International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02255-9

Language-Guided Hierarchical Fine-Grained Image Forgery Detection and Localization

Xiao Guo[1] · **Xiaohong Liu**[2] · **Iacopo Masi**[3] · **Xiaoming Liu**[1]

Received: 23 September 2023 / Accepted: 23 September 2024

Abstract

Differences in forgery attributes of images generated in CNN-synthesized and image-editing domains are large, and such differences make a unified image forgery detection and localization (IFDL) challenging. To this end, we present a hierarchical fine-grained formulation for IFDL representation learning. Specifically, we first represent forgery attributes of a manipulated image with multiple labels at different levels. Then, we perform fine-grained classification at these levels using the hierarchical dependency between them. As a result, the algorithm is encouraged to learn both comprehensive features and the inherent hierarchical nature of different forgery attributes, thereby improving the IFDL representation. In this work, we propose a Language-guided Hierarchical Fine-grained IFDL, denoted as HiFi-Net++. Specifically, HiFi-Net++ contains four components: multi-branch feature extractor, language-guided forgery localization enhancer, as well as classification and localization modules. Each branch of the multi-branch feature extractor learns to classify forgery attributes at one level, while localization and classification modules segment the pixel-level forgery region and detect image-level forgery, respectively. In addition, the language-guided forgery localization enhancer (LFLE), containing image and text encoders learned by contrastive language-image pre-training (CLIP), is used to further enrich the IFDL representation. LFLE takes specifically designed texts and the given image as multi-modal inputs and then generates the visual embedding and manipulation score maps, which are used to further improve HiFi-Net++ manipulation localization performance. Lastly, we construct a hierarchical fine-grained dataset to facilitate our study. We demonstrate the effectiveness of our method on 8 different benchmarks for both tasks of IFDL and forgery attribute classification. Our source code and dataset can be found: github.com/CHELSEA234/HiFi-IFDL.

Keywords Forgery detection · Manipulation localization · Contrastive language-image pre-training · Large language model

1 Introduction

Chaotic and pervasive multimedia information sharing offers better means for spreading misinformation cnn (2010), and the forged image content could, in principle, sustain recent "infodemics" inf (2022). Firstly, Convolutional Neural Networks (CNN) synthesized images made extraordinary leaps culminating in recent synthesis methods—Dall·E Ramesh et al. (2022) or Google ImageN Saharia et al. (2022)—based on diffusion models (DDPM) Ho et al. (2020), which even generate realistic videos from text Singer et al. (2022); Ho et al. (2022). Secondly, the availability of image editing tool-kits produced the substantially low-cost access to image forgery or tampering (e.g., splicing and inpainting). In response to such an issue of image forgery, the computer vision community has made considerable efforts, which however branch separately into two directions: detecting either CNN synthesis Zhang et al. (2019b); Wang et al. (2020b);

Communicated by Sergio Escalera.

✉ Xiao Guo
guoxia11@msu.edu

Xiaohong Liu
xiaohongliu@sjtu.edu.cn

Iacopo Masi
masi@di.uniroma1.it

Xiaoming Liu
liuxm@msu.edu

[1] Department of Computer Science and Engineering, Michigan State University, East Lansing, USA

[2] John Hopcroft Center for Computer Science, Shanghai Jiao Tong University, Shanghai, China

[3] Department of Computer Science, Sapienza, University of Rome, Rome, Italy

Published online: 10 December 2024

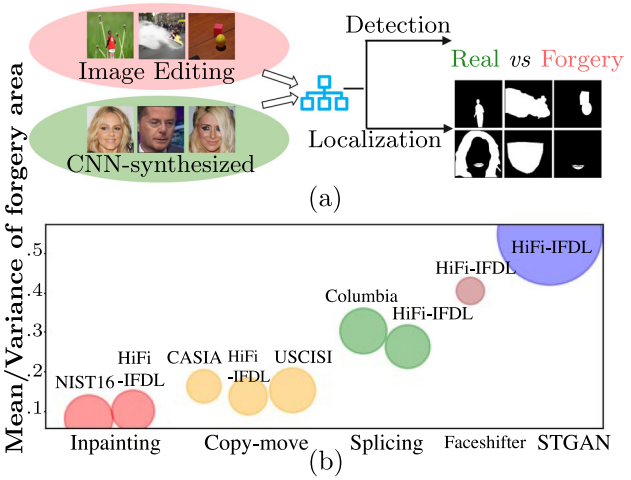


Fig. 1 **a** In this work, we study image forgery detection and localization (IFDL), regardless of forgery method domains. **b** The distribution of forgery region depends on individual forgery methods. Each color represents one forgery category (x-axis). Each bubble represents one image forgery dataset. The y-axis denotes the average of forgery area. The bubble's area is proportional to the variance of the forgery area

Stehouwer et al. (2020) or conventional image editing Wu et al. (2019); Hu et al. (2020); Liu et al. (2022); Dong et al. (2022); Wang et al. (2022). As a result, these methods may be ineffective when deploying to real-life scenarios, where forged images can possibly be generated from either CNN-synthesized or image-editing domains.

To push the frontier of image forensics Sencar et al. (2022), we study the image forgery detection and localization problem (IFDL)—Fig. 1a—regardless of the forgery method domains, i.e., CNN-synthesized or image editing. It is challenging to develop a unified algorithm for two domains, as images, generated by different forgery methods, differ largely from each other in terms of various forgery attributes. For example, a forgery attribute can indicate whether a forged image is fully synthesized or partially manipulated, or whether the forgery method used is the diffusion model generating images from the Gaussian noise or an image editing process that splices two images via Poisson editing Pérez et al. (2003). Therefore, to model such complex forgery attributes, we first represent forgery attribute of each forged image with multiple labels at different levels. Then, we present a hierarchical fine-grained formulation for IFDL, which requires the algorithm to classify fine-grained forgery attributes of each image at different levels, via the inherent hierarchical nature of different forgery attributes.

Figure 2a shows the interpretation of the forgery attribute with a hierarchy, which evolves from the general forgery attribute, fully-synthesized vs partial-manipulated, to specific individual forgery methods, such as DDPM Ho et al. (2020) and DDIM Song et al. (2021). Then, given an input image, our method performs fine-grained forgery attribute classification at different levels (see Fig. 2b). The image-level forgery detection benefits from this hierarchy as the fine-grained classification learns the comprehensive IFDL representation to differentiate individual forgery methods. Also, for the pixel-level localization, the fine-grained classification features can serve as a prior to improve the localization. This holds since the distribution of the forgery area is prominently correlated with forgery methods, as depicted in Fig. 1b.

In Fig. 2c, we leverage the hierarchical dependency between forgery attributes in fine-grained classification. Each node's classification probability is conditioned on the path from the root to itself. For example, the classification probability at a node of `DDPM` is conditioned on the classification probability of all nodes in the path of `Forgery`→ `Fully Synthesis`→`Diffusion`→`Unconditional`→`DDPM`. This differs from prior work Wu et al. (2019); Marra et al. (2018); Yu et al. (2019); Marra et al. (2019), which assume a "flat" structure in which attributes are mutually exclusive. Predicting the entire hierarchical path helps understand forgery attributes from coarse to fine, thereby capturing dependencies among individual forgery attributes.

Therefore, we propose Hierarchical Fine-grained Network (HiFi-Net) Guo et al. (2023b), which is the preliminary version of this work and published in Proceeding of the IEEE/CVF Computer Vision and Pattern Recognition (CVPR 2023). Specifically, HiFi-Net has three components: a multi-branch feature extractor, a localization module, and a detection module. Each branch of the multi-branch extractor classifies images at one forgery attribute level. The localization module generates the forgery mask with the help of a deep-metric learning-based objective, which improves the separation between real and forged pixels.

Although the proposed HiFi-Net Guo et al. (2023b) achieves the state-of-the-art IFDL performance, its performance becomes suboptimal when the manipulation area is small, as detailed in Fig. 9 of the work Guo et al. (2023b). Furthermore, it is always desirable to improve the generalization capacity of IFDL solution Wang et al. (2020b); Ojha et al. (2023). Given these two considerations, we empirically observe that leveraging the fixed visual embedding of the pre-trained CLIP image encoder can improve the HiFi-Net's localization performance on images with small manipulation regions and boost the overall generalization ability. We believe this is because the pre-trained CLIP image encoder has two merits. First, CLIP is trained to match images to corresponding texts, which means the pre-trained CLIP visual embeddings excel at locating objects of interest, regardless of the object's spatial size—Fig. 3a. Secondly, the pre-trained CLIP is exposed to a vast number of web data, and its visual embedding is observed to generalize well for differentiating real and fake images Ojha et al. (2023). The Fig. 2 in work Wu et al. (2023) further supports this hypothesis.

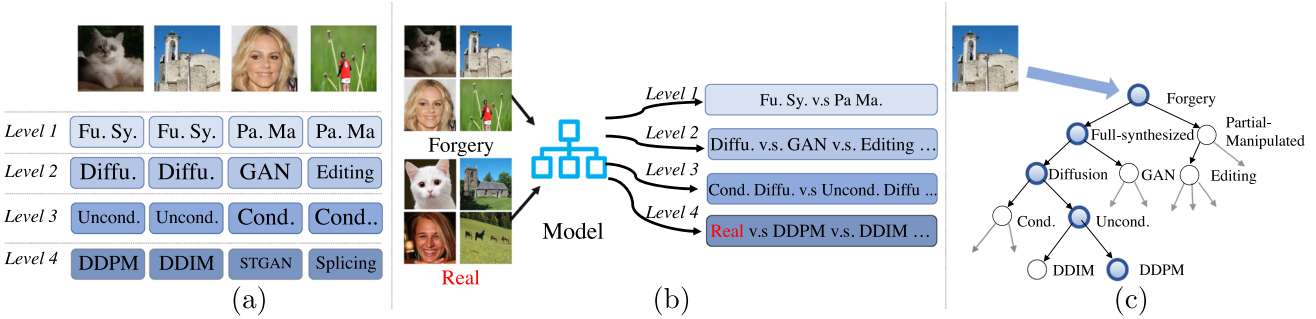


Fig. 2 **a** We represent the forgery attribute of each manipulated image with multiple labels at different levels. **b** For an input image, we encourage the algorithm to classify its fine-grained forgery attributes at different levels, i.e. a 2-way classification (fully synthesized or partially manipulated) on level 1. **c** We perform the fine-grained classification via the hierarchical nature of different forgery attributes, where each depth l node's classification probability is conditioned on classification probabilities of neighbor nodes at depth $(l-1)$. [Key: Fu. Sy.: Fully Synthesized; Pa. Ma.: Partially manipulated; Diff.: Diffusion model; Cond.: Conditional; Uncond.: Unconditional]

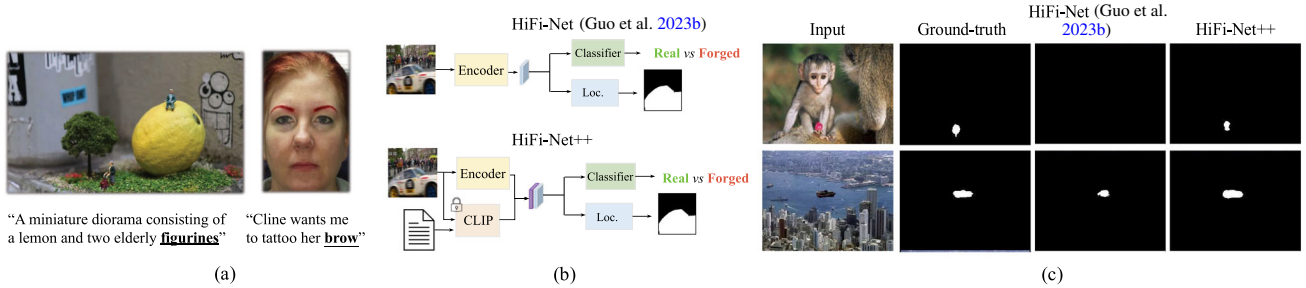


Fig. 3 **a** The pre-trained CLIP has a powerful visual embedding that can localize and recognize objects of interest, described by the text query, regardless the objects' spatial size. For example, the pre-trained CLIP can understand the existence of figurines and eyebrows, given the corresponding text query. These images and text queries are acquired via the public CLIP website https://rom1504.github.io/clip-retrieval/. **b** *Upper*: The HiFi-Net contains three modules, which are multi-branch feature extractor, classification module, and localization module. *Bottom*: We integrated the Language-guided Forgery Localization Enhancer (LFLE) into the existing HiFi-Net, which is then denoted as HiFi-Net++. **c** The HiFi-Net++ can localize manipulation accurately, even when the manipulation area is small

Also, although the pre-trained CLIP model is trained to maximize the correlation between same text-image vector pairs, the computer vision community recently, in fact, adapts this pre-trained foundation model for text-pixel matching. This adaptation involves leveraging pre-trained CLIP to achieve remarkable outcomes in pixel-wise classification tasks, such as image segmentation Zhou et al. (2022a); Xu et al. (2022, 2023); Ghiasi et al. (2022); Li et al. (2022a) and object detection Zhong et al. (2022); Rao et al. (2022). This success is attributed to the fact that the pre-trained CLIP has a powerful representation ability that matches, or grounds, the rich semantic text embedding to certain visual contents expressed by the visual embedding produced by the pre-trained CLIP image encoder.

Motivated by the above discussions, we therefore propose HiFi-Net++, which has an additional Language-guided Forgery Localization Enhancer (LFLE) module on the top of the HiFi-Net Guo et al. (2023b), as shown in Fig. 3b. This LFLE module consists of the pre-trained CLIP image and text encoder, as well as a refinement block that helps optimize the text embedding to better adapt to the current IFDL task. Specifically, given the input image, the pre-trained CLIP image encoder provides the *visual embedding*. After that, we use specifically designed templates with the name of each aforementioned forgery attribute (e.g., `fully-synthesized`, `partial-manipulated`, `GAN`, etc.) to construct text inputs, and then apply the pre-trained CLIP text encoder on these text inputs to obtain corresponding text embeddings. Such text embeddings are used with the visual features output from the pre-trained CLIP image encoder to generate a 2D manipulation score map. This manipulation score map denotes the spatial location manipulated by a given forgery method. As depicted in Fig. 4, the highly generalizable visual embedding and manipulation score maps are used altogether to help localize the manipulation. It is worth noting that, being inspired by the prior works Zhou et al. (2022b); Gao et al. (2023); Zhang et al. (2021); Yao et al. (2021), we also use a refinement block to enhance the text embedding, such that we can adapt the pre-trained CLIP text encoder and its generated text embedding for our manipulation localization tasks.

