images maintain SFW status, which can be attributed to both the strict initial filtering and the inherent safety mechanisms in the underlying VLMs and T2I generators.

Potential Biases Analysis. To conduct a comprehensive bias analysis of our dataset, we employed two sophisticated detection frameworks: EasyFace [2] for human demographic analysis and Florence-2 for general object detection.

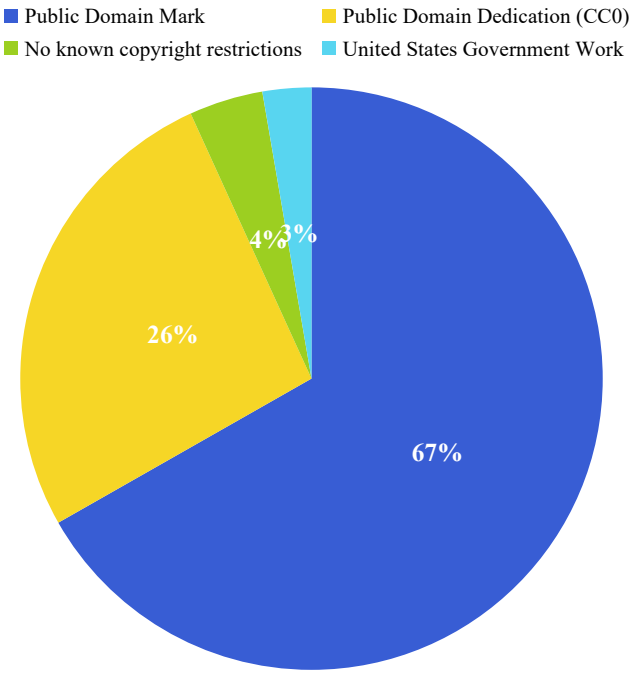

Figure 2. Distribution of copyright statuses for real images in OpenSDID.

We utilized the EasyFace detection framework to assess representation across multiple demographic dimensions. Our analysis encompassed binary gender classification, seven distinct racial/ethnic categories, and nine age groups spanning from infancy to elderly, providing a granular view of demographic representation within the dataset. The face detection algorithm was applied across the entire dataset, utilizing pre-trained models optimized for multi-attribute recognition. Table 2 presents the results. While this analysis provides valuable insights into dataset representation, we acknowledge the inherent limitations of automated demographic classification systems, particularly when dealing with edge cases and intersectional identities.

To complement the demographic analysis, we employed Florence-2 for open-world object detection, providing insights into the distribution of general categories within the dataset. The results are visualized in Figure 3.

3. More Related Work Discussion

For training and evaluation, we use the implementation of the IMDL-BenCo framework [28], which includes several state-of-the-art methods for comparison. Here, we briefly describe the methods included in our paper.

CAT-Net (Compression Artifact Tracing Network) [19] is a dual-stream neural network that simultaneously pro-

Dataset	Type	# Real	# Fake	Generator	Adv. T2I	Glo.	Loc.	Users	# Model
UADFV [50]	Face	241	252	GAN	✗	✗	✓	✗	1
DFFD [7]	Face	58K	240K	GAN	✗	✓	✓	✗	7
FaceForensics++ [37]	Face	1K	4K	GAN	✗	✗	✓	✗	1
DFDC [8]	Face	19K	100K	GAN	✗	✗	✓	✗	2
DeeperForensics [15]	Face	50K	10K	GAN	✗	✗	✓	✗	1
CNNSpot [45]	General	362K	362K	GAN	✗	✓	✗	✗	13
GenImage [52]	General	1.3M	1.4M	GAN & Diff.	✓	✓	✗	✗	8
DiffusionDB [46]	General	-	14M	Diff.	✓	✓	✗	✓	1
Columbia [30]	General	933	912	Trad.	✗	✗	✓	✗	-
CASIA [9]	General	7.2K	5.1K	Trad.	✗	✗	✓	✗	-
IMD2020 [31]	General	35K	35K	Trad.	✗	✗	✓	✗	-
NIST16 [11]	General	-	564	Trad.	✗	✗	✓	✗	-
Coverage [47]	General	100	100	Trad.	✗	✗	✓	✗	-
AutoSplice [14]	General	2.3K	3.6K	Diff.	✓	✓	✓	✗	1
CocoGlide [12]	General	–	512	Diff.	✓	✗	✓	✗	1
Dolos [42]	Face	20K	105K	Diff.	✓	✓	✓	✗	4
HiFi-Net [13]*	General	–	1M	GAN & Diff.	✗	✓	✓	✗	10
GIM [5]	General	300K	1.1M	Diff.	✓	✗	✓	✗	3
TGIF [29]	General	3.1K	75K	Diff.	✓	✗	✓	✗	3
OpenSDID	General	300K	450K	Diff.	✓	✓	✓	✓	5

Table 1. Overview of more comprehensive image forgery datasets. "Type" indicates the content category (Face or General). "# Real & # Fake" indicates the number of real and fake images. "Generator" indicates synthesis method type (GAN, Diffusion (Diff.), or Traditional (Trad.)). "Adv. T2I" indicates whether advanced Text-to-Image models (e.g., Stable Diffusion) are used. "Glo." indicates the global image manipulations. "Loc." indicates the local image manipulations. "Users" indicates whether multiple users participate in dataset creation. "# Model" indicates the number of distinct models used for generation. HiFi-Net's locally manipulated images are primarily sourced from previous traditional manipulation datasets and facial deepfake detection datasets.

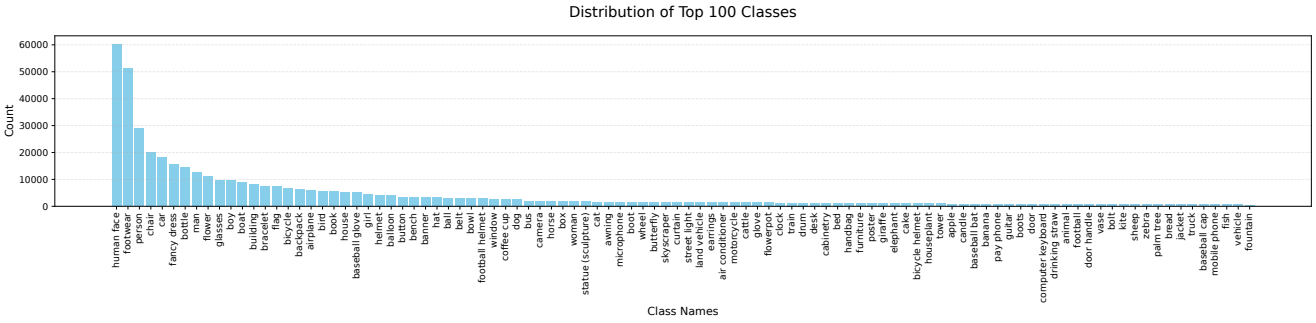

Figure 3. Distribution of top-100 classes in the OpenSDID

cesses both RGB images and JPEG compression artifacts (DCT coefficients) to detect image manipulations. By leveraging both visual content and compression artifacts, it is particularly effective at identifying forgeries, even in compressed images.

MVSS-Net (Multi-View Multi-Scale Supervised Networks) [4] is a dual-branch architecture that simultaneously processes both RGB domain features and noise patterns. The edge-supervised branch utilizes edge residual blocks, while the multi-scale feature learning branch analyzes tampering edge artifacts and noise views of input images.

