In the appendix, we provide a survey of face forgery technologies and face forgery detection techniques (1), a comprehensive statistical analysis of the DeepFaceGen dataset (2), and detailed descriptions of prompt construction (3). We further include evaluation setting details (4), information on the detail extraction module (5), and generalization ability verification experiments for different methods (6). Additionally, we present fine-grained analyses of forgery detection features (7) and fine-grained attribute statistical analyses for different forgery techniques (8). Then, we provide detailed discussions on challenges and future directions (9) and potential negative social impacts (10). In addition to the above, we include a summary of key findings and discussed topics (11) and detailed numerical results for experiments (12) .

1 Survey of Face Forgery Technology and Face Forgery Detection Technology

In this section, we present a comprehensive overview of both face forgery technologies and face forgery detection technologies. Regarding the former, we categorize face forgery methods into task-oriented and prompt-guided generation techniques based on their image/video generation approach. Subsequently, we discuss the forgery detection techniques designed specifically for these two types of forgery methods.

1.1 Task-oriented Based Face Forgery Technology

Task-oriented based face forgery involves modifying specific facial features, such as expressions and movements. Traditional facial Photoshop (PS) techniques, which involve manual image manipulation, also fall within this scope. However, traditional PS techniques often leave detectable traces that can be identified by the naked eye. Therefore, survey of task-oriented based face forgery focus on advanced deepfake methods including face swapping, face reenactment, and face alteration.

Face Swapping. Face swapping involves transferring the facial identity from a source image to a target image while preserving the expressions, movements, and background of the target image. Early face swapping techniques primarily relied on autoencoders. One such tool, Deepfake [16], popularized by Reddit users, trains the facial images of the source and target persons separately, allowing the decoder to accurately reproduce their faces. In face swapping, the encoder extracts the source person's facial features and inserts them into the target person's image using the decoder. [48] introduces FaceswapGAN, which employs a face swapping attention mechanism to enhance image realism. This method also addresses occlusion issues using segmentation masks. RSGAN [32] is designed for face swapping using two autoencoders to represent the hair and face regions. It replaces the face's latent representation and reconstructs the image, effectively addressing issues such as mismatched face orientation and lighting. [34] introduces FSGAN, which uses RNN-based methods to transfer expressions and movements from the target face to the source face. FSGAN demonstrates good generalization and requires fewer training samples. [25] introduces Faceshifter, a two-stage face-swapping method. It uses adaptive attention denormalization (AAD) for feature integration and employs a heuristic error acknowledgment refinement network (HEAR-Net) to address occlusion issues. [9] introduces an identity injection module to eliminate identity constraints, and enhances the loss function with weak feature matching loss to improve face synthesis quality.

Face Reenactment. Face reenactment preserves the target image's facial identity while replicating expressions, facial orientation, and body movements from the source image.. [60] introduces Imaginator, which uses a spatiotemporal feature fusion mechanism to decode continuous video from spatial features and motion. They employ two discriminators: one to evaluate the realism of facial appearances and the other to assess the realism of motions. [50] introduces Monkey-Net, which separates appearance and motion information in images, enabling motion-driven animation. Monkey-Net includes a motion transfer network, an unsupervised keypoint detector, and a motion prediction network. It predicts the visual flow map for each keypoint by distinguishing keypoints in target and source images, thereby generating forged images. [51] improves on Monkey-Net by introducing local affine transformations around keypoints, which better reproduce large pose variations. [39] uses action unit annotations combined with unsupervised training and attention mechanisms to enhance model robustness. [56] uses action units to represent facial expressions, processing the face and background separately to improve image quality and reduce identity information leakage. CycleGAN [75] is widely used in face reenactment due to its flexible training capabilities between source and target domains. [65] proposes a full-image reenactment method based on CycleGAN, which uses various receptive field specifications and PatchGAN to enhance image quality. [3] uses CycleGAN for data-driven, unsupervised video retargeting, effectively transferring continuous information for expression-driven animation. [64] introduces ReenactGAN, which extracts facial contours using an encoder and maps them via CycleGAN. A pix2pix generator then reconstructs the image. This method uses only feedforward neural networks, enabling real-time expression reenactment.

Face Alteration. Face alteration modifies specific attributes like hair color, gender, and glasses without altering facial identity. Most face alteration techniques use GAN structures. The StyleGAN series [21–23] are notable for editing facial features, while StarGAN [10] and StarGANV2 [11] enable transformations across multiple image domains, offering better scalability. Another notable method is GANnotation [45], which contains a triple continuity loss function for GAN-based face alteration and a direct facial expression alteration synthesis method. [24] introduces a CAM consistency loss function based on CycleGAN's cycle consistency loss function, which helps retain feature-independent positional information and can be applied to models like StarGAN. To address scalability and diversity issues in face alteration, [27] introduces hierarchical style disentanglement(HiSD), a hierarchical model that represents facial features as labels and attributes. Using an unsupervised approach, HiSD decouples these features, allowing for more precise modifications of target attributes.

1.2 Prompt-guided Generation Based Face Forgery Technology

Based on the differences in network architecture, prompt-guided generation face forgery techniques can be categorized into gan-based models, autoregressive-based models, and diffusion-based models.