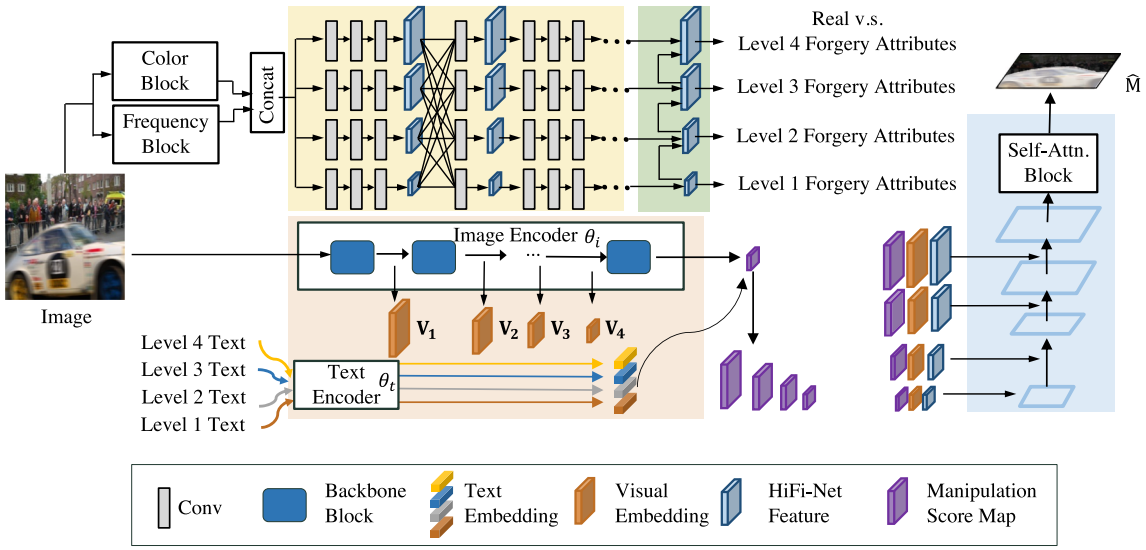


Fig. 4 Given the input image, we first leverage color and frequency blocks to extract features. The multi-branch feature extractor ( ) learns feature maps of different resolutions, and these feature maps are used for the fine-grained classification at different levels via the classification module ( ), detailed in Sect. 3.2. The language-guided localization enhancer ( ), containing the pre-trained CLIP image and text encoders (denoted as $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_t$ respectively), takes the input image and pre-defined text input, and then produce the visual embedding and the manipulation score map. The entire process is detailed in Sect. 3.3. In the end, the localization module ( ) in Sect. 3.4 jointly takes HiFi-Net feature map, visual embedding, and manipulation score map to generate the binary mask $\hat{\mathbf{M}}$ that indicates the manipulation area

Lastly, to facilitate our study of the hierarchical fine-grained formulation, we construct a new dataset, termed Hierarchical Fine-grained (HiFi) IFDL dataset. It contains 13 forgery methods, which are either the latest CNN-synthesized methods or representative image editing methods. HiFi-IFDL dataset also induces a hierarchical structure on forgery categories to enable learning a classifier for various forgery attributes. Each forged image is also paired with a high-resolution ground truth forgery mask for the localization task.

In summary, we extend the preliminary version of this work Guo et al. (2023b) in two aspects: (i) by utilizing the fixed visual embedding from the pre-trained CLIP image encoder, the HiFi-Net++ has improved the localization performance of spatially small manipulations as well as the generalization ability to unseen manipulation; (ii) we adapt the idea of text-pixel matching for the manipulation localization task—use the pre-trained CLIP text embeddings with visual features to produce the manipulation score map, which serves as an auxiliary signal to help the localization performance.

The contributions of this work are as follows:

◇ We study the task of image forgery detection and localization (IFDL) for both image editing and CNN-synthesized domains. We propose a hierarchical fine-grained formulation to learn a comprehensive representation of IFDL and forgery attribute classification.

◇ We propose an IFDL algorithm, named HiFi-Net, which not only performs well on forgery detection and localization, but also identifies a diverse spectrum of forgery attributes. Based on this proposed HiFi-Net in the preliminary version of this work, we further introduce a more advanced version IFDL algorithm, namely HiFi-Net++. HiFi-Net++ utilizes an additional language-guided forgery localization enhancer (LFLE) to improve the manipulation localization performance and its generalization ability.

◇ We construct a new dataset (HiFi-IFDL) to facilitate the hierarchical fine-grained IFDL study. When evaluating over 8 benchmarks, our method outperforms the state of the art (SoTA) on the tasks of IFDL and achieves the competitive performance on the forgery attribute classifications.

2 Related Work

Image Forgery Detection and Localization In the generic image forgery detection, it is required to distinguish real images from ones generated by a CNN: Zhang et al.Zhang et al. (2019b) report that it is difficult for classifiers to generalize across different GANs and leverage upsampling artifacts as a strong discriminator for GAN detection. On the contrary, against expectation, the work by Wang et al. Wang et al. (2020b) shows that a baseline classifier *can* actually generalize in detecting different GAN models contingent to being trained on synthesized images from ProGAN Karras et al. (2018). Another thread is facial forgery detection Rössler et al. (2019); Dufour et al. (2019); Dolhansky et al. (2019); Li et al. (2020b); Jiang et al. (2020); Sabir et al. (2019); Deb et al.

Table 1 Comparison to previous works

Method	Detection	Localization	Forgery type	Attribute learning	Fixed CLIP visual embed	Text-pixel matching
Wu et al. (2019)	✘	✔	Editing	✘	N/A	N/A
Hu et al. (2020)	✘	✔	Editing	✘	N/A	N/A
Liu et al. (2022)	✔	✔	Editing	✘	N/A	N/A
Dong et al. (2022)	✔	✔	Editing	✘	N/A	N/A
Wang et al. (2022)	✔	✔	Editing	✘	N/A	N/A
Zhang et al. (2019b)	✔	✘	CNN-based	✘	N/A	N/A
Wang et al. (2020b)	✔	✘	CNN-based	✘	N/A	N/A
Asnani et al. (2021)	✔	✘	CNN-based	Syn.-based	N/A	N/A
Yu et al. (2019)	✔	✘	CNN-based	Syn.-based	N/A	N/A
Stehouwer et al. (2020)	✔	✔	CNN-based	✘	N/A	N/A
Huang et al. (2022)	✔	✔	CNN-based	✘	N/A	N/A
Guo et al. (2023b)	✔	✔	Both	✔	N/A	N/A
Wu et al. (2023)	✔	✘	CNN-based	✘	✘	✘
Sun et al. (2023)	✔	✘	CNN-based	✘	✘	✘
Ojha et al. (2023)	✔	✘	CNN-based	✘	✔	✘
Sha et al. (2023)	✔	✘	CNN-based	✘	✔	✘
Rao et al. (2022)	N/A	N/A	N/A	N/A	✘	✔
Zhou et al. (2022a)	N/A	N/A	N/A	N/A	✔	✔
Xu et al. (2022)	N/A	N/A	N/A	N/A	✘	✔
Ghiasi et al. (2022)	N/A	N/A	N/A	N/A	✘	✔
HiFi-Net++	✔	✔	Both	✔	✔	✔

(2023); Bui et al. (2022); Guo et al. (2023a); Yao et al. (2024). For example, DD-VQA Zhang et al. (2024) introduces a novel approach by generating interpretative explanations for deepfake detection, which significantly enhances the interpretability of deepfake detection models by not only identifying the forgeries but also providing clear, human-understandable explanations behind the model's decision. All these works specialize in the image-level forgery detection, which however does not meet the need of knowing where the forgery occurs on the pixel level. Therefore, we perform both image forgery detection and localization, as reported in Table 1.

As for the forgery localization, most of existing methods perform pixel-wise classification to identify forged regions Wu et al. (2019); Hu et al. (2020); Wang et al. (2022) while early ones use a region Zhou et al. (2018) or patch-based Mayer and Stamm (2018) approach. The idea of localizing forgery is also adopted in the DeepFake Detection community by segmenting the artifacts in facial images Zhao et al. (2021); Chai et al. (2020); Cozzolino et al. (2018). Zhou et al. Zhou et al. (2020) improve the localization by focusing on object boundary artifacts. The MVSS-Net Chen et al. (2021); Dong et al. (2022) uses multi-level supervision to balance between sensitivity and specificity.

Attribute Learning CNN-synthesized image attributes can be observed in the frequency domain Zhang et al. (2019b); Wang et al. (2020b), where different GAN generation methods have distinct high-frequency patterns. The task of "GAN discovery and attribution" attempts to identify the exact generative model Marra et al. (2018); Yu et al. (2019); Marra et al. (2019) while "model parsing" identifies both the model and the objective function Asnani et al. (2021). These works differ from ours in two aspects. Firstly, the prior work concentrates on the attribute used in the digital synthesis method (synthesis-based), yet our work studies forgery-based attribute, i.e., to classify GAN-based fully-synthesized or partial manipulation from the image editing process. Secondly, unlike the prior work that assumes a "flat" structure between different attributes, we represent all forgery attributes in a hierarchical way, exploring dependencies among them.

CLIP in Image Understanding In the last few years, CLIP Radford et al. (2021) is employed by the computer vision community for core vision tasks, such as image classification, segmentation Zhou et al. (2022a); Xu et al. (2022, 2023); Ghiasi et al. (2022); Li et al. (2022a) and object detection Zhong et al. (2022); Rao et al. (2022). In this work, we are interested in two topics: (1) CLIP-based image forensic methods and (2) CLIP-based pixel-matching work.

For CLIP-based image forensic methods, Uni-Det Ojha et al. (2023) applies k-NN and linear probing on the fixed pre-trained CLIP visual embedding. This approach demonstrates

remarkable image-level detection performance that generalizes well to many unseen forgery types. DE-FAKE Sha et al. (2023) leverages pre-trained CLIP visual and textual embeddings to effectively identify images generated from text-to-image diffusion models. However, these prior methods have two major differences from our proposed HiFi-Net++. First, prior works apply simple architectures (e.g., ResNet18, 2 MLP layers, etc.) on pre-trained CLIP visual embeddings. In contrast, we meticulously devise a multi-branch feature extractor that comprehensively learns forgery attributes across multiple scales and hierarchical relations among these forgery attributes, empirically resulting in better image-level forgery detection performance. Secondly, prior works can only perform the image-level forgery detection, but our method not only detects image-level forgeries but also localizes manipulated pixels, which is a pixel-level classification task.

On the other hand, the pre-trained CLIP can also be effective on the pixel-wise vision tasks: Zhou et al. (2022a) offers a study that finetunes CLIP for semantic segmentation. Ghiasi et al. (2022) points out that CLIP and ALIGN Jia et al. (2021) models can also roughly learn to group objects and are not able to perform localization. Li et al. (2022a) introduces LSeg, a language-driven semantic image segmentation method that replicates CLIP contrastive training at the pixel level and spatial regularization blocks. MaskCLIP+ Dong et al. (2023) has been recently outperformed by CLIP-S^4 He et al. (2023), not requiring unknown classes to be specified in the training phase. Other very recent works are Xu et al. (2023), performing open-vocabulary, panoptic segmentation and DenseCLIP Rao et al. (2022) that uses the CLIP to serve as the auxiliary input signal to enhance dense prediction. Finally, Xu et al. (2022) proposes a hierarchical Grouping Vision Transformer for image semantic segmentation supervised only by natural language. Please note that, unlike MaskCLIP+, which uses pre-trained CLIP text embedding to segment common objects (e.g., dog, cyclist, and tree), we are working on more challenging tasks that segment manipulated pixels by a forgery method. Therefore, we use multi-head attention to refine the pre-trained text embedding for our image forensic tasks.

3 HiFi-Net++

In this section, we introduce HiFi-Net++ (as shown in Fig. 4) and highlight the modifications made to the original model presented in Guo et al. (2023b). We start by defining the image forgery detection and localization (IFDL) task and hierarchical fine-grained formulation. In IFDL, an image $\mathbf{X} \in \mathbb{R}^{W \times H \times 3}_{|[0,255]}$ needs to be mapped to a binary variable $\mathbf{y}$ to predict if it has been forged or is a pristine image. The method also can perform localization of the manipulation at the pixel level, thereby performing binary segmentation and outputting a binary mask $\mathbf{M} \in \mathbb{R}^{W \times H}_{|[0,1]}$, where the $\mathbf{M}_{ij}$ indicates if the ij-th pixel has been manipulated or not.

In the hierarchical fine-grained formulation, we train the given IFDL algorithm towards fine-grained classifications, and in the inference we evaluate the binary classification results on the image-level forgery detection. Specifically, we denote a categorical variable $\hat{\mathbf{y}}_b$ at branch b, where its value depends on which level we conduct the fine-grained forgery attribute classification. For example, as depicted in Fig. 2b, two forgery attribute categories at level 1 are full-synthesized, partial-manipulated; four categories at level 2 are diffusion model, GAN-based method, image editing, CNN-based partial-manipulated method; categories at level 3 discriminate whether forgery methods are conditional or unconditional; 14 classes at level 4 are real and 13 specific forgery methods. We detail this in Sect. 4 and Fig. 8.

In our preliminary work, the proposed HiFi-Net consists of a multi-branch feature extractor (Sect. 3.1) that performs fine-grained classifications at different specific forgery attribute levels and two modules (Sect. 3.2 and Sect. 3.4) that help the forgery detection and localization, respectively. In the HiFi-Net++, we introduce an additional Language-guided Forgery Localization Enhancer (IFLE) on the top of the HiFi-Net. IFLE leverages the pre-trained CLIP Radford et al. (2021) image encoder $\boldsymbol{\theta}_i$ and text encoder $\boldsymbol{\theta}_t$ to improve manipulation localization performance and the model's generalization ability. The details are reported in Sect. 3.3.

