AnimeDL-2M: Million-Scale AI-Generated Anime Image Detection and Localization in Diffusion Era

Chenyang Zhu[1,2], Xing Zhang[2], Yuyang Sun[1,2], Ching-Chun Chang[2], Isao Echizen[1,2]
[1]The University of Tokyo, [2]National Institute of Informatics, Japan

ABSTRACT

Recent advances in image generation, particularly diffusion models, have significantly lowered the barrier for creating sophisticated forgeries, making image manipulation detection and localization (IMDL) increasingly challenging. While prior work in IMDL has focused largely on natural images, the anime domain remains underexplored—despite its growing vulnerability to AI-generated forgeries. Misrepresentations of AI-generated images as hand-drawn artwork, copyright violations, and inappropriate content modifications pose serious threats to the anime community and industry. To address this gap, we propose **AnimeDL-2M**, the first large-scale benchmark for anime IMDL with comprehensive annotations. which comprises over two million images including real, partially manipulated, and fully AI-generated samples. Experiments indicate that models trained on existing IMDL datasets of natural images perform poorly when applied to anime images, highlighting a clear domain gap between anime and natural images. To better handle IMDL tasks in anime domain, we further propose **AniXplore**, a novel model tailored to the visual characteristics of anime imagery. Extensive evaluations demonstrate that AniXplore achieves superior performance compared to existing methods. Dataset and code can be found in https://github.com/FlyTweety/AnimeDL2M.

1 INTRODUCTION

The rapid advancements in AI-based image generation and editing methods, especially diffusion models [60], have made image forgery increasingly accessible, sophisticated, and challenging to detect. Traditionally, image manipulations were primarily performed manually using tools like Photoshop [71]. However, AI-based editing methods have significantly simplified the process [3], resulting in highly realistic and difficult-to-detect forgeries [19].

Although researchers have been aware of this threat and new datasets have been proposed, existing image manipulation detection and localization (IMDL) datasets and methods [15, 24, 65] are primarily tailored towards natural scenes and real-world photographs, neglecting domains such as anime imagery. Nonetheless, forged anime images are attracting increasing attention in areas such as copyright protection and content moderation [52]. Given their widespread popularity and extensive use across online communities and commercial markets, addressing forgery in anime images has become a crucial topic [19, 51]. The absence of specialized forgery detection research in this domain represents a notable gap.

Unlike daily images, anime images have distinct visual characteristics such as unique color distributions, line patterns, texture styles, and structural details [25]. Our experiment illustrates that existing forgery detection methods trained on daily images typically exhibit reduced performance in detecting and localizing forgeries within anime images. This limitation underscores the need for specialized datasets designed for IMDL tasks in anime domain.

To address these challenges, we introduce **AnimeDL-2M**, the first large-scale anime-specific image forgery dataset. In addition to its novel domain focus, AnimeDL-2M offers significant advantages in scale, generation variety, annotation richness, and content diversity. It comprises over 2 million images, including real, partially manipulated, and fully AI-generated samples. Fake images are created using six AI-based methods derived from three base models, ensuring both realism and variation and achieving high aesthetic quality scored state-of-the-art perceptual metrics. Each image is paired with comprehensive annotations, including image captions, objects, masks, mask labels and editing methods, enabling a broad range of downstream tasks. AnimeDL-2M also features rich diversity, with a broad set of object categories and manipulation scenarios, thereby providing a comprehensive benchmark for advancing research in AI-generated content detection.

Extensive experiments on AnimeDL-2M demonstrate a significant domain gap between the anime images and daily images. Considering the unique visual characteristics of anime, to better handle IMDL tasks on anime images, we propose **AniXplore**, an IMDL model designed for anime images. It first employs a Mixed Feature Extractor to leverage texture information and object semantics in anime images. Then Dual-Perception Encoder is further introduced to encode and fuse texture-level cues with object-level semantics in two branches. Finally, feature maps are sent to Localization and Classification Predictor to get the prediction result. Through extensive comparative experiments, AniXplore achieves superior performance compared to six leading SOTA models.

Our main contributions are summarized as follows: **(1)** We introduce **AnimeDL-2M**, the first large-scale anime-specific IMDL dataset, featuring over 2 million images with rich annotations and high diversity. **(2)** We propose **AniXplore**, a novel model tailored to synthetic anime detection with generalizability to in-the-wild images. **(3)** We demonstrate the domain gap between anime and daily images, which sheds light on future IMDL research. Our findings underscore the need for domain-specific solutions to support real-world applications such as copyright protection, content moderation, and intellectual property enforcement.

2 RELATED WORK

This section reviews the studies in IMDL, with a focus on both datasets and model designs. We first summarize existing datasets, highlighting the lack of resources dedicated to anime imagery. We then introduce prior models, discussing their feature extraction strategies, backbone networks, and decoder designs, which offer design insights but also highlight the need for specialized solutions in the anime domain.

Dataset	Year	# Images		Domain	Manipulation Types
		Real	Edited		
Columbia [22]	2004	183	180	Daily	Random
CASIAv1 [12]	2013	800	921	Daily	Manual
CASIAv2 [12]	2013	7,491	5,123	Daily	Manual
DSO-1 [11]	2013	100	100	Daily	Manual
Coverage [71]	2016	100	100	Daily	Manual
NIST16 [14]	2016	875	564	Daily	Manual
Fantastic Reality [28]	2019	16,592	19,423	Daily	Manual
IMD20 Manual [54]	2020	-	2,000	Internet	Unknown
IMD20 Synthetic [54]	2020	-	35,000	Daily	Random, Synthetic AI
tampered COCO [30]	2022	-	400,000	Daily	Random
tampered RAISE [30]	2022	24,462	400,000	Daily	Random
COCOGlide [15]	2022	-	512	Daily	Synthetic AI
AutoSplice [24]	2023	2,273	3,621	News	Synthetic AI
MIML [57]	2024	-	123,150	Internet	Unknown
GRE [65]	2024	-	228,650	Daily, News	Synthetic AI
CIMD [81]	2025	-	600	Daily	Manual
AnimeDL-2M (Real & Inpaint Subset)	**2025**	**639,268**	**779,502**	**Anime**	**Synthetic AI**

Table 1: Summary of public image IMDL datasets. AnimeDL-2M is the first IMDL dataset built with the latest diffusion models for anime images. Apart from real images and edited images, AnimeDL-2M also includes 884,129 fully AI-generated images.

2.1 IMDL Datasets

Table 1 summarizes the widely used datasets in IMDL research. Traditionally, most IMDL benchmarks [12, 14, 22, 28, 31, 54, 71] employ classical manipulation techniques such as copy-move, splicing, and object removal, which are often referred to as Photoshop-based methods [6], only a limited number of datasets [15, 37, 49] include large-scale inpainting manipulations. With the rapid progress in generative modeling, text-guided image inpainting has become an emerging trend in dataset construction [6, 50]. For instance, Jia et al.[24] proposed one of the earliest pipelines using DALL·E 2 to generate inpainted samples, while Sun et al.[65] expanded this approach by incorporating a wider range of generative models. In addition, several recent datasets [23, 36, 38, 61, 66, 73] integrate auxiliary metadata or text annotations generated by large language models (LLMs), enabling the exploration of multimodal approaches and strengthening links to the broader domain of disinformation detection.