PSCC-Net (Progressive Spatio-Channel Correlation Network) [25] leverages multi-scale feature learning through dense cross-connections and progressive feature fusion strategies. The model utilizes different sizes of convolutions and perceptual fields to extract valuable information about tampered locations, making it particularly effective at detecting various types of image manipulations.

TruFor [12] uses a Noiseprint++ extractor to process RGB images and obtain learned noise-sensitive fingerprints. These fingerprints, along with the original RGB input, are fed into an encoder that jointly computes features for two

Category	Attribute	Percentage (%)
Gender	Male	53.4
	Female	46.6
Race/Ethnicity	White	46.5
	Black	7.0
	Latino Hispanic	4.4
	East Asian	20.5
	Southeast Asian	2.4
	Indian	1.7
	Middle Eastern	17.5
Age	0-2	0.4
	3-9	6.4
	10-19	6.7
	20-29	36.4
	30-39	18.6
	40-49	14.2
	50-59	11.6
	60-69	5.0
	70+	0.7

Table 2. Demographic distribution analysis results on the OpenSDID dataset.

parallel decoding paths: an anomaly decoder for pixel-level forgery localization and a confidence decoder for detection.

ObjectFormer [43] is a transformer-based architecture designed for image manipulation detection and localization. It takes both RGB domain and frequency domain (DCT) as input.

IML-ViT [27] employs a specialized architecture that combines a windowed Vision Transformer (ViT) backbone, which alternates between windowed and global attention blocks to process high-resolution (1024×1024) input images, with a Simple Feature Pyramid Network (SFPN) for multi-scale feature extraction.

DeCLIP [39] leverages CLIP ViT-L/14 as its image encoder and employs a convolutional-based architecture as its decoder. The CLIP image encoder is kept frozen, while only the mask decoder is trained for deepfake localization.

CNNDet [45] is a standard image classifier trained on images generated by a single CNN generator (ProGAN). Through careful data augmentations, it can successfully detect AI-generated images across multiple different architectures and datasets.

UniFD [32] detects AI-generated fake images by extracting features from a frozen CLIP-ViT model and then classifying these features using either nearest neighbor classification or an MLP classifier.

NPR [40] detects synthetic images by analyzing patterns in the relationships between neighboring pixels, which result from upsampling operations in generative networks.

GramNet [26] detects fake faces by analyzing global texture patterns in images, utilizing a specialized neural network that extracts and compares statistical features from real and AI-generated facial textures.

FreqNet [41] integrates a high-frequency representation module and frequency convolutional layer into a lightweight CNN architecture, processing both phase and amplitude spectra between FFT and IFFT operations while forcing continuous focus on high-frequency information to effectively detect deepfakes.

RINE [18] detects AI-generated images by extracting features from multiple intermediate layers of CLIP's Vision Transformer. These extracted features are then used to train a binary classifier that distinguishes between real and fake images.

Methods like CAT-Net and MVSS-Net significantly enhance detection accuracy through their sophisticated integration of multi-domain data, including RGB images, frequency domain information, and noise patterns. Approaches such as ObjectFormer and IML-ViT leverage transformer-based architectures to achieve precise detection and localization capabilities. More recent innovations, including DeCLIP and UniFD, harness the power of pre-trained CLIP models for robust feature extraction, pairing them with specialized decoders or classifiers to achieve state-of-the-art detection performance.

The open-world nature of the OpenSDI introduces several challenges. First, the diversity of user preferences and the constant innovation in diffusion models make it difficult to train models that generalize well. Second, the wide range of manipulation scopes, from global image synthesis to local edits, requires models to be robust across different scales and types of modifications.

MaskCLIP advances beyond existing approaches through several key architectural innovations. Unlike methods such as CAT-Net and MVSS-Net that primarily focus on specific artifacts or dual-stream architectures, MaskCLIP leverages a more comprehensive approach through its SPM framework. This framework uniquely combines CLIP's semantic understanding capabilities with MAE's reconstruction power, creating a more robust foundation for detection and localization tasks.

Though some previous studies also leverage pre-trained models, such as CLIP [23, 39] and SAM [22], we integrate both CLIP and MAE to achieve more generalized outcomes. The efficacy of our design is substantiated through comprehensive experimental validation on the OpenSDID dataset, illustrating notable enhancements over current methods in terms of both detection accuracy and localization precision.

4. More Quantitative Results

In this section, in addition to assessing the efficiency and the scalability of MaskCLIP and other state-of-the-art meth-

ods, we focus on conducting a series of key additional experiments to examine the performance and capabilities of our proposed method, MaskCLIP, thoroughly across various domains and scenarios. These experiments aim to provide a comprehensive understanding of MaskCLIP's effectiveness in diverse contexts, ranging from traditional image manipulation detection to more advanced challenges posed by AI-generated content.

The main paper focuses on studying the proposed OpenSDI challenge and evaluates crossing the constructed OpenSDI datasets (e.g., SD1.5, Flux.1). In contrast, the evaluation across OpenSDI and non-OpenSDI (other public) datasets is used to study the broader benefits of our OpenSDI dataset and therefore is included in the suppl. material (Tables 5 and 6).

Scalability Analysis. To evaluate the data scalability of MaskCLIP, we conducted experiments by training MaskCLIP and the SOTA method TruFor on different proportions of the OpenSDID dataset. Table 3 presents the pixel-level F1 scores for both methods when trained on 25%, 50%, and 100% of the data. As shown, the performance of MaskCLIP improves as the training data size increases from 25% to 100%, demonstrating its favorable data scaling properties and ability to leverage larger datasets for enhanced performance in OpenSDI tasks.

Data Proportion	TruFor	MaskCLIP
25%	0.5294	0.5737
50%	0.6062	0.5906
100%	0.7100	0.7563

Table 3. Pixel-level F1 scores of MaskCLIP & TruFor (SOTA) on different proportions of the OpenSDID dataset.

Cross-Dataset Evaluation on Traditional Image Forgery Detection and Localization (IMDL) Benchmarks. Table 4 presents a comprehensive comparison between our proposed MaskCLIP method and state-of-the-art image manipulation detection methods across five established forensics datasets (COVERAGE, Columbia, NIST16, CASIAv1, and IMD2020). Following the IMDL-BenCo framework [28], we adopt Protocol-MVSS, where all models are trained exclusively on the CASIAv2 dataset and evaluated directly on other datasets without fine-tuning, enabling a true assessment of zero-shot domain generalization capabilities. The experimental results demonstrate the superior performance of our method across most evaluation scenarios. Notably, our approach achieves the highest average F1 score and outperforms existing methods on three out of five benchmarks. It is important to note that all test benchmarks focus on traditional image manipulation detection tasks, specifically addressing conventional manipulation techniques such as copy-paste and splicing operations. The primary variations across these datasets stem from differences in image content domains and resolutions. But our OpenSDI challenge tackles a fundamentally different problem: detecting and localizing manipulations generated by advanced AI models, particularly diffusion-based T2I generation methods.

Cross-Dataset Evaluation on Another AI-generated Image Benchmark Dataset (GenImage). To further evaluate the performance of our proposed method, we conduct experiments on the recently introduced GenImage benchmark. This benchmark allows us to compare methods trained on the GenImage SDv1.4 subset with those trained on our OpenSDID. The evaluation is divided into two groups: in-dataset evaluation and cross-dataset zero-shot evaluation.

In the in-dataset evaluation, the methods are trained and tested on the same distribution. In contrast, the cross-dataset zero-shot evaluation testing models trained on OpenSDID on GenImage dataset. Specifically, GenImage uses ImageNet as its real source data, while OpenSDID is composed of web-collected images, introducing a significant domain gap between the two datasets.