3.1 Multi-branch Feature Extractor

We first extract the feature of the given input image via the color and frequency blocks, and this frequency block applies a Laplacian of Gaussian (LoG) Burt and Adelson (1987) onto the CNN feature map. This architecture design is similar to the method in Masi et al. (2020), which exploits image generation artifacts that can exist in both RGB and frequency domain Wang et al. (2022); Dong et al. (2022); Wang et al. (2020b); Zhang et al. (2019b).

Then, we propose a multi-branch feature extractor, whose branch is denoted as $\boldsymbol{\theta}_b$ with $b \in \{1 \ldots 4\}$. Specifically, each $\boldsymbol{\theta}_b$ generates the feature map of a specific resolution, and such a feature map helps $\boldsymbol{\theta}_b$ conduct the fine-grained classification at the corresponding level. For example, for the finest level (i.e. identifying the individual forgery methods), one needs to model contents at all spatial locations, which requires a high-resolution feature map. In contrast, it is reasonable to have low-resolution feature maps for the coarsest level (i.e. binary) classification.

We observe that different forgery methods generate manipulated areas with different distributions (Fig. 1b), and different patterns, e.g., deepfake methods Rössler et al. (2019); Li et al. (2020a) manipulate the whole inner part

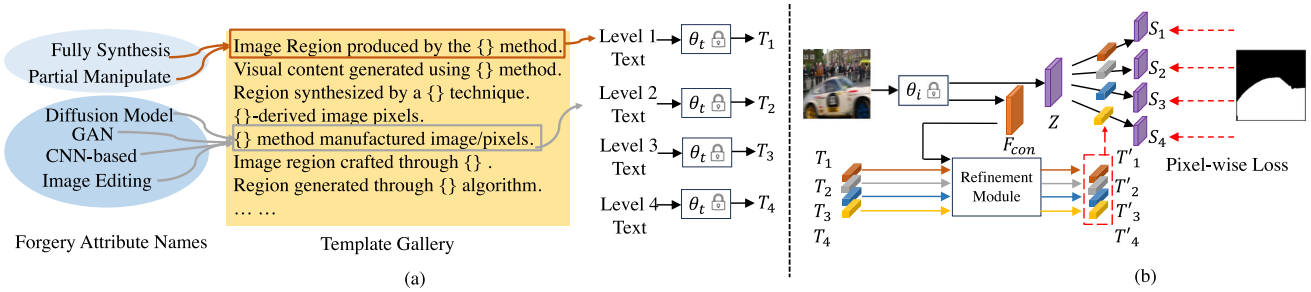


Fig. 5 **a** Two forgery attribute names on level 1 of the hierarchical fine-grained formulation (Fig. 8) are `Fully Synthesis` and `Partial Manipulation`. We combine these forgery attribute names with a template (e.g., `"Image region is synthesized by {} method"`), which is randomly chosen from the template gallery. Therefore, the level 1 text input has two sentences: `"image is manipulated by fully-synthesized method"`, `"image is manipulated by partial-manipulated methods"`. Consequently, the level 2, level 3 and level 4 have 4, 6, and 13 sentences as the text input. We apply the pre-trained CLIP text encoder (e.g., $\boldsymbol{\theta}_t$) on text inputs at different levels and obtain text embedding $\mathbf{T}_1$, $\mathbf{T}_2$, $\mathbf{T}_3$ and $\mathbf{T}_4$. **b** The pre-trained CLIP image encoder (e.g., $\boldsymbol{\theta}_i$) takes the input image, we obtain $\mathbf{F}_{con}$ to represent the visual content, and feature map $\mathbf{Z}$ that maintains the ability to be aligned with text embedding. After that, a refinement module utilizes $\mathbf{F}_{con}$ to conduct refinements on text embeddings at different levels. These refined text embedding ($\mathbf{T}'_b$ with $b \in \{1 \dots 4\}$) along with $\mathbf{Z}$ generates manipulation score map $\mathbf{S}_b$ with $b \in \{1 \dots 4\}$, as the auxiliary signal to help the localization. During the training, we only keep the refinement module as trainable, while pre-trained CLIP image and text encoders are frozen

of the face, whereas STGAN Liu et al. (2019) changes sparse facial attributes such as mouth and eyes. Therefore, we believe that features used for fine-grained classification can serve as a prior for localization. It is important to have such a design for localizing both manipulated images with CNNs or classic image editing.

3.2 Classification Module

Hierarchical Path Prediction We intend to learn the hierarchical dependency between different forgery attributes. Given the image $\mathbf{X}$, we denote output logits and predicted probability of the branch $\boldsymbol{\theta}_b$ as $\boldsymbol{\theta}_b(\mathbf{X})$ and $p(\mathbf{y}_b|\mathbf{X})$, respectively. Then, we have:

$$p(\mathbf{y}_b|\mathbf{X}) \doteq \text{softmax}\left(\boldsymbol{\theta}_b(\mathbf{X}) \odot (1 + p(\mathbf{y}_{b-1}|\mathbf{X}))\right) \tag{1}$$

Before computing the probability $p(\mathbf{y}_b|\mathbf{X})$ at branch $\boldsymbol{\theta}_b$, we scale logits $\boldsymbol{\theta}_b(\mathbf{X})$ based on the previous branch probability $p(\mathbf{y}_{b-1}|\mathbf{X})$. Then, we enforce the algorithm to learn hierarchical dependency. Specifically, in Eq. (1), we repeat the probability of the coarse level $b-1$ for all the logits output by the branch at level b, following the hierarchical structure. Figure 6 shows that the logits associated with predicting `DDPM` or `DDIM` are multiplied by probability for the image to be `Unconditional (Diffusion)` in the last level, according to the tree structure in the hierarchical fine-grained formulation.

3.3 Language-Guided Forgery Localization Enhancer

The overall Language-guided Forgery Localization Enhancer module contains the pre-trained CLIP image $\boldsymbol{\theta}_i$ and text encoders $\boldsymbol{\theta}_t$, as well as a refinement block that makes the generated text embeddings better adapted to the manipulation localization task.

Text Input Construction As depicted in Fig. 5a, we first use the name of forgery attributes and pre-defined templates to create text inputs. To improve the variability of the text input, we form a template gallery with many text templates generated by a publicly available Large Language Model (LLM) Ouyang et al. (2022). Also, we empirically use multiple alternative names for each forgery attribute, for example, `"GAN"`, `"Generative Adversarial Networks"` are the alternative names of the same manipulation method (e.g., forgery attribute). After that, the pre-trained text encoder $\boldsymbol{\theta}_t$ takes text inputs to produce the text embeddings.

Formally, we denote the text embedding at different forgery levels as: $\mathbf{T}_1 \in \mathbb{R}^{2\times C}$, $\mathbf{T}_2 \in \mathbb{R}^{4\times C}$, $\mathbf{T}_3 \in \mathbb{R}^{6\times C}$ and $\mathbf{T}_4 \in \mathbb{R}^{13\times C}$, where $\mathbf{T}_b$ with $b \in \{1 \dots 4\}$ indicates the text embedding of the prompt generated at level b and C is the dimension of the text embedding.

Architecture When the pre-trained CLIP image encoder $\boldsymbol{\theta}_i$ takes the given image as input, we obtain b intermediate features maps, denoted as $\mathbf{V}_b \in \mathbb{R}^{W_b \times H_b \times C}$ with $b \in \{1 \dots 4\}$. These feature maps are visual embeddings that are depicted in Fig. 4, and we empirically find them have a high generalization ability in distinguishing real and manipulated pixels. More details can be found in Table 2 and Fig. 11. After that, we concatenate these feature maps into $\mathbf{F}_{con}$. Given the powerful representation ability of the pre-trained CLIP model, $\mathbf{F}_{con}$ can contain rich information for visual semantics. Secondly, we extract the feature map before the final attention pooling in $\boldsymbol{\theta}_i$, and denote this feature map as $\mathbf{Z} \in \mathbb{R}^{W\times H\times C}$. In general, in the pre-trained CLIP image encoder, the feature

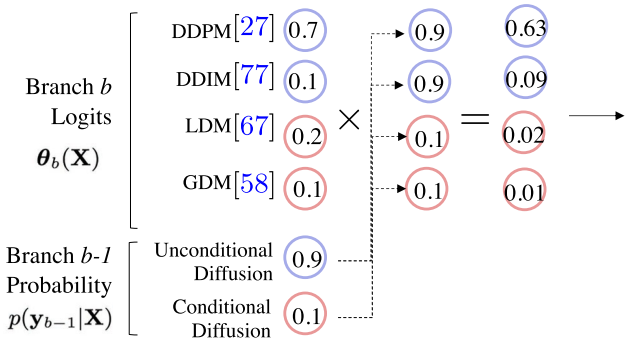


Fig. 6 The classification probability output from branch $\boldsymbol{\theta}_b$ depends on the predicted probability at branch $\boldsymbol{\theta}_{b-1}$, following the definition of the hierarchical forgery attributes tree

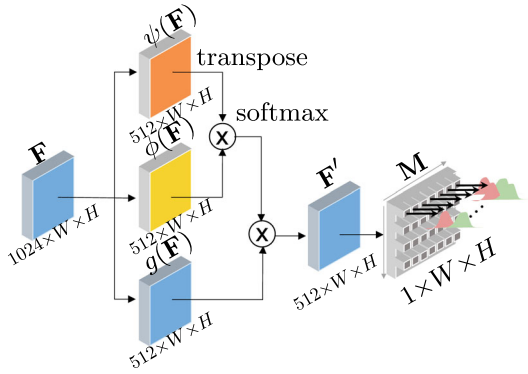


Fig. 7 The localization module adopts the self-attention mechanism to transfer the feature map $\mathbf{F}$ to the localization mask $\mathbf{M}$

after the final attention pooling is used to correlate with language. However, as indicated in DenseCLIP work Rao et al. (2022), even $\mathbf{Z}$ also has the ability to correlate or align with the given text input.

Similar to the previous method that refines the fixed pre-trained CLIP text embedding for specific downstream tasks, we apply a refinement module, which is based on the multi-head attention mechanism. Formally, such a refinement module is applied on the $\mathbf{F}_{con}$ and $\mathbf{T}_b$ with $b \in \{1 \ldots 4\}$, and then produces a text embedding perturbation $\Delta\mathbf{T}_b$ with $b \in \{1 \ldots 4\}$

$$\Delta\mathbf{T}_b = \texttt{RefineModule}(\mathbf{T}_b, \mathbf{F}_{con}). \tag{2}$$

This text embedding perturbation is finally added on the original input text embedding $\mathbf{T}_b$ to produce the final text embedding for detection and localization performance as:

$$\mathbf{T}'_b = \mathbf{T}_b + \Delta\mathbf{T}_b. \tag{3}$$

We then compute the cosine similarity between each refined text embedding and $\mathbf{Z}$ to yield a manipulation score map as:

$$\mathbf{s}_b \doteq \mathbf{T}'_b\,\mathbf{Z}^\top \quad b \in \{1 \ldots 4\}. \tag{4}$$

As a result, this operation yields four manipulation score maps of the spatial dimensionality: $\mathbf{s}_1 \in \mathbb{R}^{2\times W\times H}$, $\mathbf{s}_2 \in \mathbb{R}^{4\times W\times H}$, $\mathbf{s}_3 \in \mathbb{R}^{6\times W\times H}$, $\mathbf{s}_4 \in \mathbb{R}^{13\times W\times H}$. These score maps have two properties: (a) they can serve as a language prior, which enables us to leverage the powerful image-text association ability of the pre-trained CLIP model; (b) being supervised against the corresponding ground truth manipulation mask, these manipulation score maps learn to indicate the region that is manipulated by the given forgery method, offering auxiliary signals for manipulation localization.

These generated 2D score maps are incorporated in the existing HiFi-Net and help localize the manipulation area, as depicted in Fig. 4. Specifically, we first concatenate the HiFi-Net output features (e.g., $\boldsymbol{\theta}_b$), up-sampling visual embeddings $\mathbf{V}_{b\uparrow}$—where the arrow indicates upsampling operation—and clip image encoder feature $\mathbf{Z}$ into a final fused feature map $\mathbf{Y}_b$.

$$\mathbf{Y}_b \doteq [\mathbf{Z}_\uparrow, \mathbf{V}_{b\uparrow}, \boldsymbol{\theta}_b] \quad b \in \{1 \ldots 4\}. \tag{5}$$

Discussion It is important to alleviate the potential concern about why the pre-trained CLIP image and text encoders generate manipulation score maps that help the forgery localization tasks, as the pre-trained CLIP model is meant to correlate image and text based on their semantics. In fact, the manipulation score map is generated by $\mathbf{F}_{con}$ and $\mathbf{T}'_b$ via Eq. (4), and this $\mathbf{F}_{con}$ is the concatenated feature of $\mathbf{V}_b$, which are empirically generalize to the manipulation localization task (details in Table 2 of Sect. 5.2). In addition, text embedding itself contains rich semantics and we further use a refinement module to make it better adapt for our manipulation localization task. Note that we define a pixel-wise loss (e.g., cross entropy) between the manipulation score map and ground truth forgery mask. We use this pixel-wise loss to optimize the trainable refinement module.

3.4 Localization Module

Architecture We first feed the final fused $\mathbf{Y}_b$ into the feature pyramid network Lin et al. (2017), which is widely used as the effective algorithm for fusing multi-scale feature maps in the pixel-wise prediction task. After that, we have the final fused feature map, denoted as $\mathbf{F} \in \mathbb{R}^{512\times W\times H}$, which is used to output the mask $\hat{\mathbf{M}}$ for localizing the forgery.