Despite these advances, existing datasets overwhelmingly focus on natural images, leaving anime-style content—an increasingly popular and distinct visual domain—largely underexplored. To address this gap, we introduce the first large-scale IMDL dataset specifically curated for anime imagery which reflects a real-world application scenario: the growing need for automated copyright protection and content integrity verification in AI-generated anime artworks. Our dataset is characterized by its rich annotations and high diversity, we anticipate that this contribution will stimulate further research at the intersection of multimedia forensics, generative media, and copyright governance.

2.2 IMDL Models

As summarized in [48], most existing models follow a common paradigm: First, they extract auxiliary features from the input image, then feed both the raw image and these features into an encoder network to obtain multi-scale feature maps. Finally, these features are fused and decoded to predict forgery locations and classification results.

Regarding input features, although some studies have demonstrated that auxiliary features are not strictly necessary [47, 63], many state-of-the-art methods rely heavily on them. These include frequency- or edge-based representations extracted via handcrafted filters [56, 69] and noise-based features obtained through trained or learnable extractors [2, 4, 6, 9, 15, 33, 53, 85]. Other studies incorporate semantics-aware features [5, 86], model-specific artifacts [21, 67], or compression-related cues such as JPEG artifacts [30, 32]. Some works further enhance detection by combining multiple auxiliary features [26, 34, 68].

In terms of encoder architectures, traditional approaches largely utilize CNN-based backbones, while more recent efforts have explored Transformer-based designs [16, 26, 35, 47, 63, 78] or hybrid architectures that combine both paradigms [34, 69, 86]. Emerging directions also investigate the use of large vision encoders from LLMs [29, 33, 64, 74] or the integration of LLMs directly into the detection pipeline [17, 23, 36, 38, 66, 73]. Additionally, effective fusion of diverse input features has become a key research focus [4, 16, 26, 34, 82].

Decoder designs have also evolved to better support localization tasks, where multi-scale feature maps play a critical role [4]. Various fusion strategies have been proposed to enhance feature aggregation [18, 42, 56, 62, 76, 82]. In parallel, some studies focus on improving classification accuracy [15]. Moreover, contrastive learning has been employed to refine feature representation [1, 20, 29, 34, 41, 46, 83, 84], while other works introduce novel paradigms and frameworks for IMDL [35, 39, 40, 43, 45, 55, 70, 77, 80].

In this work, we conduct extensive experiments to construct a benchmark for AI-generated anime IMDL and propose the AniXplore model. Leveraging the unique visual characteristics of anime-style imagery, our approach incorporates a hybrid feature extractor and a dual-branch architecture specifically designed to integrate

texture-level cues with object-level semantic information. Extensive experiments demonstrate that this design achieves state-of-the-art performance in IMDL within the anime domain.

3 ANIMEDL-2M DATASET

This section presents AnimeDL-2M, the first million-scale IMDL dataset in the anime domain. We detail the dataset construction pipeline, including data collection, image perception, image segmentation, AI-based image generation, and dataset annotation, followed by an assessment of aesthetic quality and subject diversity. Compared to existing publicly available regional editing datasets detailed in Table 1, AnimeDL-2M dataset offers significant advantages in scale, generation variety, annotation richness, and content diversity.

3.1 Source Data Collection

We collect raw images from Danbooru [10], a widely-used art and anime platform that hosts high-quality, user-annotated images accompanied by detailed tags and textual descriptions. We apply post-processing steps, including filtering out NSFW content to avoid violent or inappropriate material. Additionally, we resize the longer side of each image to 1024 pixels to achieve a balance between visual quality and computational efficiency. To evaluate the performance of benchmark models under realistic conditions, we additionally collected AI-generated anime images from Civitai [8], a popular community platform for AI-generated artwork contributed by users worldwide. Specifically, we retrieved the top 100 highest-rated anime generative models and filtered out those labeled with "realistic" tags or marked with high NSFW content. This filtering process yielded a curated set of 9,071 models spanning 14 base model categories, including Illustrious, Pony, Stable Diffusion (SD) 1.4/1.5/2.0/2.1, FLUX.1 S/D, SDXL 0.9/1.0/Hyper/Lightning/Turbo, among others. Utilizing the image URLs embedded in the metadata of these models, we collected a total of 104,627 high-quality text-to-image (T2I) samples, which serve as a challenging testset for evaluating our manipulation detection framework.

3.2 Dataset Construction

In order to simulate real-world image forgery scenarios while achieving a balance between efficiency and quality, we developed a fully automated pipeline based on large multi-modal models. This pipeline enables the creation of large-scale annotated datasets featuring diverse types of tampering content generated by multiple models. As shown in Figure 1, the pipeline consists of three key steps: (1) *Image perception*, which generates a descriptive caption for each image; (2) *Image segmentation*, which produces region masks to guide AI-based editing; and (3) *Image generation*, which synthesizes both inpainted and fully AI-generated images.

3.2.1 Image Perception. In real-world scenarios, image manipulations are typically driven by specific intentions rather than being performed randomly or arbitrarily. Individuals must first understand the content of an image before making manipulations. In addition, manipulations often occur at the object level [86], such as by removing, adding, or modifying particular objects. Based on these insights, the first stage of our pipeline aims to simulate the process of understanding and decision-making by leveraging a multimodal large model. This stage extracts critical information for downstream tasks. Specifically, we deploy the InternVL2.5 [7], which is one of the best open-source multi-modal large language model, to generate a concise description of the image, which we refer to as image caption. The model is then prompted to enumerate the objects present in the image, which will be used in Image Segmentation stage for generating masks. Given that the downstream image generation model employs CLIP's text encoder [58], which accepts a maximum of 77 tokens, we instruct the large model to produce captions that are both clear and succinct, constrained to fewer than 40 tokens. This leaves sufficient token capacity for additional input components required by the generation model.

3.2.2 Image Segmentation. After identifying the objects for manipulation, we uses them as labels to prompt the GroundedSAM [59] to generate the mask for each object. To enhance the quality of the generated mask, we apply morphological closing operations to smooth its edges and reduce internal holes. During the mask generation process, a single label may yield multiple mask regions; in such cases, we merge them into a unified mask. Furthermore, if the Intersection-over-Union (IoU) between masks associated with different object labels exceeds 0.9, we treat them as overlapping representations and merge them into a single mask as well. As a result of the merging operations, we obtain three types of masks. The first and most common type is the single-instance mask, which contains a single instance of one object. The second type is the multi-instance mask, which includes multiple instances of the same object class. The third type is the multi-class mask, formed by merging highly overlapping masks from different object categories. The inclusion of diverse mask types further enriches the diversity of the AnimeDL-2M dataset and enhances its overall quality.

3.2.3 Image Generation. With the collected raw images, image captions, and object masks, we proceed to Text-to-Image synthesis and AI-inpainting image synthesis. The first step involves selecting appropriate generative models. Previous research has shown that different generative models may leave distinct fingerprints or artifacts in the generated images [75]. To ensure diversity within the dataset and to construct a reliable benchmark for evaluation, it is crucial to employ a variety of generative models during the image synthesis process. Besides, given the anime-specific focus of our dataset, the selected generative models must be capable of producing high-quality images that faithfully adhere to the anime style. Following extensive screening and empirical evaluation, we selected three representative methods for each of the two generation tasks. After generation, we apply image evaluating model MPS [79] to evaluate quality and filter out low quality images.