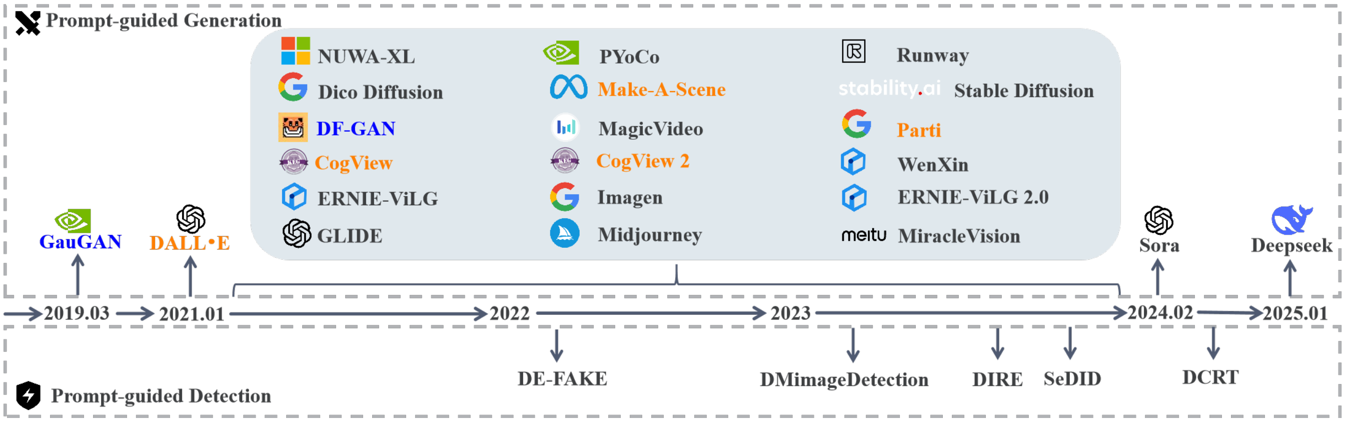

Figure 1: Prompt-guided generation methods/products (above the timeline) and forgery detection techniques (below the timeline) are shown on a chronological timeline. GAN, Autoregressive, and Diffusion are marked with blue, orange, and black fonts, respectively.

GAN-based Models. Based on their model structure, GANs can be classified into single-stage generation networks and stacked architectures. DF-GAN [55], a single-stage generation network, uses one generator, one discriminator, and a pre-trained text encoder. It maps text to images by incorporating affine transformations, enabling direct image synthesis from textual descriptions. GoGAN [30], a stacked architecture, generates higher resolution images in stages. Each branch's generator captures the image distribution, while the discriminator assesses authenticity, refining image resolution and achieving stable training results. Despite their capabilities, GANs face stability issues and mode collapse. These limitations have led to their gradual replacement by autoregressive and diffusion models, which offer improved stability and better handling of diverse data distributions.

Autoregressive-based Models. Autoregressive-based models generate images by modeling spatial relationships between pixels and high-level attributes using an Encoder-Decoder architecture with a multi-head self-attention mechanism. In Text2Image generation, these models convert text and images into token sequences. The autoregressive model predicts image sequences from these tokens, which are then decoded into final images using techniques such as Variational Autoencoders (VAEs) to enhance image quality. Autoregressive models offer explicit density modeling and stable training compared to GANs. Notable examples include DALL·E [36], which generates creative images from text prompts, CogView [14], known for its high-quality image synthesis, and Make-A-Scene [17], which enables interactive image generation. However, autoregressive models face limitations in computational resources, data requirements, and training time due to their large number of parameters. Diffusion models, which offer improved efficiency and require less data, have led to a decline in interest in autoregressive models.

Diffusion-based Models. Diffusion-based models have become the state-of-the-art in deep generative models, surpassing previous image and video synthesis techniques. Diffusion models generate images and videos by combining noise prediction models with conditional diffusion or classifier guidance. This process allows the diffusion model to create the desired output based on the provided guidance. These models excel at handling various input conditions and mitigating mode collapse, making them dominant in fields such as Text2Image, Image2Image, Text2Video, and Image2Video synthesis. Notable examples include GLIDE [33], known for its high-quality Text2Image generation; Imagen [44], which excels in photorealistic image synthesis; Sora [37], a state-of-the-art Text2Video model; and Stable Diffusion [42], which is widely used for its versatility and stability.

1.3 Detection Technique for Task-oriented Based Face Forgery

Detection techniques target task-oriented based face forgeries by identifying artifacts left in various feature spaces during the forgery process. These techniques can be categorized into spatial domain-based, frequency domain-based, and temporal domain-based detection technique.

Spatial Domain-based Detection Technique. [70] suggests that the key to distinguishing real from forged faces lies in subtle local details. They propose a texture enhancement module, an attention generation module, and a bi-linear attention pooling module to help the model focus on facial texture details. However, these methods often overfit to specific forgery artifacts, leading to a rapid decline in detection performance when faced with unseen forgery methods. To avoid overfitting, researchers have generated forged faces by applying certain operations to real faces. [26] introduces the FaceX-Ray model, which detects forgery by identifying face fusion boundaries. During training, the model predicts image authenticity and performs pixel-wise classification on the gray scale map of fusion boundaries. This method does not rely on specific forgery artifacts, showing remarkable generalization capabilities in detecting forgeries from unseen methods. [49] argues that forgeries often contain general forgery traces. They propose Self-Blended Images (SBI), synthetic forgeries created by transforming key points within the same face image, which show strong generalization against unknown forgery methods. However, this method performs poorly against prompt-guidede synthesis methods due to its reliance on the self-forgery process. [4] introduces RECCE, combining reconstruction learning and classification to help the model learn compact features of real faces and uncover essential differences between real and fake faces. Some studies have explored the interpretability of deep face forgery detection models. [15] hypothesizes that detection models identify authenticity by discerning information unrelated to facial identity. They use facial identity as an auxiliary label and designed source feature encoders and target encoders for identity recognition tasks.

Frequency Domain-based Detection Technique. Videos and images disseminated across online streaming media often undergo multiple compressions, resulting in low-quality images that obscure forgery artifacts. To address this issue, researchers have explored detection clues in the frequency domain. For instance, [40] finds that forgery artifacts can be effectively extracted in the frequency domain. They design a frequency-aware decomposition module to adaptively capture forgery clues within images. Additionally, they introduce a local frequency information statistics module to gather frequency information from each local region of an image and recombine these statistics into multi-channel feature maps for the frequency domain. Since artifacts appear in different regions of various images, [59] introduces a multi-modal and multi-scale autoregressive model (M2TR) to detect local artifact details at different spatial levels. This model incorporates frequency domain features as auxiliary information, enhancing its capability to detect forgeries in highly compressed images. While frequency domain-based methods show strong forgery detection capabilities in highly compressed images, their performance significantly declines when encountering unknown forgery methods.