To model the dependency and interactions of pixels on the large spatial area, the localization module employs the self-attention mechanism Zhang et al. (2019a); Wang et al. (2018). As shown in the localization module architecture in Fig. 7, we use 1×1 convolution to form g, ϕ and ψ, which convert input feature $\mathbf{F}$ into $\mathbf{F}_g = g(\mathbf{F})$, $\mathbf{F}_\phi = \phi(\mathbf{F})$ and $\mathbf{F}_\psi = \psi(\mathbf{F})$. Given $\mathbf{F}_\phi$ and $\mathbf{F}_\theta$, we compute the spatial attention matrix

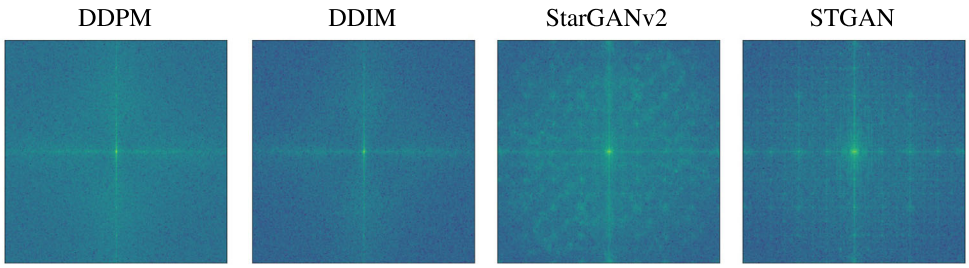
Table 2 IFDL results on HiFi-IFDL

Forgery detection	CNN-syn.		Image edit.		Overall	
	AUC	F1	AUC	F1	AUC	F1
CNN-det.* (Wang et al. 2020b)	76.5	60.5	54.8	33.5	56.5	40.5
CNN-det. (Wang et al. 2020b)	92.3	90.0	87.0	74.7	90.1	83.7
Two-bran. (Masi et al. 2020)	93.3	89.2	83.3	66.7	86.7	80.2
Att. Xce. (Stehouwer et al. 2020)	93.8	91.2	90.8	82.1	87.3	90.0
PSCC-Net (Liu et al. 2022)	94.6	93.2	90.7	82.3	93.2	91.3
Uni-Det (Ojha et al. 2023)	94.0	94.2	63.2	41.5	77.2	60.2
HiFi-Net (Guo et al. 2023b)	97.0	**96.1**	91.5	85.9	96.8	**94.1**
HiFi-Net++	**97.8**	95.0	**91.7**	**86.2**	**97.2**	93.7

(a) CNN-detector Wang et al. (2020b) has 4 variants with different augmentations, and we report the variant with the best performance. For Two-branch Masi et al. (2020), we implement this method with the help of its authors

Forgery localization	CNN-syn.		Image edit.		Overall	
	AUC	F1	AUC	F1	AUC	F1
OSN-det.* (Wu et al. 2022)	51.4	38.8	83.2	70.1	79.4	56.5
CatNet* (Kwon et al. 2022)	48.6	31.9	86.1	79.4	78.3	65.1
CatNet (Kwon et al. 2022)	92.5	81.5	92.0	88.2	92.4	86.8
Att. Xce. (Stehouwer et al. 2020)	89.1	87.7	83.3	79.3	87.1	86.5
PSCC-Net (Liu et al. 2022)	94.3	96.8	91.1	86.5	92.7	94.9
HiFi-Net (Guo et al. (2023b))	98.4	97.0	93.0	90.1	95.3	96.9
CLIP-ResNet50[1] (Radford et al. 2021)	90.2	80.6	84.5	79.3	88.7.3	84.2
CLIP-ViT-B-16[1] (Radford et al. 2021)	87.5	78.3	79.8	73.2	80.2	69.7
HiFi-Net++	**98.6**	**98.1**	**97.8**	**95.7**	**97.7**	**98.6**

(b) OSN-det Wu et al. (2022) only releases pre-trained weights with the inference script, without the training script. CLIP image encoders with [1] represents that we freeze pre-trained CLIP image encoders and only train a few transposed convolution layers



(c) Frequency artifacts in different forgery methods. DDPM Ho et al. (2020) and DDIM Song et al. (2021) do not exhbit the checkboard patterns Zhang et al. (2019b); Wang et al. (2020b) observed in GAN-based methods, such as StarGAN-v2 Choi et al. (2020) and STGAN Liu et al. (2019)

∗ means we apply author-released pre-trained models. Models without ∗ mean they are trained on HiFi-IFDL training set. [**Bold**: best result]

$\mathbf{A}_s = \text{softmax}(\mathbf{F}_\phi^T \mathbf{F}_\theta)$. We then use this transformation $\mathbf{A}_s$ to map $\mathbf{F}_g$ into a global feature map $\mathbf{F}' = \mathbf{A}_s \mathbf{F}_g \in \mathbb{R}^{512 \times W \times H}$.

Objective Function Following Masi et al. (2020), we employ a metric learning objective function for localization, which creates a wider margin between real and manipulated pixels. We first learn features of each pixel and then model the geometry of such learned features with a radial decision boundary in the hyper-sphere. Specifically, we start with pre-computing a reference center $\mathbf{c} \in \mathbb{R}^D$, by averaging the features of all pixels in real images of the training set. We use $\mathbf{F}'_{ij} \in \mathbb{R}^D$ to indicate the ij-th pixel of the final mask prediction layer. Therefore, our localization loss $\mathcal{L}_{loc}$ is:

$$\mathcal{L}_{loc} = \frac{1}{HW} \sum_{i}^{H} \sum_{j}^{W} \mathcal{L}(\mathbf{F}'_{ij}, \mathbf{M}_{ij}; \mathbf{c}, \tau), \tag{6}$$

where:

$$\mathcal{L} = \begin{cases} \left\| \mathbf{F}'_{ij} - \mathbf{c} \right\|_2 & \text{if } \mathbf{M}_{ij} \text{ real} \\ \max\left(0, \tau - \left\| \mathbf{F}'_{ij} - \mathbf{c} \right\|_2\right) & \text{if } \mathbf{M}_{ij} \text{ forged.} \end{cases}$$

Here τ is a pre-defined margin. The first term in $\mathcal{L}$ improves the feature space compactness of real pixels. The second term encourages the distribution of forged pixels to be far away from real by a margin τ. Note our method differs to Ruff et al. (2018); Masi et al. (2020) in two aspects: (1) unlike Ruff et al. (2018), we use the second term in $\mathcal{L}$ to enforce separation; (2) compared to the image-level loss in Masi et al. (2020) that has two margins, we work on the more challenging pixel-level learning. Thus, we use a single margin, which reduces the number of hyper-parameters and improves the simplicity.

3.5 Training and Inference

In the training, each branch is optimized towards the classification at the corresponding level, we use 4 classification losses, $\mathcal{L}^1_{cls}$, $\mathcal{L}^2_{cls}$, $\mathcal{L}^3_{cls}$ and $\mathcal{L}^4_{cls}$ for 4 branches. At the branch b, $\mathcal{L}^b_{cls}$ is the cross entropy distance between $p(\mathbf{y}_b|\mathbf{X})$ and a ground truth categorical $\hat{\mathbf{y}}_b$. When the input image is labeled as "real", we only apply the last branch ($\boldsymbol{\theta}_4$) loss function. Otherwise, we use all the branches. Let us denote the input image as $\mathbf{X}$.

$$\mathcal{L}_{cls} = \begin{cases} \mathcal{L}^1_{cls} + \mathcal{L}^2_{cls} + \mathcal{L}^3_{cls} + \mathcal{L}^4_{cls} & \text{if } \mathbf{X} \text{ is forged} \\ \mathcal{L}^4_{cls} & \text{if } \mathbf{X} \text{ is real.} \end{cases}$$

In addition, we compute the cross entropy between the predicted manipulation score map $\mathbf{s}_b$ and the downsampled manipulation label $\mathbf{M}_{\downarrow}$. We denote this loss as $\mathcal{L}^{score}_{loc}$. The architecture is trained end-to-end with different learning rates per layer. The detailed objective function is:

$$\mathcal{L}_{tot} = \lambda_1 \mathcal{L}_{loc} + \lambda_2 \mathcal{L}_{cls} + \lambda_3 \mathcal{L}^{score}_{loc}. \tag{7}$$

λ_1, λ_2, and λ_3 are hyper-parameters that keep different objective terms on the reasonable magnitudes.

In the inference, HiFi-Net++ takes the input image $\mathbf{X}$ and pre-defined text input $\mathbf{T}_b$, and then generate the forgery mask from the localization module, and predicts forgery attributes at different levels. We use the output probabilities at level 4 for forgery attribute classification. For binary "forged vs. real" classification, we predict as forged if the highest probability falls in any manipulation method at level 4.

4 Hierarchical Fine-Grained IFDL Dataset

We construct a fine-grained hierarchical benchmark, named HiFi-IFDL, to facilitate our study. HiFi-IFDL contains some most updated and representative forgery methods, for two reasons: (1) Image synthesis evolved into a more advanced era and artifacts become less prominent in the recent forgery method; (2) It is impossible to include all possible generative method categories, such as VAE Kingma and Welling (2014) and face morphing Scherhag et al. (2019). So we only collect the most-studied forgery types (*i.e.,* splicing) and the recent generative methods (*i.e.,* DDPM).

Specifically, HiFi-IFDL includes images generated from 13 forgery methods spanning from CNN-based manipulations to image editing, and these forged methods are based on taxonomy illustrated in Fig. 8. More formally, we use Table 9 to report more details of these forgery methods, each of which generates 100, 000 images. For the real images, we select them from 6 datasets (e.g., FFHQ Karras et al. (2019), AFHQ Choi et al. (2020), CelebaHQ Lee et al. (2020), Youtube face Rössler et al. (2019), MSCOCO Lin et al. (2014), and LSUN Yu et al. (2015)). For datasets that contain more than 100, 000 images or video frames (e.g., LSUN, and YouTube videos), we randomly select 100, 000 images from them. For datasets with a total number of images less than 100, 000 (e.g., FFHQ Karras et al. (2019), AFHQ Choi et al. (2020), and CelebaHQ Lee et al. (2020)), we utilize the entire dataset to construct the HiFi-IFDL. As a result, training, validation, and test sets have 1, 710K, 15K, and 174K images, respectively. Constructing such a large-scale dataset is essential, as it offers a wide variety of forgery patterns and realistic distributions of real images, which help train a robust detection model that performs effectively in real-world scenarios. In Fig. 9b, we show several examples taken from our dataset that represent a variety of objects, faces, and animals. All manipulations are marked by the high-resolution binary mask.

While there are different ways to design a forgery hierarchy, our hierarchy starts at the root of an image being forged, and then each level is made more and more specific to arrive at the actual generator. Our work studies *the impact of the hierarchical formulation to IFDL*. While different hierarchy definitions are possible, it is beyond the scope of this paper.

5 Experiments

In this section, we first evaluate IFDL performance of the proposed HiFi-Net++ on HiFi-IFDL dataset and report the results in Sect. 5.2. Then, in Sect. 5.3, we examine the IFDL performance on the image editing domain, following the experiment setup defined in Wang et al. (2022), and compare to various prior works Wu et al. (2019); Hu et al. (2020); Liu et al. (2022); Dong et al. (2022); Wang et al. (2022); Guo et al. (2023b) on 5 different datasets, including *Columbia* Ng et al. (2009), *Coverage* Wen et al. (2016), *CASIA* Dong et al. (2013) and *NIST16* NIS (2016). After that, in Sect. 5.4, we conduct a comparison to SoTA forgery detection methods on identifying images generated by various forgery methods. In the end, for the digital facial manipulated image, we conduct the evaluation on the Diverse Fake Face Dataset

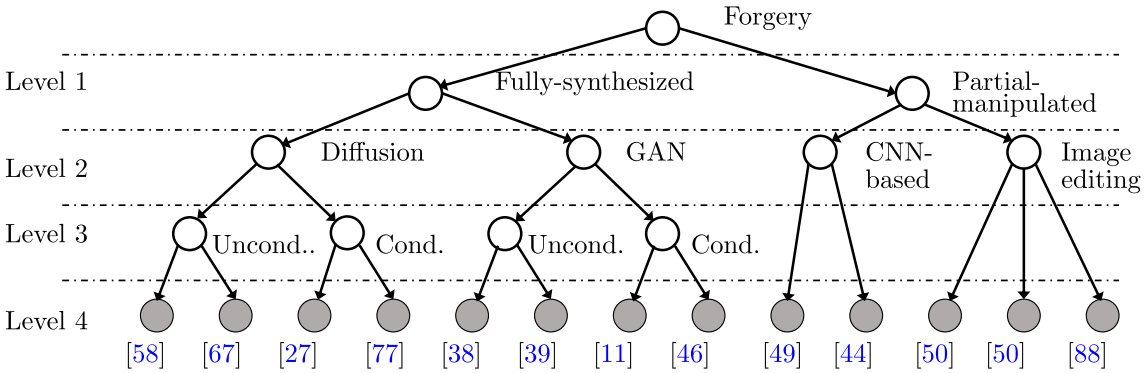
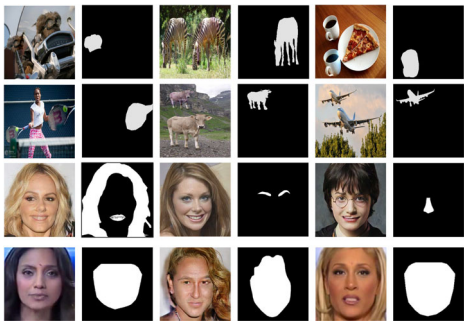


Fig. 8 Overview of HiFi-IFDL dataset. At level 1, we separate forged images into fully-synthesized and partial-manipulated. At level 2, we discriminate different forgery methodologies, e.g., image editing, CNN-based partial manipulation, Diffusion or GANs. Then, at level 3, we separate images based on whether forgery methods are conditional or unconditional. The final level 4 refers to the specific forgery method. [Key: Uncond.: unconditional, Cond.: conditional]

Forgery Method	Image Source	Images #	Source
DDPM [27]	LSUN	100k	pre-trained
DDIM [77]	LSUN	100k	pre-trained
GDM. [58]	LSUN	100k	pre-trained
LDM. [67]	LSUN	100k	pre-trained
StarGANv2 [11]	CelebaHQ	100k	pre-trained
HiSD [46]	CelebaHQ	100k	pre-trained
StGANv2-ada [38]	FFHQ, AFHQ	100k	pre-trained
StGAN3 [39]	FFHQ, AFHQ	100k	pre-trained
STGAN [49]	CelebaHQ	100k	self-train
Faceshifter [44]	Youtube video	100k	released

(a)



(b)

Fig. 9 **a** The details of the collected dataset. Each column in order shows forgery method; the image source used for the generation; the image number; if images are generated with pre-trained/self-trained models or released images. **b** We offer high resolution forgery masks on manipulated images

(DFFD) dataset Stehouwer et al. (2020), comparing it with prior works Stehouwer et al. (2020). The reason we choose DFFD is because DFFD offers the manipulation mask label on the fake facial images, which enables us to measure both detection and localization performance.