The results in Table 5 demonstrate not only our method's performance but also highlight OpenSDID's superiority in terms of image diversity and realism. The table presents accuracy metrics for various methods across different testing subsets, including Midjourney, SD V1.4, SD V1.5, ADM, GLIDE, Wukong, VQDM, and BigGAN, along with the average accuracy. The results reveal that methods trained on OpenSDID demonstrate superior performance in cross-dataset zero-shot evaluation, despite the domain gap. This suggests that OpenSDID provides a more challenging and diverse training set, enhancing model generalization capabilities. Our proposed method, MaskCLIP, achieves a competitive average accuracy of 77.4%, surpassing several state-of-the-art methods in the cross-dataset setting.

Zero-shot Evaluation on OpenSDID with Different Pretrained Dataset. We conduct comprehensive zero-shot evaluations to assess the generalization capability of various state-of-the-art forgery detection methods across different training settings. Table 6 presents the pixel-level localization performance using IoU and F1 metrics on OpenSDID test sets. We evaluate three representative methods: TruFor [12], DeCLIP [39], and IML-ViT [27], each tested with their officially released pretrained weights and trained on our OpenSDID.

Several key observations emerge: First, when using original pretrained weights, all methods show limited generalization to T2I diffusion-generated images, with performance declining significantly compared to their reported results on traditional forgery datasets. This indicates that existing methods trained on traditional manipulation data struggle to transfer to diffusion-based forgeries. Second, retraining these methods on OpenSDID substantially improves their performance, particularly for SD1.5 and SD2.1 models. For instance, TruFor's IoU improves from 0.0742

Method	COVERAGE	Columbia	NIST16	CASIAv1	IMD2020	Average
Mantra-Net [48]	0.090	0.243	0.104	0.125	0.055	0.123
MVSS-Net [4]	0.259	0.386	0.246	0.534	0.279	0.341
CAT-Net [19]	0.296	0.584	0.269	0.581	0.273	0.401
ObjectFormer [43]	0.294	0.336	0.173	0.429	0.173	0.281
PSCC-Net [25]	0.231	0.604	0.214	0.378	0.235	0.333
NCL-IML [51]	0.225	0.446	0.260	0.502	0.237	0.334
Trufor [12]	0.419	**0.865**	0.324	0.721	0.322	0.530
IML-ViT [27]	0.435	0.780	0.331	0.721	0.327	0.519
MaskCLIP	**0.451**	0.848	**0.342**	**0.725**	0.319	**0.537**

Table 4. Pixel-level F1 (localization) performance across traditional image forensics benchmark datasets. All methods are trained on CASIAv2 and do the cross-dataset evaluation on these benchmarks.

Method	Testing Subset								Avg. Acc.(%)
	Midjourney	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	
ResNet-50†	54.9	**99.9**	99.7	53.5	61.9	98.2	56.6	52.0	72.1
DeiT-S†	55.6	**99.9**	99.8	49.8	58.1	98.9	56.9	53.5	71.6
Swin-T†	62.1	**99.9**	99.8	49.8	67.6	99.1	62.3	57.6	74.8
CNNDet†	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec†	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
F3Net†	50.1	**99.9**	**99.9**	49.9	50.0	**99.9**	49.9	49.9	68.7
GramNet†	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9
PSCC-Net [25]	65.0	94.9	95.3	55.0	56.2	79.5	53.6	59.4	69.9
MVSS-Net [4]	**66.7**	84.4	85.2	60.3	62.9	69.8	67.8	70.0	70.9
TruFor [12]	55.6	46.0	45.0	66.0	60.8	51.0	70.7	70.4	58.2
DeCLIP [39]	56.3	79.8	79.0	**72.4**	**75.5**	79.9	77.5	**81.3**	75.2
MaskCLIP	51.5	95.0	96.3	70.6	**75.5**	73.8	**78.3**	77.8	**77.4**

Table 5. Image-level (detection accuracy) performance on the GenImage benchmark dataset [52]. The methods are divided into two groups based on their training data: the first group (marked with †) is trained on GenImage SDv1.4, while the second group (no marks) is trained on OpenSDID.

to 0.6342 on SD1.5 after retraining. These results highlight both the challenge and importance of developing robust detection methods specifically designed for diffusion-based forgeries, as traditional datasets show limited effectiveness in this emerging threat landscape.

More Robustness Analysis. Figure 4 shows the performance comparison of different methods under image degradation conditions, specifically Gaussian blur and JPEG compression, across three datasets (SD2.1, SDXL, and Flux.1). Under varying levels of Gaussian blur (3-23) and JPEG compression quality (60-100), MaskCLIP consistently demonstrates superior robustness compared to other approaches. Particularly notable is MaskCLIP's performance on the SD2.1 dataset, where it maintains strong F1 scores even as image quality degrades. While performance naturally decreases with more severe degradation, MaskCLIP exhibits more graceful degradation compared to competing methods, maintaining its lead across different datasets and degradation types. This consistent performance advantage highlights the robust nature of our approach in handling various real-world image quality challenges.

Comprehensive Analysis of Pretrained Model Combinations. Since our proposed Synergizing Pretrained Models (SPM) learning scheme leverages multiple pretrained models, with MaskCLIP serving as just one implementation example, we conducted extensive experiments with various model combinations. We explored different CLIP variants (ViT-B/32, OpenCLIP) and alternative pixel-wise encoders to thoroughly evaluate the SPM approach. Table 7 presents a systematic comparison of different encoder combinations. The OpenAI ViT-B/32 paired with MAE-base achieves F1 scores of 0.7227 and 0.4472 on SD1.5 and SD3 datasets, respectively. Scaling up to OpenCLIP ViT-L/14 yields improved performance, demonstrating the advantages of a larger vision transformer architecture. Notably, substituting MAE-base with Dinov2-base results in decreased performance (F1 scores of 0.6278 and

Method	Data	SD1.5 IoU	SD1.5 F1	SD2.1 IoU	SD2.1 F1	SDXL IoU	SDXL F1	SD3 IoU	SD3 F1	Flux.1 IoU	Flux.1 F1	AVG IoU	AVG F1
TruFor [12]	Trufor†	0.0742	0.1073	0.0770	0.1115	0.0704	0.1035	0.0996	0.1424	0.1019	0.1464	0.0846	0.1222
TruFor [12]	OpenSDID	0.6342	0.7100	0.5467	0.6188	0.2655	0.3185	0.3229	0.3852	0.0760	0.0970	0.3691	0.4259
DeCLIP [39]	Dolos-LDM	0.0138	0.0218	0.0131	0.0210	0.0089	0.0144	0.0145	0.0230	0.0070	0.0115	0.0115	0.0183
DeCLIP [39]	Dolos-Lama	0.0093	0.0151	0.0098	0.0158	0.0057	0.0098	0.0155	0.0251	0.0048	0.0085	0.0090	0.0149
DeCLIP [39]	Dolos-Pluralistic	0.0145	0.0233	0.0154	0.0245	0.0064	0.0108	0.0182	0.0292	0.0069	0.0116	0.0123	0.0199
DeCLIP [39]	Dolos-Repaint	0.0254	0.0377	0.0221	0.0342	0.0184	0.0284	0.0344	0.0522	0.0162	0.0253	0.0233	0.0356
DeCLIP [39]	OpenSDID	0.3718	0.4344	0.3569	0.4187	0.1459	0.1822	0.2734	0.3344	0.1121	0.1429	0.2520	0.3025
IML-ViT [27]	Trufor†	0.0806	0.1143	0.0825	0.1165	0.0746	0.1066	0.1279	0.1750	0.1295	0.1768	0.0990	0.1378
IML-ViT [27]	CASIAv2	0.0248	0.0384	0.0228	0.0366	0.0213	0.0337	0.0266	0.0418	0.0290	0.0460	0.0249	0.0393
IML-ViT [27]	OpenSDID	0.6651	0.7362	0.4479	0.5063	0.2149	0.2597	0.2363	0.2835	0.0611	0.0791	0.3251	0.3730
MaskCLIP	CASIAv2	0.0312	0.0465	0.0289	0.0442	0.0256	0.0398	0.0334	0.0502	0.0358	0.0532	0.0310	0.0468
MaskCLIP	OpenSDID	0.6712	0.7563	0.5550	0.6289	0.3098	0.3700	0.4375	0.5121	0.1622	0.2034	0.4271	0.4941