Temporal Domain-based Detection Technique. Temporal domain forgery detection focuses on identifying dynamic inconsistencies between video frames over time. [31] proposes a dual-stream branch network. One branch extracts dynamic temporal inconsistencies from consecutive video frames, and the other amplifies artifact details using a Laplacian of Gaussian (LoG) operator. Recognizing the correlation between forgery and anomaly detection tasks, [43] introduces the deep support vector data description (Deep SVDD) loss function to improve the intra-class compactness of real faces and the inter-class distinction between real and forged faces, enhancing the model's generalization capability. [73] finds that setting the temporal convolution kernel size to 1 in 3D convolutional kernels enhances the network's ability to capture temporal inconsistencies in forged videos. However, temporal inconsistencies can be compromised by noise, compression, and other factors, leading to reduced robustness in these methods.

1.4 Detection Technique for Prompt-guided Generation Based Face Forgery

Research achievements in the detection of prompt-guided generation based face forgery are currently limited. Researchers are attempting to break through the mindset of searching for clues specific to task-oriented based face forgery and instead seek the unique fingerprints produced by the prompt-guided generation based face forgery process.

[47] systematically studies the detection and attribution of fake images generated by diffusion models. They compare the results of image-only input and mixed input (images and corresponding text descriptions) to explore the detection and tracing capabilities of CNN classification models. [13] analyzes the frequency domain and model identification capabilities, concluding that diffusion-generated images have unique fingerprints similar to GAN images. [61] find that the diffusion reconstruction effect of fake images is superior to that of real images. They use the difference between the reconstructed image and the original image, called Diffusion Reconstruction Error (DIRE), for binary classification to determine authenticity, showing higher generalization ability. Based on this, [29]and [6] refine the loss construction of DIRE. However, these methods are tested on small, self-created datasets, and their experimental conclusions lack generality. Additionally, they do not specifically focus on detecting face forgeries. Currently, the detection of faces generated by diffusion models remains relatively unexplored.

2 DeepFaceGen Detailed Statistical Data

In order to construct a robust and extensive benchmark for the detection of face forgery, we carefully consider a range of critical factors including the manner of generation, generation framework, content diversity, ethnic fairness, and label richness throughout the benchmark development process. Following this, we provide detailed introduction to the forged face samples and authentic face samples in DeepFaceGen.

Forged Face Samples. The forged face samples of DeepFaceGen consists of 35 types of forgery methods. The number of forged images/videos reaches $350,264/423,548$. For content diversity, we collected $143,579$ forged images and $93,497$ forged videos from [28] and [18]. As shown in Figure **??**, the forged images contain 27 forgery methods, including task-oriented based and prompt-guided based generation. Forged samples between both generation methods are roughly balanced. The task-oriented based samples include face swapping, face reenactment and face alteration. In the prompt-guided based generation, sufficient Text2Image and Image2Image samples are generated according to the input modality. At the video-level, a rough balance is similarly maintained between the samples generated by the 16 forgery methods. In the process of generating forged video/image samples, in order to maintain ethnic fairness, we control the balance of skin color through text prompt in prompt-guided based generation. Task-oriented based samples also fit ethnic fairness by employing SkinToneClassifier [38]. Additionally, we employ YOLO [57] with manual screening to eliminate low-quality data. The detailed forged statistical data can be seen in Table 1.

Table 1: Detailed Statistical Data of DeepFaceGen.

Manner	Subset	Methods	Images	Videos	Labels
Task-oriented	Face Swapping	FaceShifter	10 500	14 387	n-way labels
		FSGAN	10 500	55 205	
		DeepFakes	10 500	6 000	
		BlendFace	10 500	13 491	
		DSS	10 500	2 866	
		SBS	10 500	-	
		MMReplacement	10 500	1 461	
		SimSwap	-	27 786	
	Face Reenactment	Talking Head Video	9 203	28 935	n-way labels
		ATVG-Net	10 500	11 273	
		Motion-cos	-	22 811	
		FOMM	10 235	42 411	
	Face Alteration	StyleGAN2	10 263	-	n-way labels
		MaskGAN	8 613	-	
		StarGAN2	10 500	-	
		SC-FEGAN	10 500	-	
		DiscoFaceGAN	10 500	-	
Prompt-guided	Text2Image	OJ	28 203	-	n-way labels prompt labels
		SD1	25 677	-	
		SD2	20 898	-	
		SDXL	22 839	-	
		Wenxin	9 989	-	
		Midjourney	9 784	-	
		DF-GAN	40 320	-	
		DALL·E	8 000	-	
		DALL·E 3	2 000	-	
	Text2Video	AnimateDiff	-	40 320	n-way labels prompt labels
		AnimateLCM	-	35 642	
		Hotshot	-	40 320	
		Zeroscope	-	40 320	
		MagicTime	-	40 320	
	Image2Image	Pix2Pix	9 620	-	n-way labels prompt labels
		SDXLR	9 990	-	
		VD	9 130	-	
Total			**350 264**	**423 548**	

Authentic Face samples. In order to ensure content diversity and ethnic fairness in the authentic face samples used in DeepFaceGen, we obtained real samples from reputable sources including [28], [18], [7], and [72]. Specifically, we collected 482 and 463, 101 real images from [28] and [18], and 19, 942, 590, 99, 630, 193, 245 real videos from [72], [28], [18], and [7]. The final collection consists of 463, 583 images and 313, 407 videos, encompassing diverse ages, genders, skin tones, expressions, hair styles, hair colors, backgrounds, dressing styles, and glasses.