5.1 Implementation Details

HiFi-Net++ is implemented on PyTorch and, during the training procedure, trained with 40 epochs and each epoch contains 100, 000 iterations. The training batch size is 16, with 8 real and 8 forged images. In our HiFi-Net++, multi-branch feature extractor, the feature map resolutions for different branches are 256, 128, 64, and 32 pixels. In the experiment on the HiFi-IFDL dataset, the fine-grained classification for 1st, 2nd, 3rd, and 4th levels are 2-way, 4-way, 6-way, and 14-way multi-class classification, respectively. In terms of the Language-guided Forgery Localization Enhancer, we have experimented with different image encoder backbones, such as ResNet-50, ResNet-101 and ViT-B-16. The overall performance from these backbones does not vary largely, and ResNet50 offers the best computation efficiency. Therefore, we select ResNet-50 as the pre-trained CLIP image encoder for the main result analyzed in the rest of the paper, except the ablation study.

As for the details of $\mathcal{L}_{loc}$ implementation, we first use the Feature Pyramid Network to convert the joint feature $\mathbf{Y}_b$ to the high-dimensional feature $\mathbf{F}'_{ij} \in R^D$, where $D = 64$. Then we average the feature $\mathbf{F}'_{ij} \in R^D$ for all pixels in the real image from the HiFi-IFDL, and this average value then is used as $\mathbf{c}$. Then, we compute the ℓ_2 distance between each pixel feature $\mathbf{F}'_{ij} \in R^D$ and $\mathbf{c}$, and denote the largest distance as D_{max}. In the Eq. 6 of the paper, we set the threshold τ as $2.5 \cdot D_{max}$.

Both HiFi-Net Guo et al. (2023b) and HiFi-Net++ are trained in an end-to-end manner. The detailed hyper-parameters in Eq. 7 are $\lambda_1 = 10$, $\lambda_2 = 1$, and $\lambda_3 = 0.01$. The feature extractor is modified based on the pre-trained HRNet Wang et al. (2020a), in which we add color and frequency-enhanced blocks, as well as more convolu-

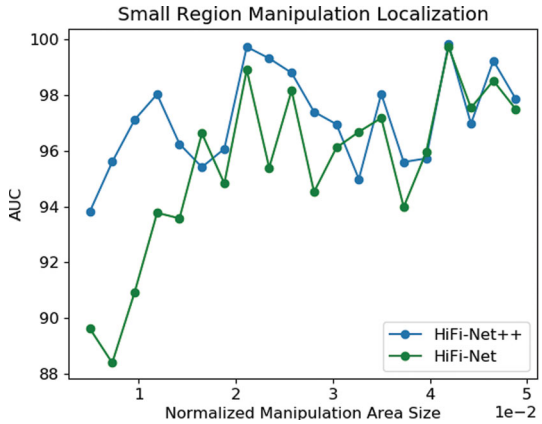


Fig. 10 The small region manipulation localization performance of HiFi-Net Guo et al. (2023b) and HiFi-Net++. The X-axis denotes the normalized manipulation area size, indicating this plot only visualizes forged images with manipulation less than 5% of the input image size. The Y-axis depicts the average AUC of two methods achieve among images with the given manipulation area size. With the help of the language-guided forgery localization enhancer, the HiFi-Net++ improves the localization performance over HiFi-Net, on images with small manipulation regions

tion layers. Consequently, each branch of our multi-branch feature extractor can have an identical number of convolutional layers, and such details can be found in our source code, which is publicly released. Moreover, we employ different learning rates for different trainable modules. For example, we use `1e-4` for the multi-branch feature extractor and classification module, `3e-4` for the localization module, and `5e-5` for the refinement module inside the language-guided forgery localization enhancer. Note that we do not change the learning strategy when different pre-trained CLIP image encoders are used in the language-guided forgery localization enhancer. During the training, we use `ReduceLROnPlateau` as the learning rate scheduler to reduce the learning rate when the value of the composite loss function on the validation set does not decrease for continuous 75 times.

5.2 HiFi-IFDL Dataset Performance

Table 2 reports the different model performance on the HiFi-IFDL dataset, in which we use AUC and F1 score as metrics on both image-level forgery detection and pixel-level localization. Specifically, in Table 2, first, we observe that the pre-trained CNN-detector Wang et al. (2020b) does not perform well because it is trained on GAN-generated images that are different from images manipulated by diffusion models. Such differences can be seen in Fig. 2c, where we visualize the frequency domain artifacts by following the routine Wang et al. (2020b) that applies the high-pass filter on the image generated by different forgery methods. Similar visualization is adopted in Wang et al. (2020b); Zhang et al. (2019b); Corvi et al. (2022); Ricker et al. (2022) also. Then, we train both prior methods on HiFi-IFDL, and they again perform worse than our model: CNN-detector uses ordinary ResNet50, but our model is specifically designed for image forensics. Two-branch processes deepfakes video by LSTM that is less effective in detecting forgery in the image editing domain. Attention Xception Stehouwer et al. (2020) and PSCC-Net are proposed for facial image forgery and image editing domains, respectively. These two methods perform worse than HiFi-Net++ by 9.9% and 4.0% AUC, respectively. In the last two rows, we can observe that HiFi-Net++ achieves 0.4% higher AUC and 0.4% lower F1 scores than the HiFi-Net. It is reasonable for HiFi-Net++ to have comparable image forgery detection performances with HiFi-Net, since the newly-proposed language-guided forgery localization enhancer mainly improves the localization performance, and features used for forgery image detection are the same in the HiFi-Net and HiFi-Net++.

In Table 2, we compare with previous methods which can perform the forgery localization. Specifically, the pre-trained OSN-detector Wu et al. (2022) and CatNet Kwon et al. (2022) do not work well on CNN-synthesized images in HiFi-IFDL dataset, since they merely train models on images manipulated by editing methods. Then, we use HiFi-IFDL dataset to train CatNet, but it still performs worse than ours: CatNet uses DCT stream to help localize areas of splicing and copy-move, but HiFi-IFDL contains more forgery types (e.g., inpainting). Meanwhile, the accurate classification performance further helps the localization as statistics and patterns of forgery regions are related to different individual forgery methods. For example, for the forgery localization, the HiFi-Net achieves 2.6% AUC and 2.0% F1 improvement over PSCC-Net. Additionally, the superior localization demonstrates that our hierarchical fine-grained formulation learns more comprehensive forgery localization features than the multi-level localization scheme proposed in PSCC-Net.

Before discussing the HiFi-Net++, we first report how the pre-trained CLIP encoder generalizes to the manipulation localization task. First, we obtain the intermediate features $\mathbf{V}_b \in \mathbb{R}^{C \times W_b \times H_b}$ with $b \in \{1 \ldots 4\}$ from the pre-trained CLIP image encoder (Sect. 3.3 and Fig. 4). We apply a 1×1 convolution (denoted as $\mathbf{w}_{1\times1}$) to convert the channel-wise dimension C of each intermediate feature into 64, and upsample $\mathbf{V}_4$, $\mathbf{V}_3$, and $\mathbf{V}_2$, such that they have the same width and height as $\mathbf{V}_1$. Then, we concatenate these transformed feature maps into a joint visual embedding, on which we apply a few transposed convolution layers for localizing the manipulation. During the training, we only optimize a few transposed convolution layers and $\mathbf{w}_{1\times1}$, while the pre-trained CLIP image encoder remains fixed. Consequently, as shown in the Table 2, CLIP-ResNet50 achieves similar localization results as Attention Xception Stehouwer et al. (2020), which is exclusively trained on the HiFi-IFDL dataset. Fur-

thermore, we show the qualitative results of its localization performance in Fig. 11, in which pre-trained CLIP image encoder can capture the rough shape and only fails to capture the specific contour of objects, (e.g., person). This demonstrates that the pre-trained CLIP image encoder has the generalization ability to help the manipulation localization. Although CLIP-ViT-B-16 does not offer as impressive results as CLIP-ResNet-50, it still yields comparable results with the pre-trained CatNet which specializes in localizing manipulations.

Therefore, with the additional language-guided forgery localization module, which contains the pre-trained CLIP image encoder, HiFi-Net++ achieves the better overall localization performance over the HiFi-Net, and this improvement mainly comes from the image editing domain where HiFi-Net++ achieves 4.8% higher AUC and 5.6% F1 score over HiFi-Net. We believe this is because, in the language-guided forgery localization module, both the pre-trained CLIP visual embedding and manipulation score maps play key roles in this improvement. More details can be found in the ablation study (Sect. 5.6). Also, even though HiFi-Net++ does not largely surpass over the HiFi-Net on the localization performance of the CNN-synthesized image, it again achieves the better performance in both AUC and F1 scores. Moreover, HiFi-Net++ delivers the better performance on localizing the small region manipulations, as illustrated in Fig. 10. Qualitatively, we showcase the manipulation localization results of HiFi-Net++ in Fig. 11 and Fig. 12. Based on these empirical

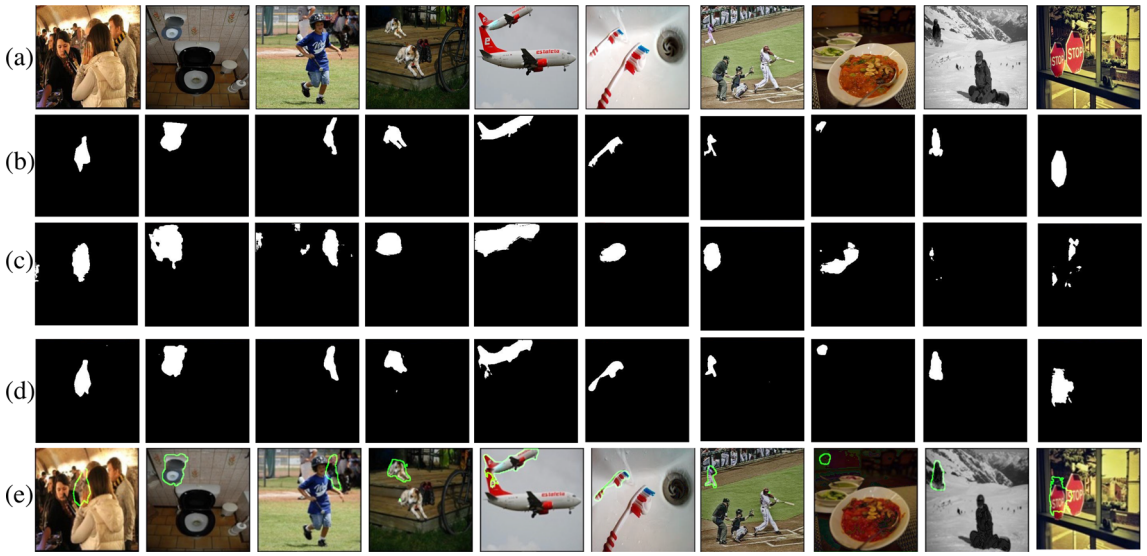


Fig. 11 The qualitative localization performance on the image editing domain data in HiFi-IFDL *test* dataset. From top to down, rows are **a** input image, **b** ground truth mask, **c** pre-trained CLIP image encoder (ResNet-50) localization result, **d** HiFi-Net++ localization result, and **e** Overlaid image based on the HiFi-Net++ localization result. From row **c**, the pre-trained CLIP image encoder can roughly locate the manipulation objects, whereas it still sometimes identifies the manipulation in the wrong way (last three images in this row). In contrast, in row **d** and **e**, we show the HiFi-Net++ can produce localization mask that is close to the ground truth and identifies the manipulation region accurately

Fig. 12 The qualitative manipulation localization performance on the facial forgery images in HiFi-IFDL *test* dataset. From top to bottom, figures are **a** input image, **b** ground truth localization mask, and **c** localization result of HiFi-Net++. As depicted in the last row, HiFi-Net++ is able to identify manipulation on different facial components, such as the face and hair of different styles. Also, the small manipulations on eyebrows (the third last column) and eyes (last two columns) can be identified

results, we can conclude that, by introducing the additional language-guided forgery localization enhancer, the localization performance can be enhanced.

5.3 Image Editing Datasets Performance

Table 3 reports IFDL results for the image editing domain. We evaluate on 5 datasets: *Columbia* Ng et al. (2009), *Coverage* Wen et al. (2016), *CASIA* Dong et al. (2013), *NIST16* NIS (2016) and *IMD*20 Novozamsky et al. (2020). Following the previous experimental setup of Wu et al. (2019); Hu et al. (2020); Liu et al. (2022); Dong et al. (2022); Wang et al. (2022), we pre-trained both HiFi-Net Guo et al. (2023b) and HiFi-Net++ on the HiFi-IFDL and then evaluate these two models on unseen test sets. After that, we continue to fine-tune the pre-trained models on the *NIST16*, *Coverage*, and *CASIA*. Table 3 reports different pre-trained models' localization performances. Specifically, NCL has a 1.3 higher AUC score than HiFi-Net Guo et al. (2023b) on the average localization performance, showcasing the effectiveness of its contrastive learning scheme that captures the non-mutual exclusive relation among different image patches, TANet utilizes a stacked multi-scale transformer with boundary operators that help achieve 4.0 higher AUC score than HiFi-Net Guo et al. (2023b) on CASIA dataset. Nevertheless, HiFi-Net++ maintains the best localization performance in this setup. For example, measured by the average performance across all different test sets, HiFi-Net++ has 0.5% and 0.8% higher AUC scores than TANet and NCL, respectively. This advantage largely comes from HiFi-Net++'s better performance on the Coverage Wen et al. (2016) dataset, which contains many samples with only a small manipulation region. HiFi-Net++ is effective in identifying such manipulation areas with small spatial sizes. On the other hand, Table 3 shows our HiFi-Net++ maintains the best localization performance when the model is fine-tuned on specific manipulation localization datasets, measured by either AUC or F1 score. We hypothesize this is because the LGLE not only improves the HiFi-Net++'s localization performance to unseen manipulation patterns but also encourages HiFi-Net++ to capture specific manipulation patterns after fine-tuning.