3.2.4 Dataset Annotations. It is worth noting that the intermediate information obtained from the first two stages of the data pipeline also constitutes valuable annotation data, which can be applied to broader evaluation and detection methodologies, such as further development of multimodal detection approaches based on text semantics or editing method attribution. Therefore, unlike IMDL datasets, we have additionally provided captions, objects, mask labels, and editing methods as extra annotations, which we anticipate to facilitate the future study.

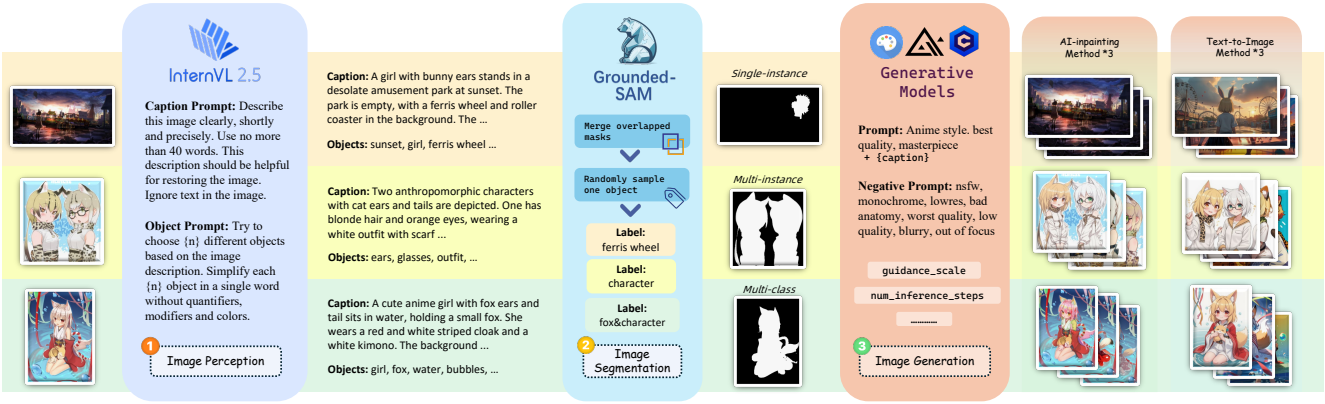

Figure 1: An overview of AnimeDL-2M's data construction pipeline and data example. Image perception component reads the image and outputs image caption as well as objects found in the image. Image segmentation component randomly picks one object and generates its mask for each image. Image generation component uses inpainting and text-to-image methods with 6 different models to create 6 fake images for each raw image. Captions, objects, mask labels, and editing methods serve as extra annotations.

3.3 Dataset Statistics

Table 2 summarizes the composition of the proposed AnimeDL-2M dataset. In Figure 2, We utilized MPS [79], a multi-dimensional preference scoring model for evaluating text-to-image generation, to access the image quaility of AnimeDL-2M dataset. We also present the most frequently manipulated subject when generating images with inpainting methods in Figure 3. To summarize, AnimeDL-2M offers key advantages as below:

New domain. Anime images exhibit significant visual differences from daily images, resulting in a substantial domain gap. Models trained on existing daily image datasets perform poorly when applied to anime images. As the first IMDL dataset in the anime domain, our dataset fills this critical gap and sheds light on many future research directions within the this field.

Large-scale. includes a large number of synthetic samples generated by different models under both Text-to-Image and AI-inpainting settings, as well as a substantial real image subset, outperforming most existing datasets and providing a comprehensive benchmark. Notably, the dataset is balanced across different generative methods and tasks. Its diverse content and balanced distribution benefit both evaluating and training IMDL models.

Synthetic AI. AI-generated image manipulations are becoming increasingly prevalent. Compared to traditional methods, AI-based edits often exhibit globally consistent styles and less distinguishable boundaries, making IMDL more challenging [65]. Unlike conventional IMDL datasets, AnimeDL-2M focuses on AI-based image manipulations and includes fully AI-generated images as well. Moreover, as shown in Figure 2, images in AnimeDL-2M generally receive high aesthetic scores, providing strong evidence of superior image quality and semantic consistency between images and annotations in the AnimeDL-2M dataset.

Rich Annotation. For each group of original, edited, or generated images, AnimeDL-2M not only provides segmentation masks as in traditional datasets, but also includes additional annotations such as image captions, object descriptions, mask labels, and editing methods. These enriched annotations enable a broader range of tasks to be conducted on this dataset and are intended to facilitate future research in related domains.

Subset	Type	# Images
Danbooru	Real	639,268
Stable Diffusion	Text2Image	259,834
Stable Diffusion XL	Text2Image	259,834
FLUX	Text2Image	259,834
Stable Diffusion	Inpainting	259,834
Stable Diffusion XL	Inpainting	259,834
FLUX	Inpainting	259,834
CivitAI	Text2Image	104,627
Total	-	**2,302,899**

Table 2: Statistics of AnimeDL-2M Dataset

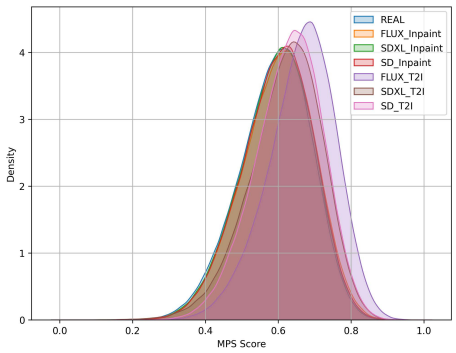

Figure 2: Aesthetic distribution of real and synthetic anime images. Note that inpainted images have a similar distribution to real images.

High Diversity. AnimeDL-2M exhibits strong diversity across the following four dimensions: (1) three distinct types of segmentation masks; (2) six different image forgery methods based on three base models; (3) not only partially manipulated images, but also fully authentic and fully synthetic ones; and (4) diverse objects varying widely in type and content. As shown in Figure 3, the distribution of manipulated subjects is fairly diverse and seemingly

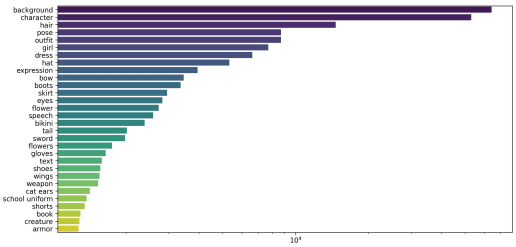

Figure 3: Top30 subject distribution of AnimeDL2M dataset. It exhibits a diverse range of subjects which highlights the open-world nature of the dataset, making it suitable for training robust and generalized IMDL models.

random, which contributes to model generalization and enables a more comprehensive evaluation of model performance.

4 ANIXPLORE MODEL

This section introduces AniXplore, our proposed IMDL model tailored for the anime domain. We present the motivation behind the model design and introduce the overall architecture. Our model consists of a Mixed Feature Extractor, a Dual-Perception Encoder, and a Localization and Classification Predictor, aiming to capture forensic information from both local textures and global semantics.