3 Detailed Descriptions of Prompts Construction

In the design of prompts, we strive to achieve both content diversity and fairness, which are accompanied by a strong emphasis on detailed prompt descriptions. Following this, we designed a complete expressive framework for each prompt sentence based on the face information that humans take into account when describing faces. The prompt sentence framework contains 9 description attributes: ages, genders, skin tones, expressions, hair styles, hair colors, backgrounds, dressing styles, and glasses. Each description attribute contains a detailed scenario situation. By iterating through the combination of 9 attributes, we can generate over 40, 000 prompts. This design ensures data balance across the various text attributes. Then, we use LoRA [19] to fine-tune the selected pretrained model and generate forged samples fine-tuned with deepfake samples. The detailed pipeline of prompts construction is shown in Figure 2.

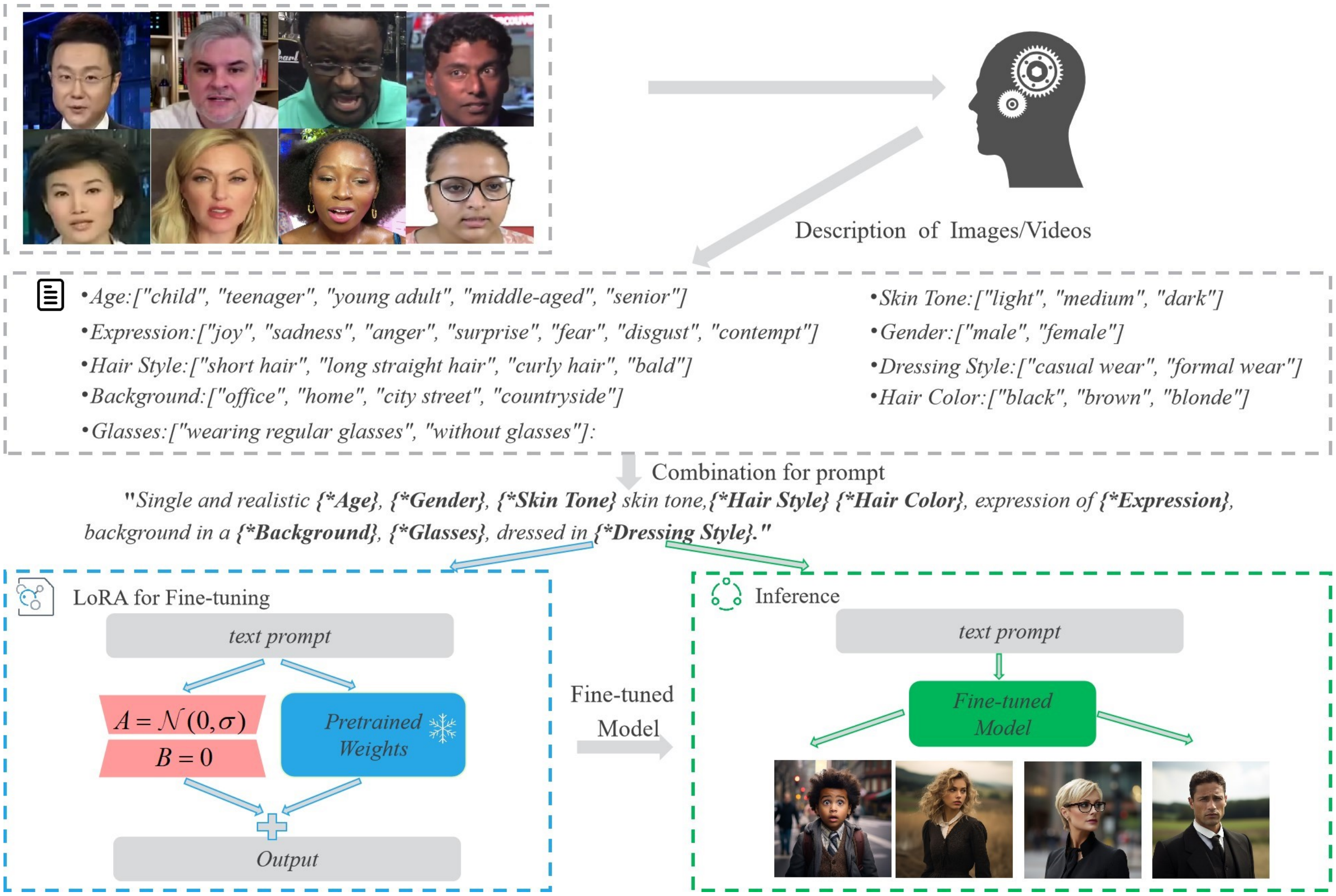
Description of Images/Videos
•Age:["child", "teenager", "young adult", "middle-aged", "senior"]
•Expression:["joy", "sadness", "anger", "surprise", "fear", "disgust", "contempt"]
•Hair Style:["short hair", "long straight hair", "curly hair", "bald"]
•Background:["office", "home", "city street", "countryside"]
•Glasses:["wearing regular glasses", "without glasses"]:
•Skin Tone:["light", "medium", "dark"]
•Gender:["male", "female"]
•Dressing Style:["casual wear", "formal wear"]
•Hair Color:["black", "brown", "blonde"]
Combination for prompt
"Single and realistic {*Age}, {*Gender}, {*Skin Tone} skin tone,{*Hair Style} {*Hair Color}, expression of {*Expression}, background in a {*Background}, {*Glasses}, dressed in {*Dressing Style}."
LoRA for Fine-tuning
text prompt
$A = \mathcal{N}(0,\sigma)$
$B = 0$
Pretrained Weights
Output
Fine-tuned Model
Inference
text prompt
Fine-tuned Model

Figure 2: Pipeline of prompts construction. It consists of four parts: the establishment of face description information, the construction of description attributes, the fine-tuning of pre-trained models and the generation of forged samples. After establishing comprehensive attributes to describe face information from images and videos, rich and comprehensive text prompts can be obtained by iterating the combination of description attributes. Then, LoRA [19] is used to fine-tune the generative model to the field of face generation for the final generation task.