Table 6. Pixel-level (localization) performance on OpenSDID. Test SOTA methods pretrained weights on various training data and do zero-shot testing on the OpenSDID's test sets. Trufor† indicates training on a combined dataset including CASIA v2, FantasticReality [17], IMD2020, and tampered versions of COCO and RAISE datasets [19], which is the training setting of Trufor.

Method			Pixel-level F1	
Encoder1	Encoder2	Params.	SD1.5	SD3
OA ViT-B/32	MAE-base	96M	0.7227	0.4472
OC ViT-L/14	MAE-base	114M	0.7363	0.4908
OA ViT-L/14	Dinov2-base	114M	0.6278	0.3521
OA ViT-L/14	MAE-base	114M	0.7563	0.3700
OA ViT-L/14	SAM-base	126M	**0.7873**	**0.5773**

Table 7. Performance comparison of different encoder combinations. We evaluate various pretrained models as Encoder1 (CLIP variants) and Encoder2 (pixel-wise encode models). OA and OC denote OpenAI and OpenCLIP respectively. The results show pixel-level F1 scores on both SD1.5 and SD3 datasets. Params. indicates the trainable parameters of the model, where we only keep the CLIP model frozen.

Loss Components			Pixel-level F1	
$\mathcal{L}_{CE}$	$\mathcal{L}_{BCE}$	$\mathcal{L}_{EDG}$	SD1.5	SD3
1.0	2.0	1.0	0.7575	0.3573
2.0	1.0	1.0	0.7620	0.3373
1.0	1.0	0.0	0.7299	0.3108
1.0	1.0	1.0	0.7563	0.3700

Table 8. Ablation study on three loss function components. $\mathcal{L}_{CE}$ denotes cross-entropy loss, $\mathcal{L}_{BCE}$ represents binary cross-entropy loss, and $\mathcal{L}_{EDG}$ is the edge-weighted loss.

Method	SD1.5	SD2.1	SDXL	SD3	Flux.1	AVG
TruFor [12]	$0.728_{\pm 0.032}$	$0.601_{\pm 0.029}$	$0.289_{\pm 0.059}$	$0.330_{\pm 0.018}$	$0.071_{\pm 0.018}$	$0.404_{\pm 0.031}$
MaskCLIP	$0.765_{\pm 0.008}$	$0.628_{\pm 0.021}$	$0.381_{\pm 0.013}$	$0.492_{\pm 0.019}$	$0.188_{\pm 0.009}$	$0.471_{\pm 0.014}$

Table 9. Performance of TruFor and MaskCLIP with multiple runs on OpenSDID.

0.3521), suggesting that MAE's self-supervised pretraining approach is more effective for our specific task. The optimal performance is achieved by combining OpenAI ViT-L/14 with SAM-base, reaching F1 scores of 0.7873 and 0.5773. This superior performance indicates that SAM's segment-focused pretraining provides particularly valuable features for our segmentation task, albeit at the cost of increased computational overhead during both training and inference. These comprehensive experiments underscore two key findings: (1) the critical importance of selecting appropriate pretrained weights, and (2) the significant performance benefits that can be achieved through larger model architectures and segment-aware pretraining strategies.

Analysis of Loss Function Components. The loss function in MaskCLIP follows established approaches from previous works [4, 12, 27], incorporating three components: cross-entropy loss ($\mathcal{L}_{CE}$), binary cross-entropy loss ($\mathcal{L}_{BCE}$), and edge-weighted loss ($\mathcal{L}_{EDG}$). While we employ a simple balanced weighting scheme for these loss terms, it is crucial to analyze the impact of each component in our objective function. Table 8 presents experimental results with various weight combinations of the three loss terms. Our findings underscore the importance of maintaining a balanced objective function that appropriately weights binary classification and edge preservation.

Multiple Runs Analysis. Although most previous works don't perform multiple runs to study robustness, we conducted three independent training runs for both MaskCLIP and the baseline TruFor method. As shown in Table 9, MaskCLIP exhibits notably smaller standard deviations across all subsets, indicating more stable and reliable performance compared to the baseline.

5. More Qualitative Results

To provide a comprehensive evaluation of our method across different Diffusion models, we present additional qualitative results on various subsets of the OpenSDID dataset. Figures 5, 6, 7, 8, and 9 demonstrate our approach's performance on images generated by SD1.5, SD2.1, SD3, SDXL, and Flux.1 models respectively.

The results showcase our method's effectiveness across various generator models. For the Flux.1 subset (Figure 9),

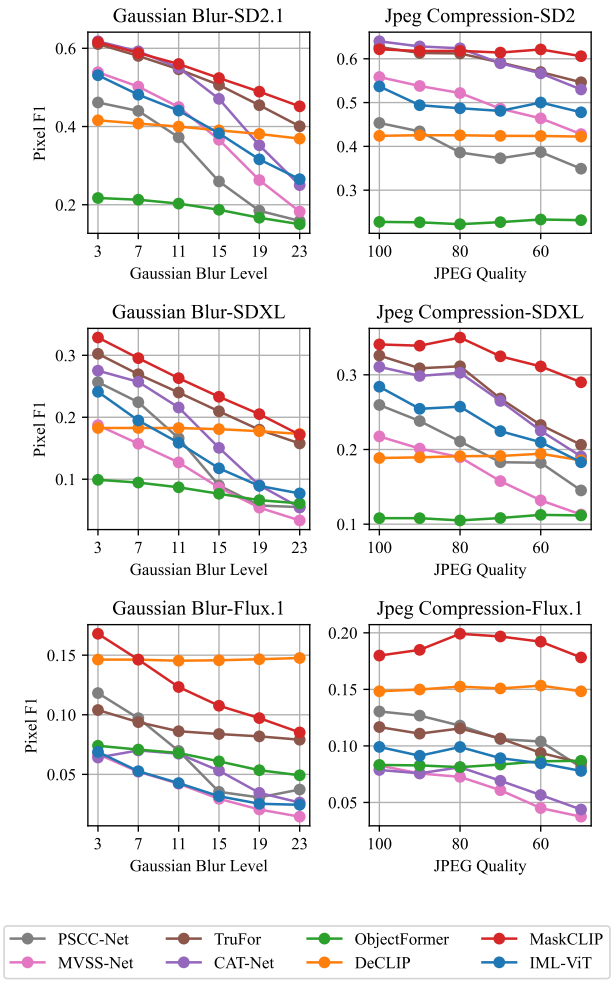

Figure 4. Robustness evaluation of different SOTA methods under image degradation on OpenSDID. It compares performance across varying levels of Gaussian Blur (left) and JPEG Compression (right).

our method exhibits strong zero-shot performance. However, in the interest of transparency, we also present several failure cases (Figure 10) that highlight our method's current limitations. These challenging cases typically involve highly photorealistic images or compositions with unique elements that closely resemble natural photographs. These observations highlight the significant advancement of current AI image generation technology. This rapid development poses higher demands on image discrimination techniques and indicates that future research needs to develop more robust detection methods.