4.1 Inspiration and Design Overview

Anime images exhibit distinctive visual characteristics that distinguish them from natural daily images, such as unrealistic lighting conditions, geometric abstractions, and the absence of sensor noise. These distinct properties underscore the necessity for specialized methods tailored to the IMDL tasks in the anime domain.

While it is commonly assumed that anime images contain fewer high-frequency components such as complex textures or stochastic noise, an overlooked yet crucial aspect is their retention of edge information in mid-to-high frequencies, especially the line contours. As anime images typically have clean and uncluttered scenes, line work in these images is generally sharp and well-defined. Furthermore, as these lines are manually drawn, they tend to exhibit a consistent artistic style across the image. Consequently, localized inconsistencies in stroke thickness, color, or drawing style may serve as effective cues for identifying image manipulations.

Additionally, prior studies have demonstrated that image manipulations frequently occur at the object level [86]. Anime images, which typically comprise a limited number of semantically salient objects with well-defined boundaries, are especially amenable to object-level semantic reasoning. Motivated by these insights, we propose **AniXplore**, a model with dual-branch architecture that integrates semantic representations with frequency-aware features to enhance the IMDL in AI-generated anime images.

4.2 Mixed Feature Extractor

We integrate the Discrete Wavelet Transform (DWT) into the feature extraction pipeline to enhance high-frequency representation. DWT excels at preserving fine-grained edge structures, making it particularly effective in capturing line-based features such as contours and brush strokes, which serve as critical visual cues in anime images. Furthermore, inspired by [86], which highlights the importance of object-level semantics in manipulation detection, we incorporate low-frequency components derived from the Discrete Cosine Transform (DCT). These features capture the global spatial structure of an image that are particularly relevant in identifying object-level manipulations. To be specific, the Mixed Feature Extractor combines DWT and DCT to process an RGB image $I \in \mathbb{R}^{3\times H\times W}$. It computes 1) High-frequency components M_{high}, derived as the average of high-frequency DWT and DCT coefficients, and 2) Low-frequency DCT components M_{low}. The frequency components are concatenated with the original image I to form mixed high-frequency input M_{high} and mixed low-frequency input M_{low} for the dual-branch encoder, where $M_{\text{high}}, M_{\text{low}} \in \mathbb{R}^{6\times H\times W}$.

$$M_{\text{high}} = I \oplus \frac{1}{2}\left(\text{DWT}_{\text{high}}(I) + \text{DCT}_{\text{high}}(I)\right), \tag{1}$$

$$M_{\text{low}} = I \oplus \text{DCT}_{\text{low}}(I), \tag{2}$$

where $\oplus$ denotes channel-wise concatenation.

4.3 Dual-Perception Encoder

To capture both localized textural patterns and global semantic information in anime images, we design a Dual-Perception Encoder comprising two complementary branches: a *Local Texture Branch* and a *Global Semantics Branch*. This dual-branch architecture ensures comprehensive feature extraction across multiple spatial scales and representation domains. The *Local Texture Branch* is optimized to capture high-frequency details, which are particularly informative in the context of hand-drawn line art and forensic artifacts. We implement it using ConvNeXt [44], a state-of-the-art convolutional architecture that effectively models local patterns. The *Global Semantics Branch* employs attention mechanisms to model long-range dependencies and contextual information. This branch facilitates semantic-level understanding, which is critical for detecting region-level inconsistencies and object-level manipulations. We implement it using Segformer [72] for semantic feature extraction. At each encoding stage, we apply a 1×1 convolutional fusion layer to integrate features from both branches. The fused representation F_i at stage i is computed as:

$$F_i = \text{Fuse}(F_i^{\text{local}} \oplus F_i^{\text{global}}), \tag{3}$$

where F_i^{local} and F_i^{global} are the outputs of the local and global branches at the i-th stage, respectively. The fused feature F_i is then propagated to the subsequent layer of the local branch for progressive refinement with integrated local and global information. The encoder comprises 3 stages, each implementing the dual-branch extraction and fusion mechanism. The final fused output $F_N \in \mathbb{R}^{384\times \frac{H}{16}\times \frac{W}{16}}$ from the third stage is forwarded to the decoder.

4.4 Localization and Classification Predictor

Following [47], we implement a Simple Feature Pyramid Network (SFPN) to transform the encoder output F_N into multi-scale feature maps $\{F_i'\}_{i=1}^5$. Each F_i' is resized to a uniform resolution of $256 \times \frac{H}{4} \times \frac{W}{4}$, after which the resized features are channel-wise concatenated and processed through a 1×1 convolutional layer, yielding a fused feature map of dimensions $256\times \frac{H}{4} \times \frac{W}{4}$. This fused representation is processed by a Multi-Layer Perceptron (MLP) to produce the predicted manipulation mask $\hat{M} \in \mathbb{R}^{1\times \frac{H}{4}\times \frac{W}{4}}$, which

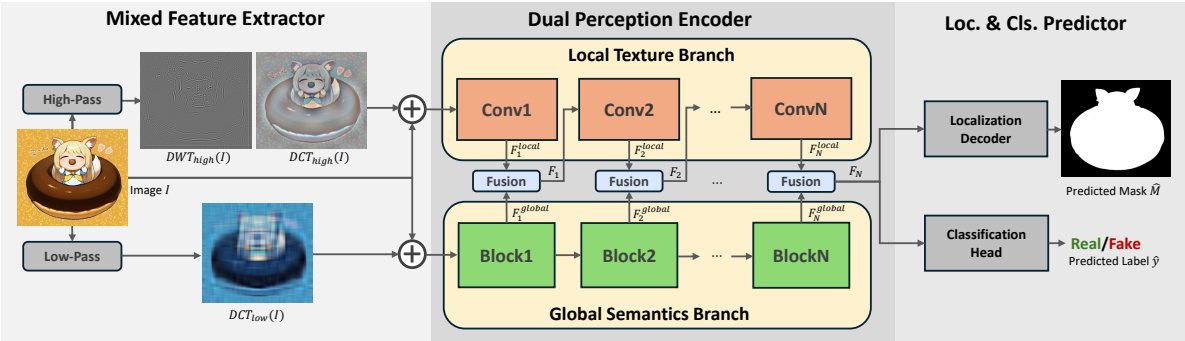

Figure 4: Overview of AniXPlore, which consists of Mixed Feature Extractor, Dual-Perception Encoder, and Localization and Classification Predictor, using information from both local textures and global semantics for anime IMDL.

is subsequently upsampled to the original resolution $H \times W$ to indicate potential forgery regions:

$$\hat{M} = \text{MLP}\Big(\text{Conv}_{1\times1}\Big(\oplus_{i=1}^{5}\text{Resize}(F_i')\Big)\Big), \{F_i'\}_{i=1}^{5} = \text{SFPN}(F_N) \quad (4)$$

For the classification head, we perform global max pooling on F_N to yield a feature vector of shape $C \times 1$, followed by a linear layer for binary prediction.

$$\hat{y} = \text{Linear}(\text{MaxPool}(F_N)) \quad (5)$$

4.5 Loss Function

We employ Binary Cross-Entropy (BCE) loss for both the localization and binary classification tasks. Our experimental analysis indicates that incorporating a classification head can adversely affect localization performance. Recent works omit the classification head [86] or use a two-stage training strategy [15, 68], both of which are either insufficient for comprehensive detection or unnecessarily complex. To address this issue, we employ an Automatic Weighted Loss (AWL) module inspired by the multi-task uncertainty weighting method [27] to mitigate the impact of loss imbalance. The overall loss is formulated as

$$\mathcal{L}_{\text{total}} = \frac{1}{2\sigma_1^2}\mathcal{L}_{\text{loc}}(M, \hat{M}) + \frac{1}{2\sigma_2^2}\mathcal{L}_{\text{cls}}(y, \hat{y}) + \log\sigma_1 + \log\sigma_2 \quad (6)$$

where M and $\hat{M}$ are ground-truth and predicted masks, y and $\hat{y}$ are ground-truth and predicted labels. σ_1 and σ_2 are trainable parameters that represent the uncertainty of each task, allowing the model to adaptively adjust the relative importance of each loss component.