4 Evaluation Details

In this section, we provide a detailed introduction to the selected forgery detection methods and disclose the implementation details during the experimental process.

4.1 Forgery Detection Models

Following the basic backone used by the 20 forgery detection methods, we introduce the forgery detection methods in detail.

- **MesoNet** [1] is a face forgery detection algorithm based on mid-level information from image noise. This approach effectively addresses the challenges of diminished image noise and the difficulty of distinguishing forged video frames using high-level semantic features. Its shallow architecture enhances sensitivity to medium and large-scale features, thereby improving the capability of detecting facial characteristics.
- **Xception** [12] is a convolutional neural network architecture entirely based on depthwise separable convolution layers, simplifies the decoupling of channel correlation and spatial correlation to derive depthwise separable convolutions. This enables efficient extraction of complex features from images and video frames.
- **EfficientNet-B0** [54] is the baseline network of the EfficientNet family, which is developed by leveraging a multi-objective neural architecture search based on mobile inverted bottleneck MBConv [46] with squeeze-and-excitation optimization [20] added to it.
- **F3-Net** [41] utilizes two complementary frequency-aware cues: frequency-aware decomposed image components and local frequency statistics. These cues are deeply explored through a dual-stream collaborative learning framework to detect subtle forgery patterns.

- **RECCE** [5] is a reconstruction and classification learning framework designed to learn common characteristics of real faces by reconstructing face images. It trains a reconstruction network using real face images and employs the latent features of this network to classify real and forged faces. Due to the inconsistency in data distribution between real and forged faces, the reconstruction errors for forged faces and can accurately highlight the forged regions.
- **DNADet** [69] adopts pre-training on image transformation classification and patchwise contrastive learning to capture globally consistent features that are invariant to semantics. It can focus on architecture-related traces and strengthen the global consistency of extracted features.
- **FreqNet** [52] is a lightweight frequency space learning network designed for generalizable forgery image detection. This approach leverages the power of frequency domain learning, providing an adaptable solution for the challenging problem of deepfake detection across diverse sources and GAN models. The methodology includes practical and compact frequency learning plugin modules that integrate with CNN classifiers to enable them to operate effectively within the frequency domain.
- **CViT** [63] is a model composed of two main components: Feature Learning (FL) and the Vision Transformer (ViT). The FL component, a stack of convolutional operations without a fully connected layer, extracts features from face images. These features are then processed by the ViT, which converts them into a sequence of image pixels for detection.
- **SLADD** [8] aims to generalize well in unseen scenarios. It operates on the principle that a generalizable detector should be sensitive to various types of forgeries. SLADD enriches the diversity of forgeries by synthesizing augmented forgeries using a pool of forgery configurations and enhances sensitivity by training the model to predict these configurations.
- **Exposing** [2] is an information bottleneck-based framework for deepfake detection that aims to extract broader forgery clues. It captures a wide range of forgery clues by extracting multiple non-overlapping local representations and fusing them into a global, semantically rich feature.
- **DIRE** [61] is based on the assumption that images generated by diffusion models can be approximately reconstructed through the diffusion process, whereas real images cannot. By applying DDIM's inversion and reconstruction process to the images under inspection, the method differentiates between forged and real samples by analyzing the reconstruction error.
- **DRCT** [6] first obtains reconstructed images for both real and fake images based on the diffusion process. It then leverages contrastive learning loss to train a classifier using the four types of images: real, real-reconstructed, fake, and fake-reconstructed. This approach helps establish a more accurate decision boundary for distinguishing between real and fake samples.
- **UnivFD** [35] analyzes the asymmetry in the decision boundary learned by the CNNSpot classifier. While it effectively distinguishes GAN-generated fake images, the feature space of real images lacks independence—i.e., all non-GAN-generated images (real and diffusion-generated images) are classified into a single category. To improve the generalization ability of the detector and enable it to distinguish real from fake images with a balanced decision boundary, a more appropriate feature space is required. To achieve this, Univdf utilizes the pre-trained CLIP model to extract the feature space.
- **NPR** [53] addresses that gap by rethinking CNN-based generator architectures to develop a generalized representation of synthetic artifacts. The research reveals that up-sampling operators, beyond generating frequency-based artifacts, introduce generalized forgery artifacts. Specifically, the local pixel interdependence created by up-sampling in GAN and diffusion-generated images is significant. To capture and characterize these artifacts, the concept of Neighboring Pixel Relationships (NPR) is introduced, providing a new method to identify structural anomalies caused by up-sampling operations.
- **TALL** [66] transforms video clips into predefined layouts to preserve both spatial and temporal dependencies, enabling effective detection of Deepfake videos. Specifically, consecutive frames are masked at fixed positions within each frame to enhance generalization performance. These frames are then rearranged into a predefined layout, effectively creating a thumbnail that retains the critical temporal and spatial features for deepfake detection.operations.
- **AltFreezing** [62] identifies that spatial artifacts are more prominent than temporal inconsistencies, leading networks to prioritize learning simpler spatial artifacts. This focus limits the model's ability to leverage all forgery features, ultimately weakening its generalization capacity. To address this, the authors divide the network weights into two groups: spatial-related and temporal-related. During training, they alternate freezing between the two sets of weights, enabling the model to learn both spatial and temporal features effectively. Additionally, a video-level data augmentation method is introduced to further enhance the model's generalization ability.
- **LSDA** [68] tackles the generalization issue in deepfake detection by reducing overfitting to forgery-specific artifacts. It expands the forgery space through variations in the latent space, enabling the model to learn a more generalizable decision boundary. This approach enhances domain-specific features and smoothens transitions between different forgery types, improving cross-domain performance.