6. Samples of OpenSDID

Figures 11, Figures 12, Figures 13, Figures 14, and Figures 15 present more samples of our datasets.

References

[1] Meta AI. Llama 3.2-vision: Multimodal large language model. Hugging Face Model Hub, 2024. Model release date: September 25, 2024. 1

[2] Sithu Aung. Easyface: Easy face analysis tool with sota models. https://github.com/sithu31296/EasyFace, 2022. 3

[3] Ollin Boer Bohan. Megalith-10m: A dataset of public domain photographs. https://huggingface.co/datasets/madebyollin/megalith-10m, 2024. Accessed: 2024-10-07. 1, 3

[4] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14185–14193, 2021. 4, 7, 8

[5] Yirui Chen, Xudong Huang, Quan Zhang, Wei Li, Mingjian Zhu, Qiangyu Yan, Simiao Li, Hanting Chen, Hailin Hu, Jie Yang, et al. Gim: A million-scale benchmark for generative image manipulation detection and localization. *arXiv preprint arXiv:2406.16531*, 2024. 4

[6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 1

[7] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020. 2, 4

[8] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 4

[9] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China summit and international conference on signal and information processing*, pages 422–426. IEEE, 2013. 4

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2

[11] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 63–72. IEEE, 2019. 4

[12] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20606–20615, 2023. 4, 6, 7, 8

CVPR
#7349

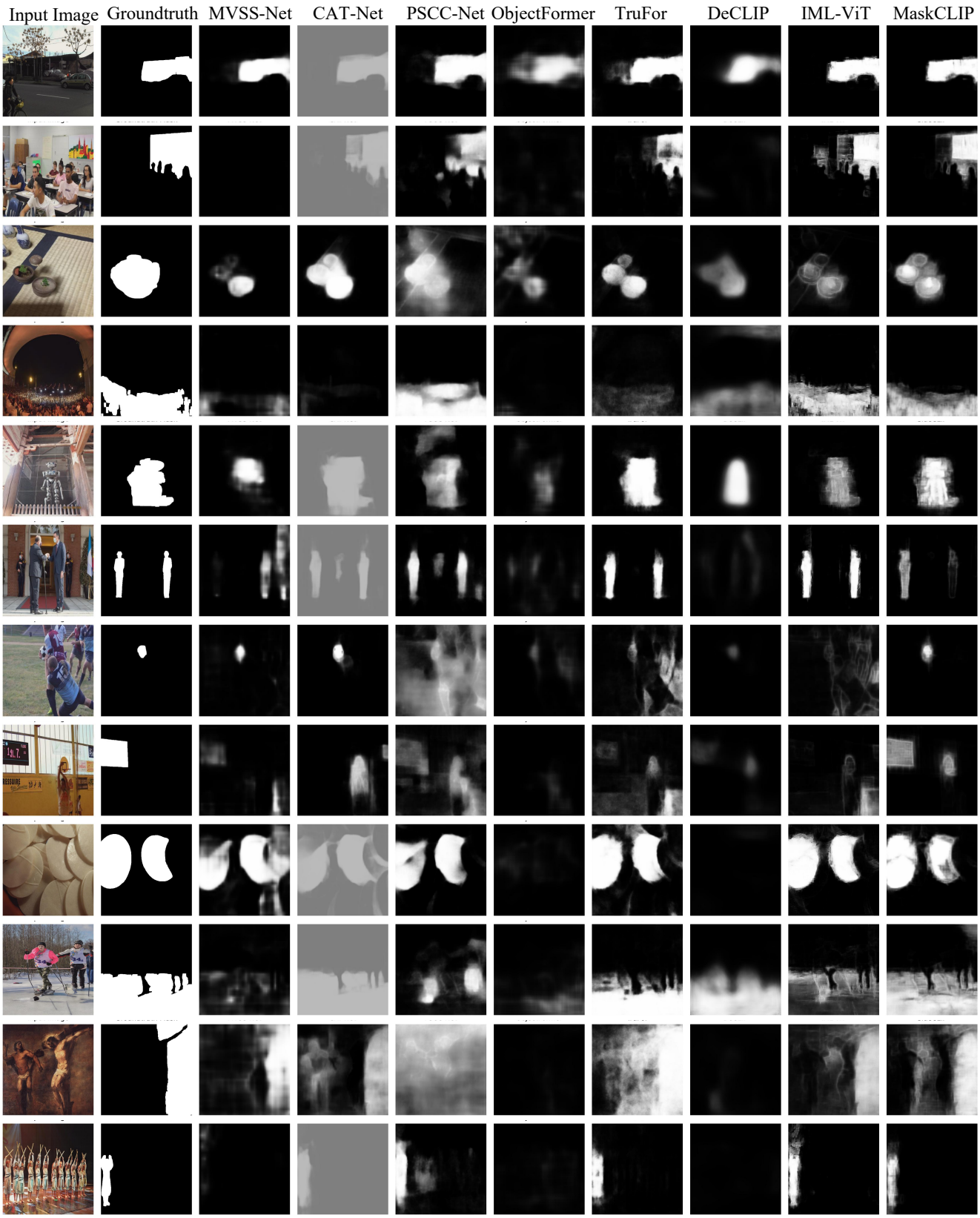

Figure 5. Qualitative results on the OpenSDID SD1.5 testset.