Image-level Forgery Detection We also report the image-level forgery detection results in Table 4, in which HiFi-Net++ achieves the best performance, while HiFi-Net achieves comparable results to ObjectFormer Wang et al. (2022). This shows the language-guided forgery localization enhancer also yields an effective representation that can potentially help detect the image manipulated by the conventional editing method. We show qualitative results in Fig. 13, where the manipulated region identified by HiFi-Net++ can capture semantically meaningful object shapes (e.g., person and airplane) and offer much more accurate localization performance than HiFi-Net and PSCC-Net.

Robustness Analysis Following previous works Wang et al. (2022); Liu et al. (2022); Hu et al. (2020), we evaluate the robustness of both HiFi-Net and HiFi-Net++ against different post-processing steps. The corresponding result is reported in Table 4. First, the proposed HiFi-Net is more robust than the previous work, except for the post-processing of resizing 0.78 times the image and JPEG compression with 50% quality. On the other hand, the robustness further improves in the HiFi-Net++, with an increase of 0.9% in the average AUC over HiFi-Net. This improvement can be explained by the fact that HiFi-Net++ uses the visual embedding of the pre-trained CLIP image encoder. This visual embedding is exposed to data collected online, which contains a large number of noisy and blurred images.

Zero-shot Localization Performance on GenAI Inpainting Methods To verify the effectiveness of the proposed localization methods on the image manipulated by recent GenAI methods, we collect images manipulated by Stable Diffusion Models and Repaint Lugmayr et al. (2022), which are held out from training, thereby testing HiFi-Net++ zero-shot localization performance. Table 5 reports the zero-shot localization performance on 4 GenAI inpainting methods. Our proposed HiFi-Net++ achieves the best localization performance among the compared methods, except in the case of Stable Diffusion 1.5 Inpainting, where HiFi-Net++ has a lower AUC score than HiFi-Net Guo et al. (2023b).

Localization Comparison with OneFormer OneFormer Jain et al. (2023), a recently proposed segmentation method, exhibits strong performance in dense prediction tasks. Therefore, we utilize OneFormer for the manipulation localization task and present the results in Table 5 to provide a more comprehensive comparative study. The best OneFormer variant is based on ConvNeXt and achieves a worse localization performance compared to PSCC-Net: 0.64 and 0.32 lower AUC scores on Columbia and Coverage, respectively. We believe the reason for OneFormer's poor manipulation localization performance is that OneFormer is a segmentation method, and there is an inherent difference between segmentation and manipulation localization tasks. As discussed in the previous work Dong et al. (2022); Zhou et al. (2020), the segmentation requires the proposed method to have abilities in object recognition and semantics understanding, yet manipulation localization focuses on learning subtle visual artifacts and frequency domain differences among real and forged pixels, which can be almost invisible to humans.

5.4 CNN Image Detection

Table 6 reports different methods' detection performances on CNN-generated images. Specifically, consistent with UniDet, we use the identical training dataset, which contains forged images generated from ProGAN Karras et al. (2018) and real images from LSUN dataset Yu et al. (2015). We

Table 3 IFDL results on the image editing domain

Localization	Col.	Cov.	NI.16	CAS.	IM20	Avg.
	Metric: AUC(%)–Pre-trained					
(a)						
ManT. (Wu et al. 2019)	82.4	81.9	79.5	81.7	74.8	80.0
SPAN (Hu et al. 2020)	93.6	92.2	84.0	79.7	75.0	84.9
PSCC-Net (Liu et al. 2022)	98.2	84.7	85.5	82.9	80.6	86.3
Ob.Fo. (Wang et al. 2022)	95.5	92.8	87.2	84.3	82.1	88.3
HiFi-Net (Guo et al. 2023b)	98.3	93.2	87.0	85.8	82.9	89.4
NCL (Zhou et al. 2023)	94.3	92.8	**91.2**	86.4	86.4	90.7
TANet (Shi et al. 2023)	**98.7**	91.4	85.3	**89.8**	84.9	90.4
HiFi-Net++	98.5	**94.7**	87.4	88.6	**86.6**	**91.2**

Localization	Cov.	CAS.	NI.16	Avg.
	Metric: AUC(%) / F1(%)–Fine-tuned			
(b)				
SPAN (Hu et al. 2020)	93.7/55.8	83.8/40.8	96.1/58.2	91.2/51.6
PSCC-Net (Liu et al. 2022)	94.1/72.3	87.5/55.4	99.6/81.9	93.7/69.8
Ob.Fo. (Wang et al. 2022)	95.7/75.8	88.2/57.9	99.6/82.4	94.5/72.0
HiFi-Net (Guo et al. 2023b)	96.1/80.1	88.5/59.6	98.9/85.0	94.6/75.5
TANet (Shi et al. 2023)	**97.8**/78.2	89.3/61.4	**99.7**/86.5	95.5/75.4
HiFi-Net++	97.5/**82.7**	**90.3**/**65.4**	**99.7**/**86.8**	**95.8**/**76.6**

(a) Localization performance of the pre-trained model. (b) Localization performance of the fine-tuned model. All results of prior works are ported from Wang et al. (2022). [Key: **Best**; Second Best.]

Table 4 (a) Detection performance on *CASIA* dataset

Detection	AUC(%)	F1(%)
(a)		
ManT. (Wu et al. 2019)	59.9	56.7
SPAN (Hu et al. 2020)	67.3	63.8
PSCC-Net (Liu et al. 2022)	99.5	97.1
Ob.Fo. (Wang et al. 2022)	99.7	97.3
HiFi-Net (Guo et al. 2023b)	99.5	97.4
HiFi-Net++	**99.7**	**98.7**

Post.	SPAN (Hu et al. 2020)	PSCC-Net (Liu et al. 2022)	Obj.Fo. (Wang et al. 2022)	HiFi-Net (Guo et al. 2023b)	HiFi-Net++
(b)					
Resize (0.78)	83.24	85.29	87.2	86.9	**87.7**
Resize (0.25)	80.32	85.01	86.3	86.5	**86.7**
Gau.Blur (3)	83.10	85.38	85.97	86.1	**86.9**
Gau.Blur (15)	79.15	79.93	80.26	**81.0**	80.1
Gau.Noi (3)	75.17	78.42	79.58	**81.9**	80.3
Gau.Noi (15)	67.28	76.65	78.15	79.5	**80.3**
J-Co. (100)	83.59	85.40	86.37	86.3	**87.4**
J-Co. (50)	80.68	85.37	**86.24**	86.0	85.6
Aver.	79.07	82.68	83.75	84.2	**85.1**

(b) Localization performance on *NIST16* with different post-processing steps. First three column results are ported from Wang et al. (2022). Key: **Best**; Gau.: Gaussian; J-Co.: JPEG Compression.]

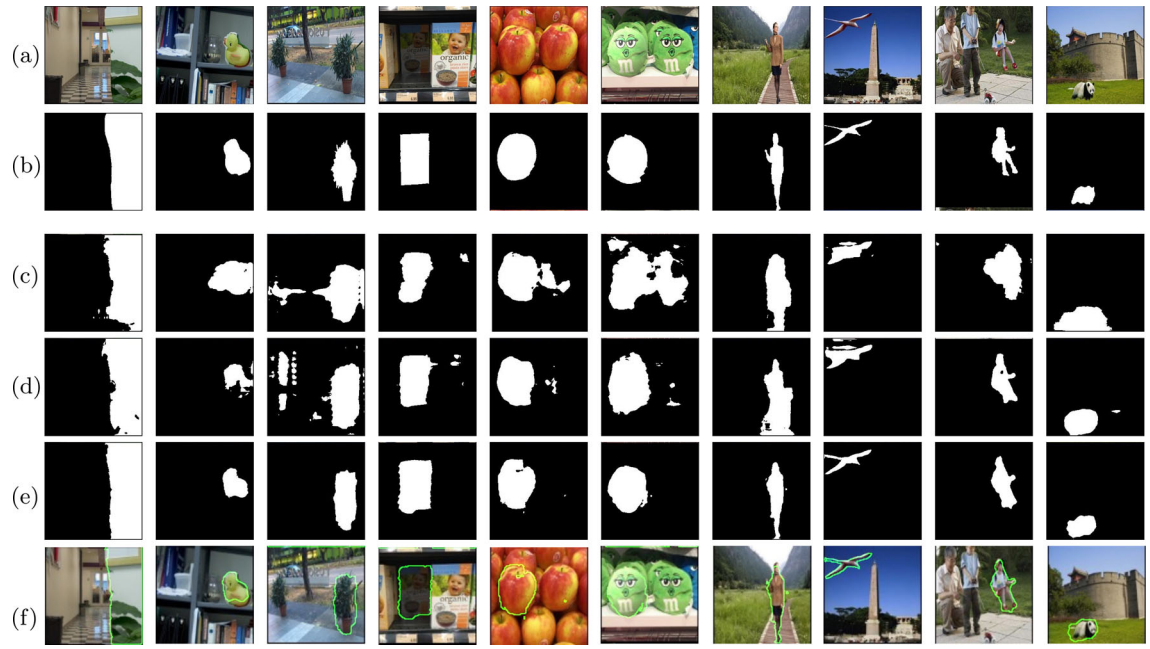


Fig. 13 Qualitative results on different image editing datasets. From top to bottom rows, these rows are **a** input image, **b** the ground truth mask, **c** PSCC-Net result, **d** HiFi-Net result Guo et al. (2023b), **e** HiFi-Net++ result, and **f** the overlaid segmentation on the images of HiFi-Net++ result

employ Average Precision (AP) as the metric, the same as the previous work, for a fair comparison. To have a thorough evaluation, we construct a forgery detection benchmark including 11 generative methods that are commonly used for measuring generalizable detection performance Wang et al. (2020b); Jeong et al. (2022b, a); Ricker et al. (2022); Ojha et al. (2023) and 7 representative GenAI methods. More formally, to have better diversity in forgery methods, these GenAI methods span from classical Stable Diffusion models Rombach et al. (2022), commercial tools (e.g., Sora, Midjourney, and DALLE-3), to the recent popular personalized method like InstantID Wang et al. (2024). Also, following the previous work Wang et al. (2020b), we add real images that have identical semantics (i.e., face and non-face objects) as forged images in this benchmark.

As indicated by Table 6, our proposed HiFi-Net++ achieves the best image forgery detection performance, surpassing the second and third best detection methods (i.e., Uni-Det and DE-FAKE) by 4.8% and 10.65% in terms of Average Precision, respectively. Although Uni-Det and DE-FAKE leverage CLIP pre-trained encoders with powerful visual representations that help detect forged images, they only achieve the best detection performance on 5 and 2 individual forgery methods, respectively. In contrast, HiFi-Net++ achieves the best detection performance on 9 individual forgery method detections. We believe this performance gap is due to two factors. First, HiFi-Net's multi-branch feature extractor is carefully devised and can learn more comprehensive artifacts from generative images, improving the overall image-level forgery detection performance. Secondly, our HiFi-Net++ is trained for both forgery detection and manipulation localization tasks, and such multi-task learning further enhances the representation's effectiveness in capturing forgery attributes. Additionally, DIRE achieves the best performance over images generated from GenAI methods (3 best performance out of 7 Gen-AI methods), whereas it suffers from poor detection precision among images generated from GANs (e.g., ProGAN, CycleGAN, etc.). This phenomenon indicates that DIRE's detection performance is limited to certain forgery types, and its overall performance is worse than HiFi-Net++.

However, it is worth mentioning that HiFi-Net++ detection performance on DALLE-3 images falls behind prior work. This is because the multi-branch feature extractor in the HiFi-Net++, which helps achieve leading average detection AP, the most important aspect for real-world applications, might be less effective in detecting DALLE-3. However, a simple modification—reducing the number of branches in HiFi-Net++—significantly improves detection AP on DALLE-3. Table 7 reports the 1-branch and 2-branch HiFi-Net++ variants, which achieve 24.7% and 11.9% higher AP than the 4-branch HiFi-Net++ variant (i.e., full model), respectively. Nevertheless, these HiFi-Net++ variants with fewer branches ineffectively learn diverse forgery patterns, decreasing averaged detection performance. This suggests that a meaningful future research direction is to develop an adaptive branch control mechanism, dynamically adjusting

Table 5 Zero-shot localization performance on held-out GenAI methods and manipulation localization comparison with OneFormer Jain et al. (2023)

Method	SD 1.5 Inpaint	SD 2.1 Inpaint	SD-XL Inpaint	RePaint	Average
	Metric: AUC/F1(%)				
(a)					
PSCC-Net (Liu et al. 2022)	.544/.395	.509/.378	.492/.436	.632/.471	.511/.428
MVSS-Net (Dong et al. 2022)	.683/.425	.638/.471	.606/.398	.589/.416	.629/.414
HiFi-Net (Guo et al. 2023b)	**.732**/.497	.678/.483	.685/.487	.751/.559	.712/.506
HiFi-Net++	.728/**.512**	**.723**/**.568**	**.713**/**.544**	**.754**/**.578**	**.730**/**.550**

Method	Backbone	Col.	Cov.	F-Shifter
	Metric: AUC(%)			
(b)				
OneFormer (Jain et al. 2023)	Swin-T	.714	.696	.755
OneFormer (Jain et al. 2023)	ConvNext	.723	.707	.805
OneFormer (Jain et al. 2023)	Dinat	.686	.652	.774
PSCC-Net (Liu et al. 2022)	HR-Net	.787	.739	.831
HiFi-Net++	Multi-branch Feature Extractor	**.821**	**.786**	**.834**

(a) The zero-shot localization performance on GenAI inpainting methods. (b) We apply OneFormer on the manipulation localization task and report the performance (AUC) in the comparison with PSCC-Net and HiFi-Net++. Different models' performances are evaluated on the HiFi-IFDL, and all methods are trained on 55, 508 images from the HiFi-IFDL dataset for a fair comparison. [Key: **Best**; SD: Stable Diffusion]

the number of branches based on the input image for better overall performance.