4.2 Implementation Details

Preproccess. The image and video datasets are divided into training, validation, and test subsets in a ratio approximately 7 : 1 : 2. To ensure fairness in evaluation, each subset maintains a ratio of real to fake instances close to 1 : 1. For video-level evaluations, the video files in the dataset need to be extracted and stored as individual video frames. Given the varying lengths of the video files we collected and generated, we standardize the number of frames extracted from each video to 24. Additionally, since the authors of SLADD [8] did not disclose the

Table 2: The Representative Features of Evaluation Methods

Modality	Method	Technical Route	Identify Objects	Modality	Method	Technical Route	Identify Objects
Image	NPR	Pixel Correlation	Task-oriented, Prompt-guided	Video	Exposing	Information Bottleneck	Task-oriented, Prompt-guided
	UnivFD	Pre-trained Feature Extraction	Task-oriented, Prompt-guided		LSDA	Latent Space Augmentation	Task-oriented, Prompt-guided
	RECCE	Reconstruction Learning	Task-oriented, Prompt-guided		SLADD	Adversarial Learning	Task-oriented, Prompt-guided
	DNADet	Contrastive Learning	Task-oriented, Prompt-guided		AltFreezing	Spatial-Temporal Learning	Task-oriented
	FreqNet	Frequency Learning	Task-oriented, Prompt-guided		TALL	Spatial-Temporal Learning	Task-oriented
	DRCT	Contrastive Learning	Diffusion-based		Xception	Spatial Learning	Task-oriented
	DIRE	Reconstruction Learning	Diffusion-based		EfficientNet	Spatial Learning	Task-oriented
	EfficientNet	Spatial Learning	Task-oriented		CViT	Spatial-Temporal Learning	Task-oriented
	Xception	Spatial Learning	Task-oriented		F3Net	Frequency Learning	Task-oriented
	F3Net	Frequency Learning	Task-oriented		MesoNet	Spatial Learning	Task-oriented

process for creating masks, we adopted the following approach: the mask for real data is set to an all-zero matrix, indicating that there are no forgery regions in the input image. For forged data, we use YOLO [57] to obtain the face bounding box, and then convert the bounding box into a binary mask image, with the forgery region set to 1 and all other areas set to 0.

Training. We all follow the original hyperparameter settings in the evaluation methods. The loss function for SLADD [8] is set to MSE, while the loss functions for MesoNet [1], EfficientNet-B0 [54], Xception [12], F3-Net [41], DNADet [69], RECCE [5], and CViT [63] are set to CrossEntropyLoss. In particular, based on CrossEntropyLoss, Exposing [2] designed the local information loss based on the theoretical analysis of mutual information to ensure the orthogonality and adequacy between local features. The optimizer for all models is Adam with a learning rate of 1×10^{-5}. The batch size is set to 128. All models are pre-trained on ImageNet. All images in the dataset were resized to a fixed resolution of 299×299 pixels and normalized to have pixel values in the range [0, 1].

Inference. We only perform single-crop inference, and directly scale the input face image to the input spatial size of the model.

5 Details on Detail Extraction Module

In this section, we first provide a forward-looking overview of the handling of detailed features within the deepfake detection domain. Following this, we conduct an in-depth analysis through multi-frequency feature analysis, texture feature analysis, and multi-feature fusion experiments. We hope these new insights will offer valuable directions for future research.

As described in Appendix A.3 and A.4, current face forgery detection methods can be categorized into three main types: Spatial Domain-based Detection Techniques, Frequency Domain-based Detection Techniques, and Temporal Domain-based Detection Techniques. Although existing detection methods for prompt-guided generation primarily focus on loss function construction centered around the diffusion process, their core approach still relies on reconstruction error from the input image, placing them within the category of Spatial Domain-based Detection Techniques. Within these three categories, forgery detection methods based on detailed features can be further classified into frequency domain analysis methods [40, 59], texture feature analysis methods [71], pixel correlation analysis methods [53, 67, 74], and pre-trained model feature extraction methods [35].

Given the current state of research, we conduct an in-depth analysis through multi-frequency feature analysis, texture feature analysis, and multi-feature fusion experiments. We hope the conclusions from these experiments will provide valuable foundational knowledge for future research, fostering deeper insights and exploration.

Multi-frequency Feature Analysis. We began by applying the Fourier transform to convert the images from the spatial domain to the frequency domain, allowing us to isolate the low, mid, and high-level frequency components using filters. We then performed an inverse Fourier transform to convert the filtered frequency-domain images back to the spatial domain, enabling us to visualize the effects of the filtering. Finally, we trained and tested the NPR [53], Xception [12], EfficientNet [54],F3Net [41], RECCE [5] and UnivFD [35] using the visualized low, mid, and high-frequency images. By comparing the detection performance across these frequency bands, we assessed their respective roles in face forgery detection.

As shown in the Table 3, utilizing features extracted from different frequency domains as inputs significantly enhances model performance compared to using the original images alone. *Mid-frequency features perform better in detecting Prompt-guided data, while high-frequency features are more effective for Task-oriented data **(Finding 9)**.* This is because Task-oriented methods often introduce subtle texture differences or edge inconsistencies, which high-frequency features are adept at capturing. Although mid-frequency features are less detailed in texture extraction, they excel in identifying artifacts from full-image generation in Prompt-guided data. In contrast, low-frequency features, which capture rough outlines, offer minimal improvement in detection performance when dealing with the high-quality forged data in deepfacegen.

Texture Feature Analysis. In the texture feature analysis experiment, we extracted texture features using both Gabor filters and LBP encoding, visualized these features, and used them as inputs for the Xception model for subsequent causal analysis based on the experimental results. The findings, as shown in Table 4, indicate that *texture features enhance the effectiveness of face forgery detection **(Finding 10)**.* Specifically, Gabor filters, with their sensitivity to image texture features across different orientations and frequencies, are effective at capturing edge and texture variations, making them well-suited for detecting Task-oriented forgery methods. On the other hand, LBP encoding is more inclined to capture global texture patterns, reflecting the overall texture distribution of the image.