5 EXPERIMENTS

This section describes our experimental setup and evaluation results. We benchmark the proposed AniXplore model and the state-of-the-art methods on AnimeDL-2M and investigate domain gaps. We further examine generalizability through cross-dataset and in-the-wild evaluations, and perform ablation studies to assess the contribution of each design component in our model.

5.1 Benchmark Settings

5.1.1 Baseline Models. We selected six well-known, state-of-the-art open-source IMDL models from literature for comparative evaluation: Mesorch_P (AAAI '25 [86]), MMFusion (MMM '24 [68]), Trufor (CVPR '23 [15]), IML-ViT ('23 [47]), CatNet (IJCV '22 [30]), PSCC (TCSVT '22 [42]), and MVSS (ICCV '21 [4]). These models have demonstrated strong performance through their innovative design, and have been widely recognized by the research community, making them solid baselines for our experiments.

5.1.2 Dataset Partition. We partitioned all image units (real images, synthetic images, and corresponding annotations) using an 8:1:1 train:validation:test ratio, excluding the Civitai subset which was reserved exclusively for evaluating cross-domain generalization. To facilitate fine-grained analysis of generative models' influence, according to the base models used for image generation, we further divide the training, validation, and test sets into three subsets: SD, SDXL, and FLUX. For evaluation purposes, we excluded fully authentic images (those without manipulated regions) since the F1 score metric becomes 0 in such cases. Additionally, we incorporated the GRE dataset [65], a recent public benchmark containing over 200K AI-inpainted images, to investigate domain gaps between photographic and anime imagery. For the GRE dataset, we adhered to the partition scheme provided by the original authors.

5.1.3 Experiment Tasks and Protocol. We designed four experimental tasks to systematically evaluate: (1) the efficacy of AnimeDL-2M and our proposed AniXplore; (2) the domain gaps between traditional and AI-based editing approaches, and between natural and anime images; and (3) the impact of architectural design on performance and generalization ability.

Task 1 & 2. Zero-shot and Training Results. These two tasks evaluate the performance of pre-trained IMDL models on the AnimeDL-2M dataset, investigating the two types of domain gaps discussed earlier. In task 1, we initialize each model with checkpoints trained with Protocol-CAT [48], a widely adopted protocol for IMDL evaluation. Three models are selected and further trained on the the GRE dataset, serving as representatives of IMDL models trained on IMDL datasets. In task 2, we use AnimeDL-2M dataset to both train each baseline model from scratch and finetune them using using the Protocol-CAT checkpoints, and compare them with AniXplore trained on AnimeDL-2M.

Task 3 & 4. Cross-dataset and In-the-wild Tests. These two tasks evaluate model performance on unseen data, providing insights into the model's generalizability and inform future improvements. Specifically, we use the four models with classification head

Pretrain	Model	Pixel-level		Image-level	
		F1	IoU	F1	Acc.
Protocol-CAT	MVSS [4]	0.0295	0.0169	0.6423	0.5231
Protocol-CAT	PSCC [42]	0.0831	0.0553	0.5295	0.3601
Protocol-CAT	CatNet [30]	0.0868	0.0561	/	/
Protocol-CAT	TruFor [15]	0.0235	0.0161	0.8470	0.7411
Protocol-CAT	MMFusion [68]	0.1106	0.0771	**1.0000**	**1.0000**
Protocol-CAT	Mesorch_P [86]	0.0227	0.0151	/	/
GRE	MVSS [4]	0.0577	0.0322	0.9993	0.9986
GRE	CatNet [30]	**0.2177**	**0.1419**	/	/
GRE	Mesorch_P [86]	0.0655	0.0431	/	/

Table 3: Zero-shot results on AnimeDL-2M. GRE refers to [65].

Model	Pixel-level		Image-level	
	F1	IoU	F1	Acc.
MVSS [4]	0.8802	0.8423	1.0000	1.0000
PSCC [42]	0.9398	0.9106	0.9994	0.9989
CatNet [30]	0.9447	0.9135	/	/
TruFor [15]	0.9614	0.9362	0.9962	0.9925
MMFusion [68]	0.8615	0.8247	1.0000	1.0000
Mesorch_P [86]	0.9661	0.9435	/	/
AniXplore	**0.9710**	**0.9506**	**1.0000**	**1.0000**
AniXplore (HR)	**0.9761**	**0.9923**	**1.0000**	**1.0000**

Table 4: Comparison of our model and existing SOTA IMDL models on AnimeDL-2M. Our model is trained from scratch. "HR" denotes high resolution version. For baseline models, since their scores when trained from scratch are either identical to finetuning results or less than 0.1% lower, we only report results finetuning results.

and initialized from checkpoints in the previous task to assess detection performance on the Civitai subset, serving as the in-the-wild test. Additionally, we retrain and evaluate both the detection and localization performance of three baseline models along with our AniXplore on different subsets of the AnimeDL-2M to further investigate domain generalization. We report the score of the checkpoint that reaches the highest average pixel-level F1 score on all subsets.

5.1.4 Metrics. We followed the same extensively used metrics for evaluation. For localization task, we use F1 score with throld = 0.5 and Intersection over Union (IoU). For detection task (for models with classification head), we use image-level F1 score and Accuracy.

5.2 Implementation Details

We train AniXplore on 8 H200 GPUs for 50 epochs with batch size of 72. All images were resized to 512×512 or padded to 1024×1024 pixels for two versions of AniXplore. We used a cosine learning rate schedule, starting at 1e-4 and decaying to 5e-7, with a 2-epoch warm-up. The AdamW optimizer was applied with a weight decay of 0.05 to reduce overfitting. Gradient accumulation was set to 2 to effectively increase the batch size and enhance generalization. We use pre-trained backbones to initialize AniXplore's two branches.

5.3 Zero-Shot and Fine-Tuned Performance

Domain Gaps. As shown in Table 3, the zero-shot performance of pretrained IMDL models on AnimeDL-2M is extremely poor. Most models achieve pixel-level F1 scores below 0.1, indicating a complete failure in localizing manipulated regions. This suggests that models trained on conventional IMDL datasets lack generalizability to the distribution of AI-edited anime images. Models pre-trained on the GRE dataset perform relatively better. This implies that certain features of AI-generated manipulations can be learned and partially transferred. However, these models still perform poorly on localization task. After fine-tuning on AnimeDL-2M, all models show significant improvements as reported in Table 4, confirming both the high training value and the annotation quality of the dataset. Some models achieve surprisingly high F1 scores in detection task. This suggests that while models cannot precisely locate manipulated regions, they can still capture global statistical cues such as unnatural frequency artifacts or noise distribution that distinguish fake images from real ones at a coarse level. In the fine-tuned setting, all models also achieve relatively high localization scores. This is mainly because anime images tend to have clean backgrounds and less noise, which makes artifacts more visually distinct. These findings collectively validate the presence of substantial domain gaps across manipulation methods and image styles, especially for localization tasks. Therefore, AnimeDL-2M serves as a necessary contribution to bridge this gap, offering a dedicated benchmark for AI-edited anime image forensics.