CVPR
#7349

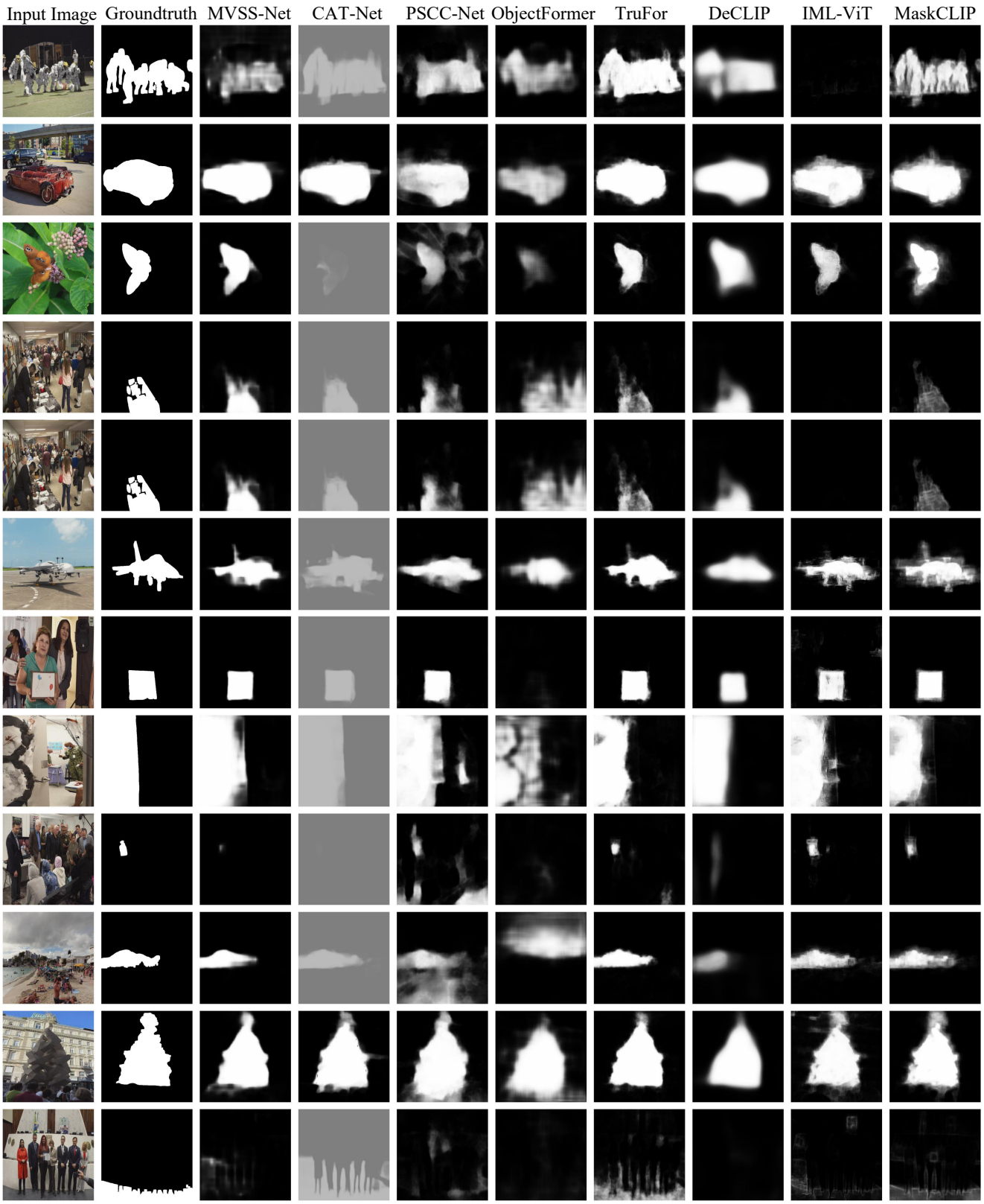

Figure 6. Qualitative results on the OpenSDID SD2.1 test set.

CVPR
#7349

Figure 7. Qualitative results on the OpenSDID SD3 test set.

CVPR
#7349

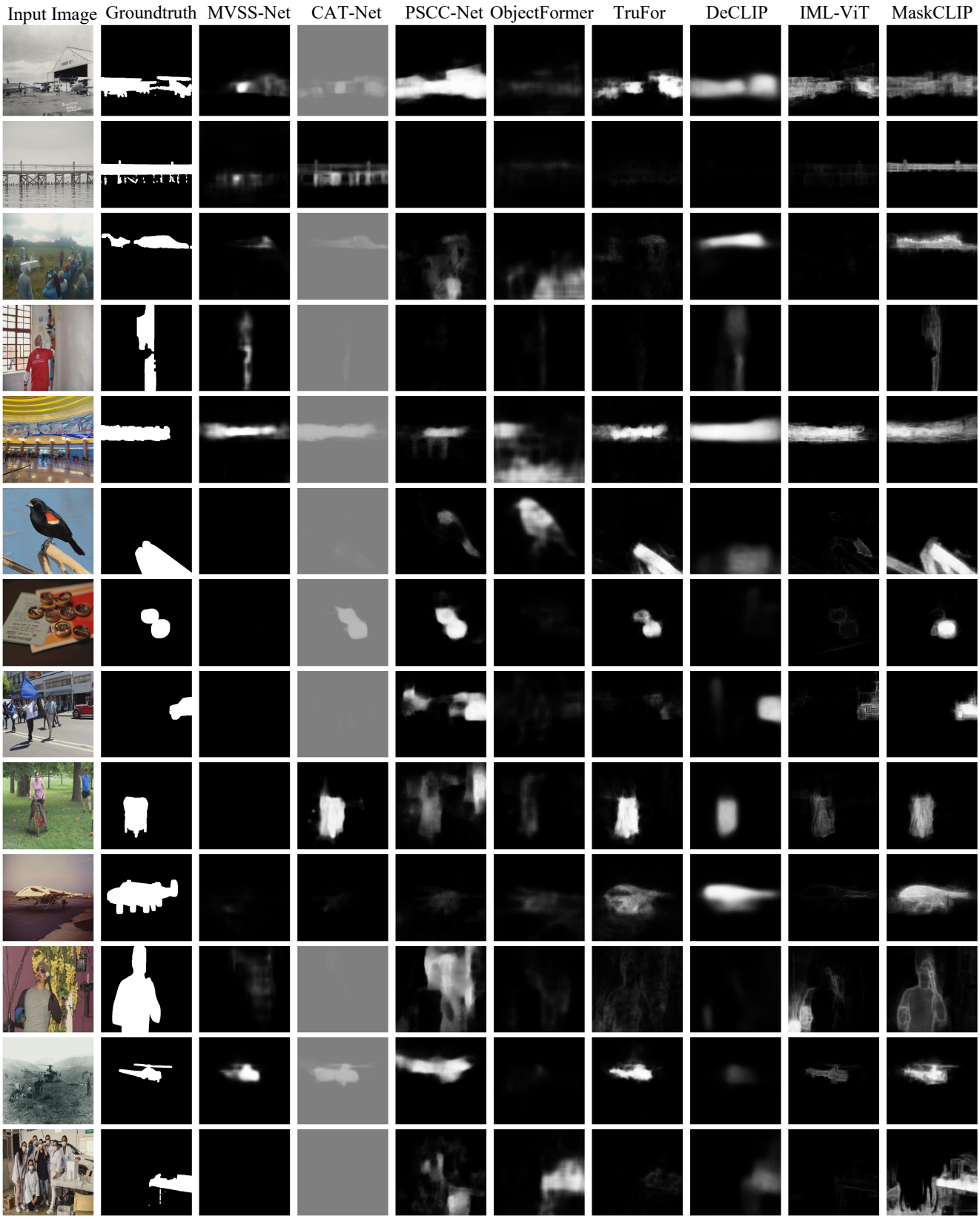

Figure 8. Qualitative results on the OpenSDID SDXL test set.