Table 3: The ACC of Multi-frequency Feature Analysis. Face Sw., Face Re., Face Al., T2I, and I2I methods are Face Swapping, Face Reenactment, Face Alteration, Text2Image, and Image2Image.

Detection	Detection	Task-oriented			Prompt-guided		Average
Feature	Method	Face Sw.	Face Re.	Face Al.	T2I	I2I	ACC
Original	Xception	65.11	62.95	58.38	73.86	69.87	66.03
	F3net	65.97	63.89	60.01	77.84	73.62	68.26
	EfficientNet	66.78	63.24	59.89	75.32	70.01	67.04
	NPR	79.51	77.32	75.56	84.02	81.65	79.61
	RECCE	76.54	76.01	75.57	82.21	80.09	77.46
	UnivFD	78.41	75.02	74.65	81.56	80.01	77.93
Low-level	Xception	64.98	63.01	59.07	74.11	70.63	66.36
	F3net	66.35	63.51	61.66	78.01	73.66	68.63
	EfficientNet	66.32	64.01	58.45	76.01	71.45	67.24
	NPR	79.66	77.4	74.98	83.99	83.65	79.93
	RECCE	78.54	76.41	75.30	81.45	80.11	78.11
	UnivFD	78.08	75.42	74.37	82.01	80.22	78.02
Mid-level	Xception	67.52	65.01	63.98	79.36	77.01	70.57
	F3net	66.01	66.32	59.01	80.01	79.45	70.16
	EfficientNet	65.45	63.78	61.01	79.65	75.78	69.13
	NPR	80.54	78.01	74.57	84.21	85.01	80.46
	RECCE	78.78	75.41	75.21	84.41	84.11	79.18
	UnivFD	78.77	74.98	74.77	83.78	82.09	78.87
High-level	Xception	69.54	68.44	67.43	75.01	72.39	70.56
	F3net	70.36	71.11	69.01	77.01	74.31	72.36
	EfficientNet	69.42	68.01	65.77	74.87	71.03	69.82
	NPR	80.77	78.64	75.01	83.71	83.87	80.40
	RECCE	79.78	75.69	76.45	82.41	80.11	79.01
	UnivFD	78.89	75.48	75.21	82.99	81.07	78.73

Multi-feature Fusion. Based on the findings from the Multi-frequency Feature Analysis and Texture Feature Analysis, we explored the potential benefits of Multi-feature Fusion to further enhance detection performance. Specifically, we selected features that demonstrated significant advantages in handling specific categories of data in the previous analyses. We then conducted experiments by concatenating these features for further analysis. The results, as shown in Table 5, indicate that the combination of Gabor Filter and High Frequency features yielded the best performance.

6 Details for Cross-generalization Ability Verification Experiments

In this section, we employ 20 forgery detection methods to evaluate the cross-generalization capabilities among sub-datasets. The forgery detection methods are first trained on the subsets that exhibited the best generalization performance in the broad capability evaluation experiments of different forgery techniques discussed in the main text (FaceShifter subset at the image level and DSS subset at the video level). Subsequently, the generalization performance is tested across various subsets. As shown in Figure **??** and Figure **??**, models with detail extraction modules, such as Exposing [2], FreqNet [52] and RECCE [5], achieve higher evaluation metrics for identifying editing forged data. During the generalization test from task-oriented forgery to prompt-guided generation forgery, it is easier to detect data generated by DF-GAN. Additionally, when using task-oriented forgery images/videos as training data, the internal generalization ability of video forgery detection models is significantly lower than that of image forgery detection models. The detailed experimental results can be viewed in Table 9 and 7.

7 Fine-grained Analysis of Forgery Detection Feature

As shown in Figure 3, we conduct a fine-grained visual analysis of forgery detection features. Based on Figure 3 (a), it is evident that the forgery features of GAN-based model are significantly different from those of Diffusion-based and Autoregressive-based models. In Figure 3 (b), the forgery feature distributions are similar when using text and image as input modalities. Additionally, Figures 3 (c) and (d) demonstrate that *the forgery features of task-oriented techniques do not show significant differences between images and videos* ***(Finding 11)***.

Table 4: The ACC of Texture Feature Analysis. Face Sw., Face Re., Face Al., T2I, and I2I methods are Face Swapping, Face Reenactment, Face Alteration, Text2Image, and Image2Image.

Detection	Detection	Task-oriented			Prompt-guided		Average
Feature	Method	Face Sw.	Face Re.	Face Al.	T2I	I2I	ACC
Original	Xception	65.11	62.95	58.38	73.86	69.87	66.03
	F3net	65.97	63.89	60.01	77.84	73.62	68.26
	EfficientNet	66.78	63.24	59.89	75.32	70.01	67.04
	NPR	79.51	77.32	75.56	84.02	81.65	79.61
	RECCE	76.54	76.01	75.57	82.21	80.09	77.46
	UnivFD	78.41	75.02	74.65	81.56	80.01	77.93
LBP	Xception	67.63	64.07	58.64	76.01	73.98	68.06
	F3net	67.45	65.01	62.03	77.78	74.01	69.25
	EfficientNet	66.81	64.01	60.54	76.42	71.77	67.91
	NPR	79.53	77.39	75.98	84.56	82.01	79.89
	RECCE	76.54	76.22	76.51	84.21	81.03	78.41
	UnivFD	78.99	75.64	74.89	81.67	78.57	77.95
Gabor	Xception	68.72	66.39	62.84	75.63	72.47	69.21
	F3net	68.54	64.55	61.89	78.04	74.63	69.53
	EfficientNet	67.05	64.47	61.47	75.99	72.35	68.26
	NPR	80.45	78.98	77.56	84.13	81.55	80.53
	RECCE	76.54	75.78	76.97	83.62	80.87	78.01
	UnivFD	79.45	76.11	76.56	82.56	80.98	79.13

8 Fine-grained Attribute Statistic Analysis for Different Forgery Techniques

In this section, we train all forgery detection models using the training samples obtained from DeepFaceGen. Subsequently, we utilize the fine-grained labels provided by DeepFaceGen to conduct a detailed analysis of the detection patterns of the forgery detection techniques across 9 attributes.