Comparison with SOTA. As presented in Table 4, our model achieves the best performance across all metrics. Although the absolute improvement over previous methods is modest, the gains are meaningful given that most baseline models already get very high scores. These results highlight the effectiveness of our model design in adapting to anime images and capturing manipulation artifacts specific to AI-generated content. Our approach offers beneficial insights for future research in anime and stylized media.

5.4 Cross-Dataset and In-the-Wild Evaluation

Factors Influencing Generalization. As shown in Table 5, all models exhibit generally poor performance on the localization task under cross-dataset settings. This is likely because localization heavily relies on identifying artifact patterns left by AI-generated regions. However, such artifacts vary significantly across generative models, making it difficult to generalize. This also suggests that training or fine-tuning on the target domain can substantially improve localization performance, which is consistent with the findings in [13]. Combining results in Table 6, we can see that generalize ability on detection task does not strongly correlate with localization performance. Some models achieve high detection accuracy across domains despite limited localization ability. This implies that detection generalization may depend more on the robustness of the model architecture than on the ability to capture specific forgery artifacts. Among all tested methods, PSCC [42], as the only model that uses RGB images as the sole input modality, demonstrates the weakest generalization, highlighting the importance of multi-modal or multi-channel feature inputs for generalizability. Meanwhile, both TruFor [15] and MMFusion [68] incorporate noise-based features, yet their performance differs significantly. This suggests that not all handcrafted features are equally effective, and the design of the feature extractor plays a critical role in mitigating overfitting. Therefore, careful selection and design of input features is essential for building more generalizable forensic models.

Task	Method	Seen (FLUX)	Unseen (SD)	Unseen (SDXL)	Seen (SD)	Unseen (SDXL)	Unseen (FLUX)
Localization	MVSS	0.9264 / 0.8893	**0.0026** / **0.0015**	0.0164 / 0.0129	0.9298 / 0.8953	**0.5581** / **0.5135**	**0.0852** / **0.0577**
	PSCC	0.9254 / 0.8883	0.0009 / 0.0006	0.0031 / 0.0022	0.9048 / 0.8684	0.0167 / 0.0108	0.0120 / 0.0074
	Mesorch_P	0.9616 / 0.9362	0.0006 / 0.0005	**0.0366** / **0.0339**	**0.9609** / **0.9372**	0.0644 / 0.0512	0.0210 / 0.0133
	AniXplore	0.9625 / 0.9375	0.0002 / 0.0001	0.0197 / 0.0175	0.8940 / 0.8535	0.1767 / 0.1584	0.0315 / 0.0206
	AniXplore (HR)	**0.9774** / **0.9603**	0.0000 / 0.0000	0.0005 / 0.0004	0.9244 / 0.8841	0.1722 / 0.1198	0.0329 / 0.0213
Detection	MVSS	0.5460 / 0.3755	0.1139 / 0.0604	0.1204 / 0.0641	0.9984 / 0.9967	0.9442 / 0.8944	0.8811 / 0.7875
	PSCC	0.5038 / 0.3367	0.0082 / 0.0041	0.0101 / 0.0051	0.9984 / 0.9969	0.6824 / 0.5179	0.6702 / 0.5039
	AniXplore	**1.0000** / **1.0000**	**1.0000** / **1.0000**	**1.0000** / **1.0000**	**1.0000** / **1.0000**	**1.0000** / **1.0000**	**1.0000** / **1.0000**
	AniXplore (HR)	**1.0000** / **1.0000**	**1.0000** / **1.0000**	**1.0000** / **1.0000**	**1.0000** / **1.0000**	**1.0000** / **1.0000**	**1.0000** / **1.0000**

Table 5: Cross-dataset evaluation results. Left columns show performance when trained on FLUX; right columns show performance when trained on SD. Metrics are F1 / IoU for localization and F1 / Accuracy for detection. HR denotes high resolution version.

Model	Image-level	
	F1	Accuracy
MVSS [4]	0.9991	0.9983
PSCC [42]	0.1233	0.0657
TruFor [15]	0.3908	0.2428
MMFusion [68]	1.0000	1.0000
AniXplore	**1.0000**	**1.0000**

Table 6: Detection results on images collected from Civitai. Each models are trained on AnimeDL-2M.

Comparison with SOTA. AniXplore integrates both DWT and DCT as frequency-domain auxiliary features. This design enhances the model's sensitivity to subtle traces and provides highly discriminative representations. AniXplore achieves outstanding generalization in the detection task, obtaining perfect F1-score and accuracy in all sub datasets. These results demonstrate that AniXplore can reliably identify fake images across a wide variety of generation models, which further confirms the robustness, versatility, and strong deployment potential of the proposed approach.

5.5 Ablation Study

Instead of examining designs that have been extensively validated in prior work, such as multiview feature maps or initializing the backbone with pretrained weights, we focus on evaluating the validity of three main components in AniXplore. All experiments are conducted on the AnimeDL-2M dataset, with input images resized to 512×512. The results are presented in Table 7.

Contribution of frequency features. We observe a obvious performance improvement after introducing frequency-domain features, which demonstrates that our Mixed Feature Extractor effectively enhances the model's perceptual ability.

Feature fusion across latent levels. We compare different strategies for feature fusion, including 1) Late Concat (LC): concat feature maps from all layers in two branches at once, 2) Progressive Concat (PC): fuse the feature map from each layer in two branches, and concat all fused feature maps, 3) Multiview Fusion (MF): fuse the feature map from each layer in two branches progressively and take fused feature from the last layer, then apply SFPN to obtain multiview feature maps from the fused feature. The results show that fusion between branches could be helpful, but fused feature maps may not always be the best feature map for decoder. It indicates that there may not be universally optimal fusion strategies, and the choice of fusion method should be tailored to the specific model architecture and task.

Freq.	Fuse scheme	Task & Loss	Pixel-level		Image-level	
			F1	IoU	F1	Acc.
–	LC	Loc.	0.9456	0.9165	–	–
✓	LC	Loc.	0.9697	0.9485	–	–
✓	LC	Loc. + Cls.	0.9633	0.9394	1.000	1.000
✓	PC	Loc.	0.9670	0.9447	–	–
✓	MF	Loc. + Cls. + AW	**0.9710**	**0.9506**	**1.000**	**1.000**

Table 7: Ablation study on module-wise configurations.

Trade-off in multi-task learning. We found that directly introducing in classification head could cause a slight drop in localization performance, and the convergence of the overall loss becomes slower, suggesting a potential optimization conflict between the classification and localization objectives. Therefore, special measurements such as auto weight (AW) for loss should be taken into account when designing the loss function to achieve an optimal balance.