CVPR
#7349

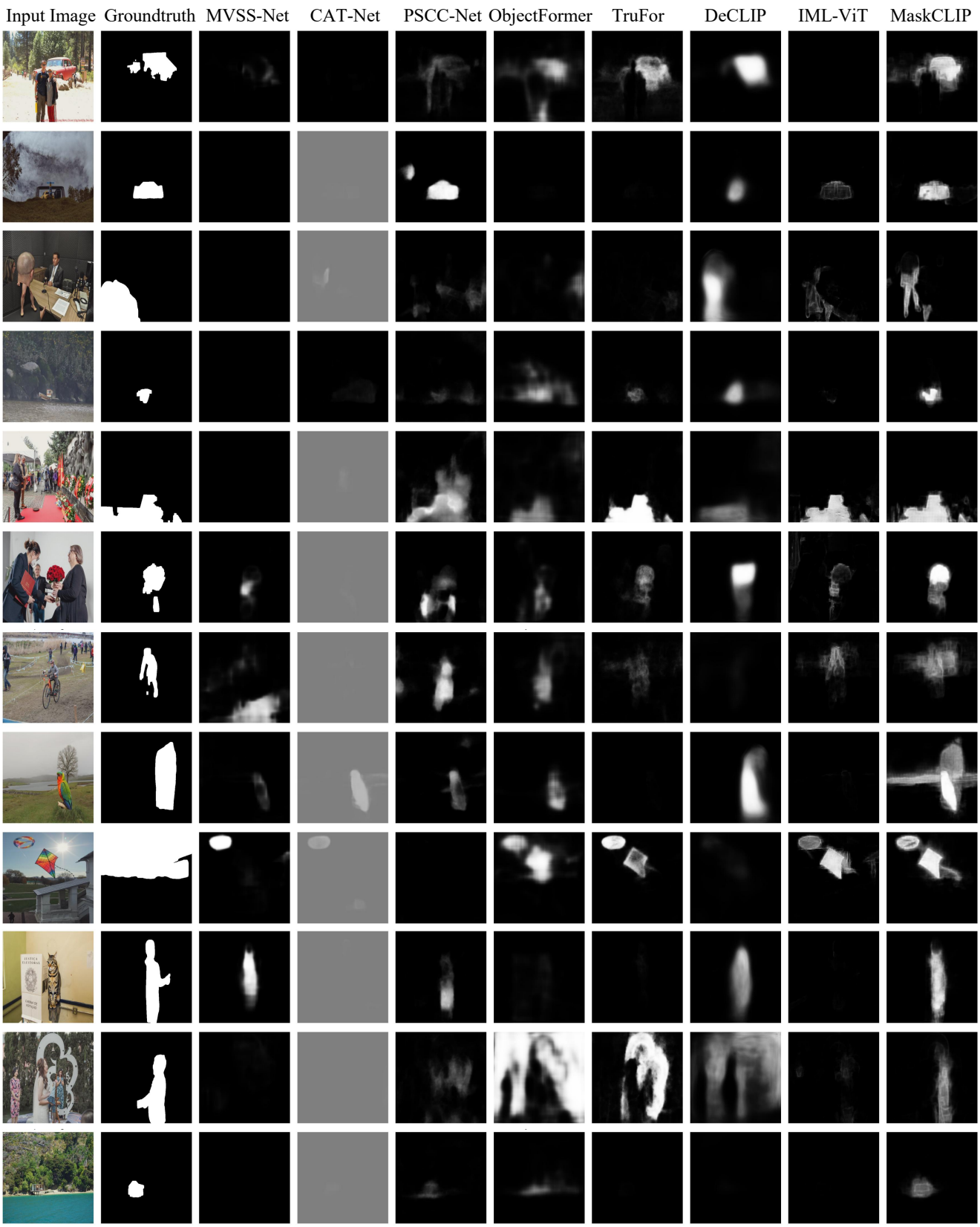

Figure 9. Qualitative results on the OpenSDID Flux.1 test set.

CVPR
#7349

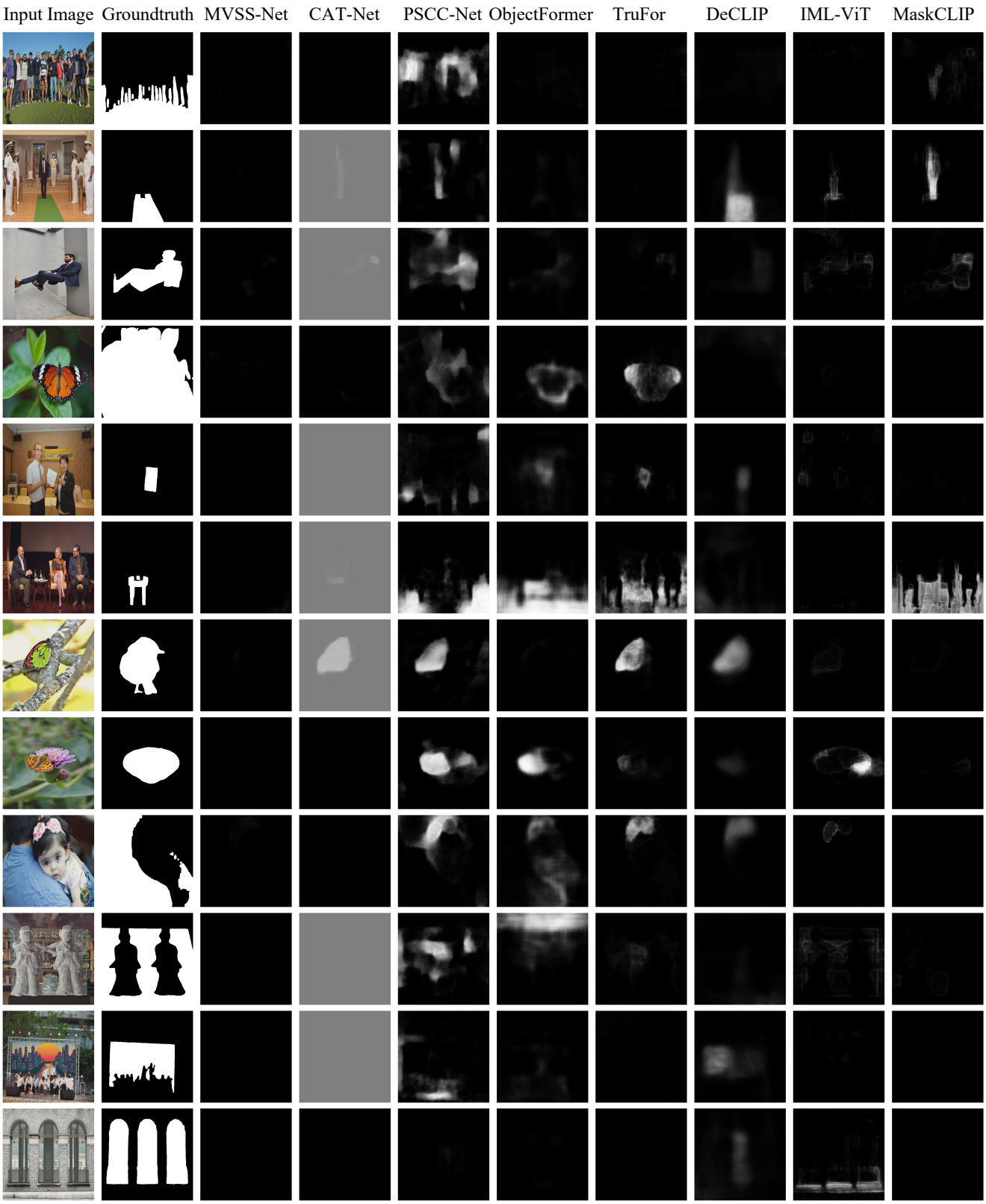

Figure 10. Some Fail Cases on the OpenSDID Flux.1 test set.

Figure 11. Sample images generated using the Stable Diffusion v1.5 (SD1.5) model.