5.5 Diverse Fake Face Dataset Performance

For the facial image forgery domain, we evaluate our methods on the Diverse Fake Face Dataset (DFFD) Stehouwer et al. (2020), which has fake facial images synthesized by different facial forgery methods and real faces from FFHQ Karras et al. (2019) as well as CelebA Liu et al. (2015). For a fair comparison, we follow the same experiment setup and metrics: for the pixel-level localization, we use IoU, Cosine Similarity, pixel-wise binary classification accuracy (PBCA) and IINC defined work Stehouwer et al. (2020); for image-level detection, we use AUC to measure. First, Table 8 show that HiFi-Net++ can produce the better localization performance than the previous work. Specifically, HiFi-Net++ achieves the best localization on fully synthesized fake facial images (0.901 IOU and 0.057 IINC). As for HiFi-Net, it also generates a localization area that has higher IOU and PBCA than the previous work on partially manipulated and fully synthesized images. Secondly, Table 8 reports that, on the detection, both HiFi-Net++ and HiFi-Net achieve comparable performance with the previous method. This is still an impressive detection result since the prior work Stehouwer et al. (2020) uses a more powerful backbone architecture and is specific to the digital facial forgery domain.

5.6 Ablation Study

To provide a more detailed view of how each proposed component contributes to the final performance, we first report the ablation study on localization and classification module in Table 9, and then we use Table 9 to report the ablation study on the language-guided forgery localization enhancer.

HiFi-Net++ Architecture We start with row 1 in Table 9, which represents the performance of the full HiFi-Net++, supervised by $\mathcal{L}_{loc}$, $\mathcal{L}_{cls}$, and $\mathcal{L}_{loc}^{score}$. From row 1 to 2, we first ablate the $\mathcal{L}_{loc}^{score}$, which reduces both detection and localization performances. This is because it is important to have $\mathcal{L}_{loc}^{score}$, which helps Language-guided Forgery Localization Enhancer (i.e., LFLE) generate accurate manipulation score maps that serve as the spatial guidance for identifying the manipulated region. Also, comparing row 1 with row 3, we can see LFLE improves the localization performance (2.4% AUC and 1.7% F1 score), which is consistent with the statement that LFLE improves the effectiveness of the proposed method in identifying manipulated pixels. Furthermore, compared to row 3's performance, we ablate the classification module and localization module in row 4 and 5, removing which causes large performance drops on the detection (24.1% F1) and localization (29.3% AUC), respectively. We evaluate the effectiveness of performing fine-grained classification at different hierarchical levels. In the 6th row, we only keep the 4th level fine-grained classification in train-

Table 6 Image forgery detection performance measured by average precision

Detection method	Generative adversarial networks						Deep- fakes	Low level vision		Perceptual loss		Gen-AI						Avg.
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN	FF++	SITD	SAN	CRN	IMLE	DALL E-3	Mid journey	SD 2.1	SD XL	IF	SORA	Instant ID	
DE-FAKE (Sha et al. 2023)	96.03	97.74	89.93	78.09	98.15	98.34	81.43	88.71	**90.14**	90, 01	95, 14	**88.75**	66.60	79.28	40.65	66.61	27.43	50.11	79.06
Uni-Det. (Ojha et al. 2023)	**100.0**	**99.46**	**99.59**	97.24	**99.98**	99.60	82.45	61.32	79.02	91.72	99.00	74.53	64.93	**95.29**	86.37	90.26	67.94	34.62	84.91
DIRE (Wang et al. 2023)	53.63	51.63	49.11	34.39	48.44	59.23	48.74	48.52	53.13	49.96	50.22	74.70	**99.99**	47.60	49.27	**100.0**	87.62	**100.0**	61.45
Ours	**100.0**	98.54	98.48	**99.60**	97.19	**99.99**	**83.28**	**96.11**	81.57	**92.41**	**99.96**	59.31	78.32	89.00	**92.35**	84.59	**93.24**	72.80	**89.71**

Forgery methods include 11 generative methods used in previous work Wang et al. (2020b); Ojha et al. (2023) and 7 representative Gen-AI methods. Previous methods' quantitative results are taken from Ojha et al. (2023). Each prior detections methods have many variants, so we select the one with the best forgery detection performance for the comparison. [**Bold**: best result.]

Table 7 Detection performance of HiFi-Net++ variants on DALLE-3 images and their averaged performance across 18 forged methods in Table 6

HiFi-Net++ variants	DALLE-3	Avg.
	Metric: AP(%)	
1-branch	**84.0**	78.1
2-branch	71.2	83.1
4-branch (*full model*)	59.3	**89.7**

[Key: **Best**, Avg.: Average detection performance, AP: Average Precision.]

ing, leading to a sensible drop of performance in detection (3.7% AUC) and localization (2.8% AUC). In the 7th row, we perform the fine-grained classification without forcing the hierarchy of Eq. 1, which decreases 3.6% AUC in the detection.

Language-guided Forgery Localization Enhancer In Table 9, row 1 denotes the HiFi-Net++ with a complete language-guided forgery localization enhancer, which has the pre-trained image encoder and uses manipulation score maps as the auxiliary signals. First, from row 1 to row 2, we ablate the manipulation score map generated by text inputs. As a result, the localization performance drops 1.8% AUC and 0.9% F1 scores in the image editing domain. Although the localization performance is not affected largely in the CNN-synthesis domain, we still believe that generated manipulation score maps help the overall localization performance. This phenomenon further shows the pre-trained CLIP model's ability to associate the input text with the corresponding visual contents. After that, we experiment with different image encoder backbones (from row 2 to 4), such as ResNet50, ResNet100 and ViT-B-16. Apparently, all these image encoder backbones provide a robust image embedding that enables the distinction between real and manipulated pixels, and these performances do not vary largely from each other. Subsequently, we replace the pre-trained CLIP image encoder with different ones that are trained for different tasks. Specifically, the performance gap between row 2 and row 5, as well as row 6, indicates the fact that visual embedding from the pre-trained CLIP image encoder is the better feature space for the manipulation localization task. In the last row, comparing to this baseline performance, we can observe that regardless of which pre-trained CLIP image encoder is used, the model always achieves the better localization performance over the baseline. For example, when using ResNet50 as the image encoder backbone, the localization performance improves 3.0% AUC and 3.5% F1 in the image editing domain, and 0.5% higher AUC in the CNN-synthesis domain.

Hierarchical Fine-grained Scheme To show how different hierarchical fine-grained schemes (HiFi-schemes) contribute

Table 8 IFDL results on DFFD dataset

IoU (↑) / PBCA (↑)	Real	Fu. Syn	Par. Man
Att. (Stehouwer et al. 2020)	−/**0.998**	0.847/0.847	0.401/0.786
HiFi-Net (Guo et al. 2023b)	−/0.978	0.893/0.893	0.411/0.801
HiFi-Net++	−/0.974	**0.901**/**0.901**	**0.412**/**0.830**

(a) Localization performance measured by IoU and PBCA, which are the higher the better.

IINC (↓) / C.S. (↓)	Real	Fu. Syn	Par. Man
Att. (Stehouwer et al. 2020)	0.015/−	0.077/0.095	**0.311**/0.429
HiFi-Net (Guo et al. 2023b)	**0.010**/−	0.060/0.107	0.323/0.410
HiFi-Net++	0.012/−	**0.057**/**0.092**	0.313/**0.402**

(b) Localization performance measured by IINC and Cosine Similarity, which are the lower the better.

Metric: AUC	
Att.Xce. (Stehouwer et al. 2020)	99.69
HiFi-Net (Guo et al. 2023b)	99.45
HiFi-Net++	**99.71**

(c) Detection Performance

[Keys: Key: **Best**. Fu. Syn.: Fully-synthesized; Par. Man.: Partially-manipulated. C.S.: consine similarity]

Table 9 Ablation study on HiFi-Net++

	Modules	Loss	Detection		Localization	
			AUC	F1	AUC	F1
1	**L**,**C**, **LFLE**	$\mathcal{L}_{cls}$, $\mathcal{L}_{loc}$, $\mathcal{L}_{loc}^{score}$	**97.0**	**94.3**	**97.7**	**98.6**
2	**L**,**C**, **LFLE**	$\mathcal{L}_{cls}$, $\mathcal{L}_{loc}$	96.6	93.9	95.7	97.4
3	**L**,**C**	$\mathcal{L}_{cls}$, $\mathcal{L}_{loc}$	96.8	94.1	95.3	96.9
4	**L**	$\mathcal{L}_{loc}$	65.0	70.0	93.4	95.0
5	**C**	$\mathcal{L}_{cls}$	95.8	92.4	66.0	58.0
6	**L**,**C**	$\mathcal{L}_{cls}^{4}$,$\mathcal{L}_{loc}$	93.1	91.7	92.5	93.9
7	**L**,**C**	$\mathcal{L}_{cls}^{ind}$,$\mathcal{L}_{loc}$	93.2	92.8	93.2	94.8

(a) Ablation study on the Classification module, Localization modules, and Language-guided Forgery Localization Enhancer, denoted as **C**, **L**, and **LFLE**, respectively. $\mathcal{L}_{cls}$, $\mathcal{L}_{loc}$, and $\mathcal{L}_{loc}^{score}$ are loss functions for classification, localization, and manipulation score maps, respectively. $\mathcal{L}_{cls}^{4}$ and $\mathcal{L}_{cls}^{ind}$ denote we only perform the fine-grained classification on 4^{th} level and classification without hierarchical path prediction. These model variants' performances are based on the image feature extracted by the proposed multi-branch feature extractor, which captures comprehensive generation artifacts.

Local	Image encoder	Pre-training	Mani. scr. mask	Image edit.		CNN-syn.	
				AUC	F1	AUC	F1
1	ResNet-50	CLIP (Radford et al. 2021)	✓	**97.8**	**94.7**	98.6	**98.1**
2	ResNet-50	CLIP (Radford et al. 2021)		96.0	93.6	**98.9**	97.6
3	ResNet-101	CLIP (Radford et al. 2021)		95.6	93.3	98.1	98.1
4	ViT-B-16	CLIP (Radford et al. 2021)		95.8	93.8	98.1	97.4
5	ResNet-50[1]	ImageNet (Deng et al. 2009)		74.0	56.2	87.0	70.5
6	HR-Net[2]	Cityscapes (Cordts et al. 2016)		89.8	77.6	93.4	79.5
Baseline	–	–		93.0	90.1	98.4	97.0

(b) Ablation study on alternative designs of language-guided forgery localization enhancer. ResNet-50[1] is pre-trained on ImageNet dataset for image recognition. HR-Net[2] (Wang et al. 2020a) is pre-trained on Cityscapes dataset for the semantic segmentation.

[Key: **Best**, Mani. Scr. Mask: manipulation score masks]

Table 10 IFDL performance of HiFi-Net++ using different hierarchical fine-grained schemes

	Method	Detection		Localization	
		AUC	F1	AUC	F1
1	–	93.2	92.8	93.2	94.8
2	S,$\mathcal{L}_{hier.}$ (Chen et al. 2022)	95.5	**94.2**	94.9	95.9
3	$\mathcal{L}_{path}$ (ours)	**96.8**	94.1	**95.3**	**96.9**

[Key: **Bold**.]

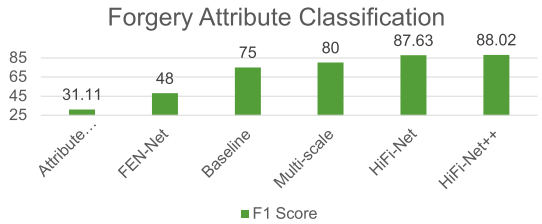


Fig. 14 The forgery attribute classification results

to the IFDL performance, we compare our proposed Hierarchical Path Prediction (e.g., $\mathcal{L}_{path}$) in HiFi-Net++ with the HiFi-scheme proposed in HRN Chen et al. (2022) and evaluate their performances on the HiFi-IFDL dataset. Specifically, HRN's HiFi-scheme employs a probabilistic classification $\mathcal{L}_{Hier.}$ with a pre-defined state space S to help capture the hierarchy among different tree nodes. Table 10 reports IFDL performance of two HiFi-schemes. The line #1 indicates the performance without the HiFi-scheme, while the line #2 and line #3 indicate performances using HiFi-schemes from HRN (i.e., state space S with $\mathcal{L}_{Hier.}$) and HiFi-Net++ (i.e., $\mathcal{L}_{path}$), respectively. Comparing the performance of line #1 against line #2 and #3, we can conclude that the HiFi-scheme contributes to IFDL performance. Moreover, using the $\mathbf{L}_{path}$ can achieve slightly better localization performance and comparable detection performance as HRN's $\mathbf{L}_{hier.}$ and S_F. This indicates the effectiveness of our proposed $\mathbf{L}_{path}$ in capturing the inherent hierarchical correlation among various forgery attributes.

5.7 Forgery Attribute Performance

We perform the fine-grained classification among real images and 13 forgery categories on 4 different levels, and the most challenging scenario is the fine-grained classification on the 4-th level. The result is reported in Fig. 14. Specifically, we train HiFi-Net 4 times, and at each time, only classifies the fine-grained forgery attributes at one level, denoted as *Baseline*. Then, we train a HiFi-Net Guo et al. (2023b) to classify all 4 levels but without the hierarchical dependency via Eq. 1, denoted as *multi-scale*. Also, we compare to the pre-trained image attribution works Asnani et al. (2021); Yu et al. (2019) and the new proposed HiFi-Net++. Both HiFi-Net and HiFi-Net++ have improvements over previous works Yu et al. (2019); Asnani et al. (2021), which we believe is because the previous works only learn to attribute CNN-synthesized images, yet do not consider attributing image editing methods. Again, the performance difference between HiFi-Net and HiFi-Net++ is not large, since the language-guided forgery localization enhancer mainly improves the localization performance.

6 Conclusion

In this work, we develop an IFDL method for both CNN-synthesized and image-editing forgery domains. We formulate the IFDL as a hierarchical fine-grained classification problem that requires the algorithm to classify the individual forgery method of given images via predicting the entire hierarchical path. Moreover, we improve our algorithm generalization ability and performance on localizing small-region manipulations via a language-guided forgery localization enhancer. Lastly, HiFi-IFDL dataset is proposed to further help the community in developing forgery detection algorithms.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-024-02255-9.