6 CONCLUSION

We present AnimeDL-2M, a large-scale dataset addressing the gap in IMDL datasets for anime images. Distinguished by its multiple generation methods, rich annotations, and high content diversity, AnimeDL-2M establishes a novel and expansive benchmark for anime-oriented IMDL tasks. Based on unique visual characteristics of anime images, we propose AniXplore, a novel framework optimized for IMDL challenges in this domain. Our experiments reveal significant domain gaps between image styles and editing methods. Experimental results also show that AniXplore outperforms existing SOTA methods in both detection and localization tasks on anime images, while exhibiting strong generalization capabilities in detection tasks. We aim to use AnimeDL-2M and AniXplore to foster future innovations in this field.

REFERENCES

[1] Ruyi Bai. 2024. Image manipulation detection and localization using multi-scale contrastive learning. *Applied Soft Computing* 163 (2024), 111914.

[2] Belhassen Bayar and Matthew C Stamm. 2016. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM workshop on information hiding and multimedia security*. 5–10.

[3] Giulia Bertazzini, Chiara Albisani, Daniele Baracchi, Dasara Shullani, and Alessandro Piva. 2024. Beyond the Brush: Fully-automated Crafting of Realistic Inpainted Images. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 1–6.

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

[5] Yuwei Chen, Ming-Ching Chang, and Xin Li. 2024. Leveraging Semantic Segmentation for Image Manipulation Detection and Localization. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 95–101.

[6] Yirui Chen, Xudong Huang, Quan Zhang, Wei Li, Mingjian Zhu, Qiangyu Yan, Simiao Li, Hanting Chen, Hailin Hu, Jie Yang, et al. 2024. GIM: A Million-scale Benchmark for Generative Image Manipulation Detection and Localization. *arXiv preprint arXiv:2406.16531* (2024).

[7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271* (2024).

[8] Civitai Community. 2025. Civitai. https://civitai.com/ Accessed: 2025-04-11.

[9] Davide Cozzolino and Luisa Verdoliva. 2019. Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security* 15 (2019), 144–159.

[10] Danbooru Community. 2025. Danbooru. https://danbooru.donmai.us/ Accessed: 2025-04-11.

[11] Tiago José De Carvalho, Christian Riess, Elli Angelopoulou, Helio Pedrini, and Anderson de Rezende Rocha. 2013. Exposing digital image forgeries by illumination color classification. *IEEE Transactions on Information Forensics and Security* 8, 7 (2013), 1182–1194.

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

[13] David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. 2023. Online detection of ai-generated images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 382–392.

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

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

[16] Kun Guo, Haochen Zhu, and Gang Cao. 2024. Effective image tampering localization via enhanced transformer and co-attention fusion. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4895–4899.

[17] Xiao Guo, Xiaohong Liu, Iacopo Masi, and Xiaoming Liu. 2024. Language-guided hierarchical fine-grained image forgery detection and localization. *International Journal of Computer Vision* (2024), 1–22.

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

[19] Anna Yoo Jeong Ha, Josephine Passananti, Ronik Bhaskar, Shawn Shan, Reid Southen, Haitao Zheng, and Ben Y Zhao. 2024. Organic or diffused: Can we distinguish human art from ai-generated images?. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 4822–4836.

[20] Qixian Hao, Ruyong Ren, Kai Wang, Shaozhang Niu, Jiwei Zhang, and Maosen Wang. 2024. EC-Net: General image tampering localization network based on edge distribution guidance and contrastive learning. *Knowledge-Based Systems* 293 (2024), 111656.

[21] Zhiyuan He, Pin-Yu Chen, and Tsung-Yi Ho. 2024. Rigid: A training-free and model-agnostic framework for robust ai-generated image detection. *arXiv preprint arXiv:2405.20112* (2024).

[22] Yu-Feng Hsu and Shih-Fu Chang. 2006. Detecting image splicing using geometry invariants and camera characteristics consistency. In *2006 IEEE international conference on multimedia and expo*. IEEE, 549–552.

[23] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. 2024. SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model. *arXiv preprint arXiv:2412.04292* (2024).

[24] Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. 2023. Autosplice: A text-prompt manipulated image dataset for media forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 893–903.

[25] Xun Jin, Junwei Tan, et al. 2025. Plagiarism detection of anime character portraits. *Expert Systems with Applications* 261 (2025), 125566.

[26] Dimitrios Karageorgiou, Giorgos Kordopatis-Zilos, and Symeon Papadopoulos. 2024. Fusion transformer with object mask guidance for image forgery analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4345–4355.

[27] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7482–7491.

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

[29] Myung-Joon Kwon, Wonjun Lee, Seung-Hun Nam, Minji Son, and Changick Kim. 2024. SAFIRE: Segment Any Forged Image Region. *arXiv preprint arXiv:2412.08197* (2024).

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

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

[32] Etienne Levecque, Jan Butora, and Patrick Bas. 2024. Dual JPEG Compatibility: a Reliable and Explainable Tool for Image Forensics. *arXiv preprint arXiv:2408.17106* (2024).

[33] Dong Li, Jiaying Zhu, Xueyang Fu, Xun Guo, Yidi Liu, Gang Yang, Jiawei Liu, and Zheng-Jun Zha. 2024. Noise-Assisted Prompt Learning for Image Forgery Detection and Localization. In *European Conference on Computer Vision*. Springer, 18–36.

[34] Shuaibo Li, Wei Ma, Jianwei Guo, Shibiao Xu, Benchong Li, and Xiaopeng Zhang. 2024. Unionformer: Unified-learning transformer with multi-view representation for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12523–12533.

[35] Yuxi Li, Fuyuan Cheng, Wangbo Yu, Guangshuo Wang, Guibo Luo, and Yuesheng Zhu. 2024. AdaIFL: Adaptive Image Forgery Localization via a Dynamic and Importance-Aware Transformer Network. In *European Conference on Computer Vision*. Springer, 477–493.

[36] Jingchun Lian, Lingyu Liu, Yaxiong Wang, Yujiao Wu, Li Zhu, and Zhedong Zheng. 2024. A Large-scale Interpretable Multi-modality Benchmark for Facial Image Forgery Localization. *arXiv preprint arXiv:2412.19685* (2024).

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 740–755.

[38] Jiawei Liu, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. 2024. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. *arXiv preprint arXiv:2410.10238* (2024).

[39] Weihuang Liu, Xiaodong Cun, and Chi-Man Pun. 2024. DH-GAN: Image manipulation localization via a dual homology-aware generative adversarial network. *Pattern Recognition* 155 (2024), 110658.

[40] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. 2024. Forgeryttt: Zero-shot image manipulation localization with test-time training. *arXiv preprint arXiv:2410.04032* (2024).

[41] Wenxi Liu, Hao Zhang, Xinyang Lin, Qing Zhang, Qi Li, Xiaoxiang Liu, and Ying Cao. 2024. Attentive and contrastive image manipulation localization with boundary guidance. *IEEE Transactions on Information Forensics and Security* (2024).

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

[43] Xuntao Liu, Yuzhou Yang, Haoyue Wang, Qichao Ying, Zhenxing Qian, Xinpeng Zhang, and Sheng Li. 2024. Multi-view Feature Extraction via Tunable Prompts is Enough for Image Manipulation Localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 9999–10007.

[44] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11976–11986.