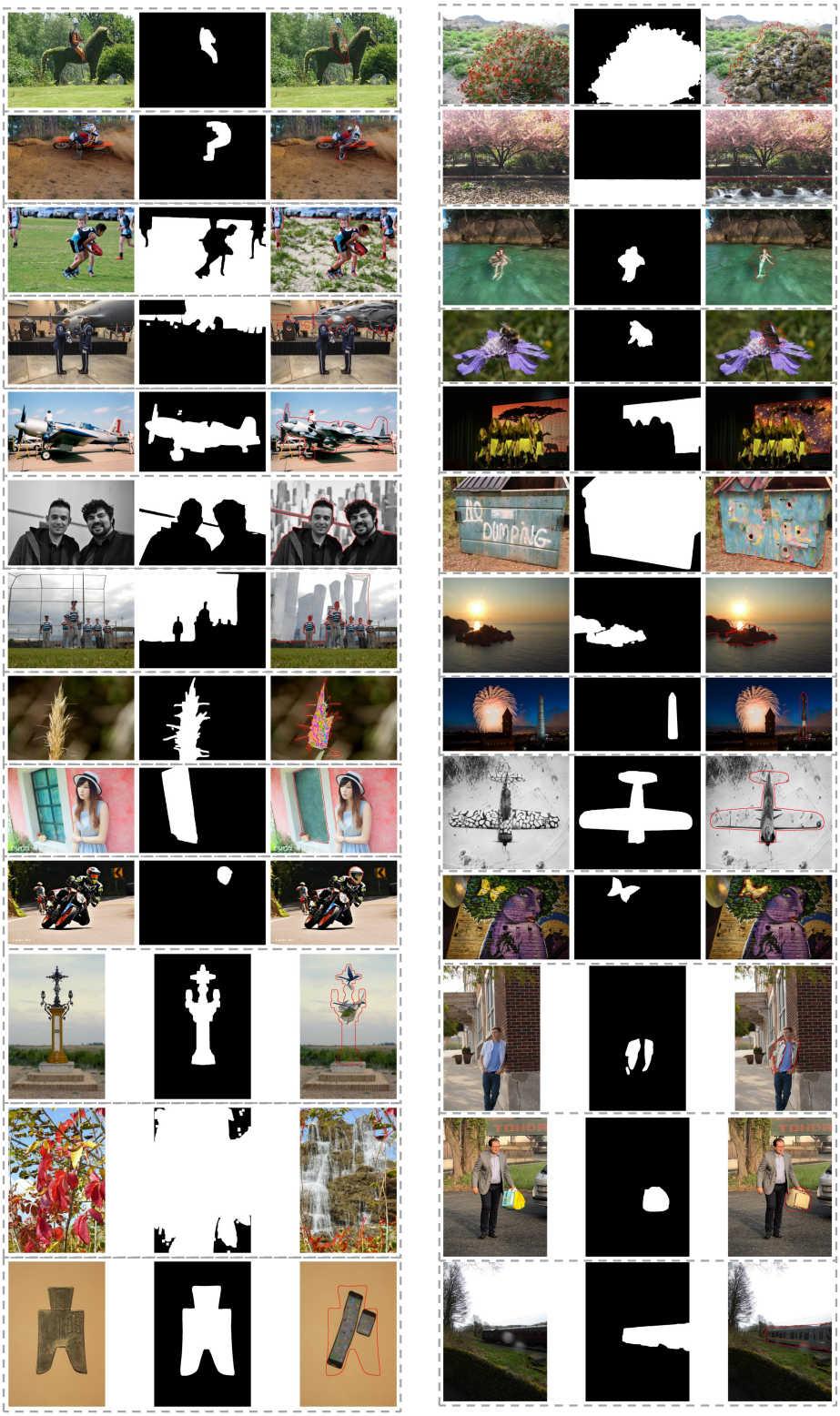

Figure 12. Sample images generated using the Stable Diffusion v2.1 (SD2.1) model.

CVPR
#7349

Figure 13. Sample images generated using the Stable Diffusion v3 model.

Figure 14. Sample images generated using the Stable Diffusion XL (SDXL) model.

Figure 15. Sample images generated using the Flux.1 model.

[13] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023. 2, 4

[14] Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. Autosplice: A text-prompt manipulated image dataset for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 893–903, 2023. 4

[15] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2889–2898, 2020. 4

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[17] Vladimir V Kniaz, Vladimir Knyaz, and Fabio Remondino. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. *Advances in neural information processing systems*, 32, 2019. 8

[18] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *Computer Vision – ECCV 2024*, pages 394–411, Cham, 2025. Springer Nature Switzerland. 5

[19] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022. 3, 7, 8

[20] Gant Laborde. Deep nn for nsfw detection. https://github.com/GantMan/nsfw_model, 2019. 3

[21] Black Forest Labs. Flux.1: A 12 billion parameter rectified flow transformer for text-to-image generation. https://github.com/black-forest-labs/flux, 2024. A 12 billion parameter rectified flow transformer capable of generating images from text descriptions. 2

[22] Yingxin Lai, Zhiming Luo, and Zitong Yu. Detect any deepfakes: Segment anything meets face forgery detection and localization. In *Chinese Conference on Biometric Recognition*, pages 180–190. Springer, 2023. 5

[23] Dong Li, Jiaying Zhu, Xueyang Fu, Xun Guo, Yidi Liu, Gang Yang, Jiawei Liu, and Zheng-Jun Zha. Noise-assisted prompt learning forimage forgery detection and localization. In *Computer Vision – ECCV 2024*, pages 18–36, Cham, 2025. Springer Nature Switzerland. 5

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[25] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11):7505–7517, 2022. 4, 7

[26] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8060–8069, 2020. 5

[27] Xiaochen Ma, Bo Du, Xianggen Liu, Ahmed Y Al Hammadi, and Jizhe Zhou. Iml-vit: Image manipulation localization by vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 5, 6, 7, 8

[28] Xiaochen Ma, Xuekang Zhu, Lei Su, Bo Du, Zhuohang Jiang, Bingkui Tong, Zeyu Lei, Xinyu Yang, Chi-Man Pun, Jiancheng Lv, et al. Imdl-benco: A comprehensive benchmark and codebase for image manipulation detection & localization. *arXiv preprint arXiv:2406.10580*, 2024. 3, 6

[29] Hannes Mareen, Dimitrios Karageorgiou, Glenn Van Wallendael, Peter Lambert, and Symeon Papadopoulos. Tgif: Text-guided inpainting forgery dataset. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2024. 4

[30] Tian-Tsong Ng, Shih-Fu Chang, and Q Sun. A data set of authentic and spliced image blocks. *Columbia University, ADVENT Technical Report*, 4, 2004. 4

[31] Adam Novozamsky, Babak Mahdian, and Stanislav Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 71–80, 2020. 4

[32] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 5

[33] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[35] Javier Rando, Daniel Paleka, David Lindner, Lennard Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. 3

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[37] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 4

[38] Christoph Schuhmann and LAION-AI. Clip-based nsfw detector. https://github.com/LAION-AI/CLIP-based-NSFW-Detector, 2022. 3

[39] Stefan Smeu, Elisabeta Oneata, and Dan Oneata. Declip: Decoding clip representations for deepfake localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 5, 6, 7, 8

[40] Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection, 2023. 5

[41] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space learning, 2024. 5

[42] Dragos-Constantin Tantaru, Elisabeta Oneta, and Dan Oneta. Weakly-supervised deepfake localization in diffusion-generated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6258–6268, 2024. 4

[43] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022. 5, 7

[44] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[45] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 4, 5

[46] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–911, 2023. 2, 4

[47] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. Coverage—a novel database for copy-move forgery detection. In *2016 IEEE international conference on image processing (ICIP)*, pages 161–165. IEEE, 2016. 4

[48] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[49] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 1

[50] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019. 4

[51] Jizhe Zhou, Xiaochen Ma, Xia Du, Ahmed Y. Alhammadi, and Wentao Feng. Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22346–22356, 2023. 7

[52] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4, 7