Acknowledgements This work was supported by the following projects. One is the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112090131 at Michigan State University, the others are Sapienza research projects "Prebunking", "Adagio", and "Risk and Resilience factors in disadvantaged young people: a multi-method study in ecological and virtual environments".

Funding Defense Sciences Office, DARPA (00112090131).

Data Availability The Hierarchical Fine-grained IFDL dataset—HiFi-IFDL dataset—is publicly available at the following repository github.com/CHELSEA234/HiFi-IFDL.

References

(2010). Survey: More americans get news from internet than newspapers or radio. http://www.cnn.com/2010/TECH/03/01/social.network.news/index.html

(2016). Nist: Nist nimble 2016 datasets. https://www.nist.gov/itl/iad/mig/

(2022). Infodemic–world health organization. https://www.who.int/health-topics/infodemic

Asnani, V., Yin, X., Hassner, T., et al. (2021). Reverse engineering of generative models: Inferring model hyperparameters from generated images. arXiv preprint arXiv:2106.7873

Bui, T., Yu, N., & Collomosse, J. (2022). Repmix: Representation mixing for robust attribution of synthesized images. In ECCV.

Burt, P. J., & Adelson, E. H. (1987). The Laplacian pyramid as a compact image code. *Readings in computer vision.* Elsevier.

Chai, L., Bau, D., Lim, S. N., et al. (2020). What makes fake images detectable? understanding properties that generalize. In ECCV.
Chen, J., Wang, P., Liu, J., et al. (2022). Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4858–4867).
Chen, X., Dong, C., Ji, J., et al. (2021). Image manipulation detection by multi-view multi-scale supervision. In ICCV.
Choi, Y., Uh, Y., Yoo, J., et al. (2020). Stargan v2: Diverse image synthesis for multiple domains. In CVPR.
Cordts, M., Omran, M., Ramos, S., et al. (2016). The cityscapes dataset for semantic urban scene understanding. In CVPR (pp. 3213–3223).
Corvi, R., Cozzolino, D., Zingarini, G., et al. (2022). On the detection of synthetic images generated by diffusion models. arXiv preprint arXiv:2211.0680
Cozzolino, D., Thies, J., Rössler, A., et al. (2018). Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.2510
Deb, D., Liu, X., & Jain, A. (2023). Unified detection of digital and physical face attacks. In FG.
Deng, J., Dong, W., Socher, R., et al. (2009). Imagenet: A large-scale hierarchical image database. In CVPR (pp. 248–255).
Dolhansky, B., Howes, R., Pflaum, B., et al. (2019). The deepfake detection challenge (DFDC) preview dataset. arXiv preprint arXiv:1910.8854
Dong, C., Chen, X., Hu, R., et al. (2022). Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. In TPAMI.
Dong, J., Wang, W., & Tan, T. (2013). Casia image tampering detection evaluation database. In 2013 IEEE China summit and ICSIP.
Dong, X., Bao, J., Zheng, Y., et al. (2023). Maskclip: Masked self-distillation advances contrastive language-image pretraining. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10995–11005).
Dufour, N., Gully, A., Karlsson, P., et al. (2019). Deepfakes detection dataset by Google & Jigsaw.
Gao, P., Geng, S., Zhang, R., et al. (2023). Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision, 132*, 1–15.
Ghiasi, G., Gu, X., Cui, Y., et al. (2022). Scaling open-vocabulary image segmentation with image-level labels. In ECCV (pp. 540–557).
Guo, X., Asnani, V., Liu, S., et al. (2023a). Tracing hyperparameter dependencies for model parsing via learnable graph pooling network. arXiv preprint arXiv:2312.2224
Guo, X., Liu, X., Ren, Z., et al. (2023b). Hierarchical fine-grained image forgery detection and localization. In CVPR (pp. 3155–3165).
He, W., Jamonnak, S., Gou, L., et al. (2023). Clip-s4: Language-guided self-supervised semantic segmentation. In CVPR.
Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In NeurIPS.
Ho, J., Chan, W., Saharia, C., et al. (2022). Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.2303
Hu, X., Zhang, Z., Jiang, Z., et al. (2020). Span: Spatial pyramid attention network for image manipulation localization. In ECCV.
Huang, Y., Juefei-Xu, F., Guo, Q., et al. (2022). Fakelocator: Robust localization of GAN-based face manipulations. In TIFS.
Jain, J., Li, J., Chiu, M.T., et al. (2023). Oneformer: One transformer to rule universal image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2989–2998).
Jeong, Y., Kim, D., Min, S., et al. (2022a). Bihpf: Bilateral high-pass filters for robust Deepfake detection. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 48–57).
Jeong, Y., Kim, D., Ro, Y., et al. (2022b). Frepgan: Robust Deepfake detection using frequency-level perturbations. In Proceedings of the AAAI conference on artificial intelligence (pp. 1060–1068).
Jia, C., Yang, Y., Xia, Y., et al. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In ICML (pp. 4904–4916).
Jiang, L., Li, R., Wu. W., et al. (2020). Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In CVPR.
Karras, T., Aila, T., Laine, S., et al. (2018). Progressive growing of GANS for improved quality, stability, and variation. In ICLR.
Karras, T., Laine, S., Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In CVPR.
Karras, T., Aittala, M., Hellsten, J., et al. (2020). Training generative adversarial networks with limited data. In NeurIPS.
Karras, T., Aittala, M., Laine, S., et al. (2021). Alias-free generative adversarial networks. In NeurIPS.
Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In ICLR.
Kwon, M. J., Nam, S. H., Yu, I. J., et al. (2022). Learning jpeg compression artifacts for image manipulation detection and localization. In IJCV.
Lee, C. H., Liu, Z., Wu, L., et al. (2020). Maskgan: Towards diverse and interactive facial image manipulation. In CVPR.
Li, B., Weinberger, K. Q., Belongie, S., et al. (2022a). Language-driven semantic segmentation. In ICLR.
Li, L., Bao, J., Yang, H., et al. (2020a). Faceshifter: Towards high fidelity and occlusion aware face swapping. In CVPR.
Li, L., Bao, J., Zhang, T., et al. (2020b). Face x-ray for more general face forgery detection. In CVPR.
Li, X., Zhang, S., Hu, J., et al. (2022b). Image-to-image translation via hierarchical style disentanglement. In CVPR.
Lin, T .Y., Maire, M., Belongie, S., et al. (2014). Microsoft coco: Common objects in context. In ECCV.
Lin, T. Y., Dollár, P., Girshick, R., et al. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117–2125).
Liu, M., Ding, Y., Xia, M., et al. (2019). Stgan: A unified selective transfer network for arbitrary image attribute editing. In CVPR.
Liu, X., Liu, Y., Chen, J., et al. (2022). Pscc-net: Progressive Spatio-channel correlation network for image manipulation detection and localization. In TCSVT.
Liu, Z., Luo, P., Wang, X., et al. (2015). Deep learning face attributes in the wild. In ICCV.
Lugmayr, A., Danelljan, M., Romero, A., et al. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In CVPR
Marra, F., Gragnaniello, D., Cozzolino, D., et al. (2018). Detection of GAN-generated fake images over social networks. In MIPR.
Marra, F., Gragnaniello, D., Verdoliva, L., et al. (2019). Do GANS leave artificial fingerprints? In MIPR.
Masi, I., Killekar, A., Mascarenhas, R.M., et al. (2020). Two-branch recurrent network for isolating Deepfakes in videos. In ECCV.
Mayer, O., & Stamm, M. C. (2018). Learned forensic source similarity for unknown camera models. In ICASSP.
Ng, T. T., Hsu, J., & Chang, S. F. (2009). Columbia image splicing detection evaluation dataset. DVMM lab Columbia Univ CalPhotos Digit Libr.
Nichol, A., Dhariwal, P., Ramesh, A., et al. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In ICML.
Novozamsky A., Mahdian B., Saic S. (2020). Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In WACV workshop.
Ojha U., Li Y., Lee Y. J. (2023). Towards universal fake image detectors that generalize across generative models. In CVPR (pp. 24480–24489).

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. In Advances in neural information processing systems (vol. 35, pp. 27730–27744).
Pérez, P., Gangnet, M., & Blake, A. (2003). Poisson image editing. In ACM SIGGRAPH.
Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In ICML (pp. 8748–8763).
Ramesh, A., Dhariwal, P., Nichol, A., et al. (2022). Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.6125
Rao, Y., Zhao, W., Chen, G., et al. (2022). Denseclip: Language-guided dense prediction with context-aware prompting. In CVPR (pp. 18082–18091).
Ricker, J., Damm, S., Holz, T., et al. (2022). Towards the detection of diffusion model deepfakes. arXiv preprint arXiv:2210.14571
Rombach, R., Blattmann, A., Lorenz, D., et al. (2022). High-resolution image synthesis with latent diffusion models. In CVPR.
Rössler, A., Cozzolino, D., Verdoliva, L., et al. (2019). Faceforensics++: Learning to detect manipulated facial images. In ICCV.
Ruff, L., Vandermeulen, R., Goernitz, N., et al. (2018). Deep one-class classification. In ICML.
Sabir, E., Cheng, J., Jaiswal. A., et al. (2019). Recurrent convolutional strategies for face manipulation detection in videos. In Media forensics CVPR workshop.
Saharia, C., Chan, W., Saxena, S., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487
Scherhag, U., Rathgeb, C., Merkle, J., et al. (2019). Face recognition systems under morphing attacks: A survey. *IEEE Access, 7*, 23012–23026.
Sencar, H. T., Verdoliva, L., & Memon, N. (2022). *Multimedia forensics*. Springer.
Sha, Z., Li, Z., Yu, N., et al. (2023). De-fake: Detection and attribution of fake images generated by text-to-image generation models. In Proceedings of the 2023 ACM SIGSAC conference on computer and communications security (pp. 3418–3432).
Shi, Z., Chen, H., & Zhang, D. (2023). Transformer-auxiliary neural networks for image manipulation localization by operator inductions. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(9), 4907–4920.
Singer, U., Polyak, A., Hayes, T., et al. (2022). Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792
Song, J., Meng, C., Ermon, S. (2021). Denoising diffusion implicit models. In ICLR.
Stehouwer, J., Dang, H., Liu, F., et al. (2020). On the detection of digital face manipulation. In CVPR.
Sun, K., Chen, S., Yao, T., et al. (2023). Towards general visual-linguistic face forgery detection. arXiv preprint arXiv:2307.16545
Wang, J., Sun, K., Cheng, T., et al. (2020). Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*, 3349–3364.
Wang, J., Wu, Z., Chen, J., et al. (2022). Objectformer for image manipulation detection and localization. In CVPR.
Wang, Q., Bai, X., Wang, H., et al. (2024). Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.7519
Wang, S. Y., Wang, O., Zhang, R., et al. (2020b). CNN-generated images are surprisingly easy to spot...for now. In CVPR.
Wang, X., Girshick, R., Gupta, A., et al. (2018). Non-local neural networks. In CVPR.
Wang, Z., Bao, J., Zhou, W., et al. (2023). Dire for diffusion-generated image detection. arXiv preprint arXiv:2303.9295
Wen, B., Zhu, Y., Subramanian, R., et al. (2016). Coverage–a novel database for copy-move forgery detection. In ICIP.
Wu, H., Zhou, J., & Zhang, S. (2023). Generalizable synthetic image detection via language-guided contrastive learning. arXiv preprint arXiv:2305.13800
Wu, Y., Abd-Almageed, W., & Natarajan, P. (2018). Busternet: Detecting copy-move image forgery with source/target localization. In ECCV.
Wu, Y., Abd Almageed, W., & Natarajan, P. (2019). Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In CVPR.
Wu, H., et al. (2022). Robust image forgery detection over online social network shared images. In CVPR.
Xu, J., De Mello, S., Liu, S., et al. (2022). Groupvit: Semantic segmentation emerges from text supervision. In CVPR (pp. 18134–18144).
Xu, J., Liu, S., Vahdat, A., & et al. (2023). Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: CVPR, pp 2955–2966
Yao, Y., Zhang, A., Zhang, Z., et al. (2021). Cpt: Colorful prompt tuning for pre-trained vision-language models. arXiv preprint arXiv:2109.11797
Yao, Y., Guo, X., Asnani, V., et al. (2024). Reverse engineering of deceptions on machine-and human-centric attacks. *Foundations and Trends® in Privacy and Security, 6*(2), 53–152.
Yu, F., Seff, A., Zhang, Y., et al. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.3365
Yu, N., Davis, L. S., & Fritz, M. (2019). Attributing fake images to GANS: Learning and analyzing GAN fingerprints. In ICCV.
Zhang, H., Goodfellow, I., Metaxas, D., & et al. (2019a). Self-attention generative adversarial networks. In: ICML
Zhang, R., Fang, R., Zhang, W., et al. (2021). Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.3930
Zhang, X., Karaman, S., & Chang, S. F. (2019b). Detecting and simulating artifacts in GAN fake images. In WIFS.
Zhang, Y., Colman, B., Guo, X., et al. (2024). Common sense reasoning for deep fake detection. In ECCV.
Zhao, T., Xu, X., Xu, M., et al. (2021). Learning self-consistency for deepfake detection. In CVPR.
Zhong, Y., Yang, J., Zhang, P., et al. (2022). Regionclip: Region-based language-image pretraining. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16793–16803).
Zhou, C., Loy. C. C., & Dai. B. (2022a). Extract free dense labels from clip. In ECCV (pp. 696–712).
Zhou, J., Ma, X., Du, X., et al. (2023). Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning. In Proceedings of the IEEE/CVF international conference on computer vision (ICCV) (pp. 22346–22356).
Zhou, K., Yang, J., Loy, C. C., et al. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision, 130*(9), 2337–2348.
Zhou, P., Han, X., Morariu. V.I., et al. (2018). Learning rich features for image manipulation detection. In CVPR.
Zhou, P., Chen, B. C., Han, X., et al. (2020). Generate, segment, and refine: Towards generic manipulation segmentation. In AAAI.