[45] Zijie Lou, Gang Cao, Kun Guo, Shaowei Weng, and Lifang Yu. 2024. Image Forgery Localization with State Space Models. *arXiv preprint arXiv:2412.11214* (2024).

[46] Zijie Lou, Gang Cao, Kun Guo, Lifang Yu, and Shaowei Weng. 2025. Exploring multi-view pixel contrast for general and robust image forgery localization. *IEEE Transactions on Information Forensics and Security* (2025).

[47] Xiaochen Ma, Bo Du, Zhuohang Jiang, Ahmed Y Al Hammadi, and Jizhe Zhou. 2023. IML-ViT: Benchmarking Image Manipulation Localization by Vision Transformer. *arXiv preprint arXiv:2307.14863* (2023).

[48] Xiaochen Ma, Xuekang Zhu, Lei Su, Bo Du, Zhuohang Jiang, Bingkui Tong, Zeyu Lei, Xinyu Yang, Chi-Man Pun, Jiancheng Lv, et al. 2024. Imdl-benco: A comprehensive benchmark and codebase for image manipulation detection & localization. *Advances in Neural Information Processing Systems* 37 (2024), 134591–134613.

[49] Gaël MAHFOUDI, Badr TAJINI, Florent RETRAINT, Frédéric MORAIN-NICOLIER, Jean Luc DUGELAY, and Marc PIC. 2019. DEFACTO: Image and Face Manipulation Dataset. In *2019 27th European Signal Processing Conference (EUSIPCO)*. 1–5. https://doi.org/10.23919/EUSIPCO.2019.8903181

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

[51] Nikkei Asia. 2024. AI Anime Flood. https://asia.nikkei.com/static/vdata/infographics/ai-anime/ Accessed: 2025-04-11.

[52] Nikkei Asia. 2024. *NIKKEI Film: Japanese anime vs. generative AI.* https://asia.nikkei.com/Business/Technology/NIKKEI-Film-Japanese-anime-vs.-generative-AI Accessed: 2025-04-12.

[53] Yakun Niu, Pei Chen, Lei Zhang, Lei Tan, and Yingjian Chen. 2024. Image Forgery Localization via Guided Noise and Multi-Scale Feature Aggregation. *arXiv preprint arXiv:2412.01622* (2024).

[54] Adam Novozamsky, Babak Mahdian, and Stanislav Saic. 2020. IMD2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops*. 71–80.

[55] Wenyan Pan, Zhihua Xia, Wentao Ma, Yuwei Wang, Lichuan Gu, Guolong Shi, and Shan Zhao. 2024. Auto-focus tracing: Image manipulation detection with artifact graph contrastive. *Knowledge-Based Systems* 304 (2024), 112545.

[56] Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. 2024. Omni-IML: Towards Unified Image Manipulation Localization. *arXiv preprint arXiv:2411.14823* (2024).

[57] Chenfan Qu, Yiwu Zhong, Chongyu Liu, Guitao Xu, Dezhi Peng, Fengjun Guo, and Lianwen Jin. 2024. Towards modern image manipulation localization: A large-scale dataset and novel methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10781–10790.

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

[59] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159* (2024).

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

[61] Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. 2024. Detecting and grounding multi-modal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[62] Ziqi Sheng, Wei Lu, Xiangyang Luo, Jiantao Zhou, and Xiaochun Cao. 2024. SUMI-IFL: An Information-Theoretic Framework for Image Forgery Localization with Sufficiency and Minimality Constraints. *arXiv preprint arXiv:2412.09981* (2024).

[63] Lei Su, Xiaochen Ma, Xuekang Zhu, Chaoqun Niu, Zeyu Lei, and Ji-Zhe Zhou. 2024. Can We Get Rid of Handcrafted Feature Extractors? SparseViT: Nonsemantics-Centered, Parameter-Efficient Image Manipulation Localization Through Spare-Coding Transformer. *arXiv preprint arXiv:2412.14598* (2024).

[64] Yang Su, Shunquan Tan, and Jiwu Huang. 2024. A Novel Universal Image Forensics Localization Model Based on Image Noise and Segment Anything Model. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*. 149–158.

[65] Zhihao Sun, Haipeng Fang, Juan Cao, Xinying Zhao, and Danding Wang. 2024. Rethinking Image Editing Detection in the Era of Generative AI Revolution. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 3538–3547.

[66] Zhihao Sun, Haoran Jiang, Haoran Chen, Yixin Cao, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. 2024. ForgerySleuth: Empowering Multimodal Large Language Models for Image Manipulation Detection. *arXiv preprint arXiv:2411.19466* (2024).

[67] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 2024. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 28130–28139.

[68] Konstantinos Triaridis and Vasileios Mezaris. 2024. Exploring multi-modal fusion for image manipulation detection and localization. In *International conference on multimedia modeling*. Springer, 198–211.

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

[70] Xudong Wang, Yuezun Li, Huiyu Zhou, Jiaran Zhou, and Junyu Dong. 2024. HRGR: Enhancing Image Manipulation Detection via Hierarchical Region-aware Graph Reasoning. *arXiv preprint arXiv:2410.21861* (2024).

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

[72] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems* 34 (2021), 12077–12090.

[73] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. 2024. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761* (2024).

[74] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2024. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435* (2024).

[75] Tianyun Yang, Juan Cao, Danding Wang, and Chang Xu. 2023. Model Synthesis for Zero-Shot Model Attribution. *arXiv preprint arXiv:2307.15977* (2023).

[76] Ye Yao, Tingfeng Han, Shan Jia, and Siwei Lyu. 2025. Dense Feature Interaction Network for Image Inpainting Localization. *IEEE Transactions on Information Forensics and Security* (2025).

[77] Zeqin Yu, Jiangqun Ni, Yuzhen Lin, Haoyi Deng, and Bin Li. 2024. Diffforensics: Leveraging diffusion prior to image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12765–12774.

[78] Kunlun Zeng, Ri Cheng, Weimin Tan, and Bo Yan. 2024. MGQFormer: Mask-guided query-based transformer for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 6944–6952.

[79] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. 2024. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8018–8027.

[80] Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. 2024. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11964–11974.

[81] Zhenfei Zhang, Mingyang Li, and Ming-Ching Chang. 2024. A new benchmark and model for challenging image manipulation detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7405–7413.

[82] Zhenfei Zhang, Mingyang Li, and Ming-Ching Chang. 2024. A new benchmark and model for challenging image manipulation detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7405–7413.

[83] Zhenfei Zhang, Mingyang Li, Xin Li, Ming-Ching Chang, and Jun-Wei Hsieh. 2024. Image Manipulation Detection with Implicit Neural Representation and Limited Supervision. In *European Conference on Computer Vision*. Springer, 255–273.

[84] Jizhe Zhou, Xiaochen Ma, Xia Du, Ahmed Y Alhammadi, and Wentao Feng. 2023. Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 22346–22356.

[85] Jiaying Zhu, Dong Li, Xueyang Fu, Gang Yang, Jie Huang, Aiping Liu, and Zheng-Jun Zha. 2024. Learning discriminative noise guidance for image forgery detection and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7739–7747.

[86] Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Jizhe Zhou. 2024. Mesoscopic Insights: Orchestrating Multi-scale & Hybrid Architecture for Image Manipulation Localization. *arXiv preprint arXiv:2412.13753* (2024).