Age Attribute. The age attribute significantly impacts the effectiveness of forgery detection models. Figures 4 (a) and 5 (a) indicate that forgery detection models face more challenges with detecting forgery samples of children, while it is easier to detect forgery data of elderly faces. This difference is due to the unique facial characteristics of children and the elderly. Children's facial features are finer and smoother, lacking prominent wrinkles and details, which makes it easier for forgery techniques to generate realistic child faces, thereby increasing the difficulty of detection. In contrast, elderly individuals often have more pronounced and complex facial features, including wrinkles, age spots, and sagging skin, which make forgery more challenging and, therefore, more likely to be detected by the model.

Skin Tone Attribute. The effectiveness of forgery detection models varies with different skin tones. Figures 4 (b) and 5 (b) show that these models have greater difficulty in accurately detecting forgeries in individuals with darker skin tones compared to those with lighter skin tones. This highlights a racial bias inherent in the forgery detection techniques. The potential cause of this bias could be linked to variations in skin tones and the influence of lighting conditions. Individuals with darker skin tones may have facial features that are harder to capture in forgery detection. Darker skin tones can result in lower contrast in facial details, such as shadows and highlights, making it difficult for forgery detection models to identify forgery artifacts. Conversely, the facial features of individuals with lighter skin tones are generally easier to capture in images. Lighter skin tones make facial details, such as wrinkles and subtle expressions, more visible and typically maintain better facial detail contrast under various lighting conditions.

Hair Style Attribute. The variety of people's hairstyles also has an impact on the effectiveness of forgery detection. As shown in Figures 4 (c) and 5 (c), detecting forgeries with the curly hair attribute is more difficult, while detecting those with the bald attribute is easier. In video-level experiments, the detection performance is relatively consistent across different attributes. We infer that curly hair, with its highly complex and irregular structure, contains rich details between strands. This complexity poses a greater challenge for forgery techniques in generating curly hair, making it easier to leave behind subtle artifacts that are difficult to detect. Consequently, detection models struggle to differentiate these subtle differences, increasing the difficulty of detecting forgeries with curly hair. In contrast, forgery techniques tend to produce more consistent results when generating bald heads due to the lack of complex hair structures, making it easier for detection models to identify forgery artifacts. Additionally, in video-level experiments, the continuity and motion information assist the forgery detection models in capturing forgery artifacts more effectively, leading to more balanced detection performance across different hair style attributes.

Hair Color Attribute. Figure 4 (d) and Figure 5 (d) show that forgery detection models perform relatively evenly when detecting forged data with the attributes of brown hair, blonde hair, and black hair. This can be attributed to similar details and contrast under lighting

Table 5: The ACC of Multi-feature Fusion. Face Sw., Face Re., Face Al., T2I, and I2I methods are Face Swapping, Face Reenactment, Face Alteration, Text2Image, and Image2Image.

Texture	Frequency	Detection	Task-oriented			Prompt-guided		Average
Feature	Level	Method	Face Sw.	Face Re.	Face Al.	T2I	I2I	ACC
LBP	Mid	Xception	66.47	67.14	62.78	81.65	80.49	71.70
		F3net	67.02	64.85	61.03	80.01	81.45	70.87
		EfficientNet	67.05	63.55	59.99	78.45	75.01	68.81
		NPR	80.57	78.26	75.27	85.29	85.16	80.91
		RECCE	79.52	76.14	74.68	84.23	84.65	80.03
		UnivFD	78.84	75.01	74.62	84.36	83.69	79.30
LBP	High	Xception	69.47	69.77	65.01	75.64	76.01	71.16
		F3net	68.79	66.45	64.35	78.45	74.36	70.48
		EfficientNet	69.05	68.45	65.34	76.21	70.15	69.84
		NPR	80.63	78.98	74.62	84.01	84.97	80.64
		RECCE	78.01	77.32	77.03	83.01	84.88	80.17
		UnivFD	78.79	75.01	75.43	83.76	82.54	79.10
Gabor	High	Xception	71.65	73.87	70.32	74.34	75.42	73.12
		F3net	68.88	66.45	63.54	78.63	75.42	70.98
		EfficientNet	69.01	70.45	65.43	76.36	71.53	70.55
		NPR	81.02	79.69	76.49	86.26	86.63	82.01
		RECCE	79.61	78.32	77.40	87.01	84.34	81.64
		UnivFD	79.25	76.15	75.46	85.99	84.63	80.29
Gabor	Mid	Xception	68.41	67.52	63.41	79.87	74.89	70.82
		F3net	67.54	65.48	62.01	80.36	81.56	69.99
		EfficientNet	68.51	65.01	60.36	79.56	76.45	69.97
		NPR	80.12	78.63	74.23	83.13	84.26	80.07
		RECCE	79.41	76.02	77.39	83.11	84.34	80.44
		UnivFD	79.65	76.05	76.48	82.69	82.01	79.37

conditions. When generating forged images, forgery techniques typically handle similar textures and lighting effects for all three hair colors. This similarity results in detection models not having significant difficulty differences in identifying these forgeries.

Expression Attribute. People's inner emotions can be externalized into different expressions. Based on (e) in Figure 4 and Figure 5, it is apparent that forgery detection models perform well when detecting forged images with the anger and surprise attributes. This may result from the facial expressions of anger and surprise attributes. They contain rich details and features that are easier to extract and recognize in image processing. Tense facial muscles and deep wrinkles are typical features of anger, while an open mouth and raised eyebrows are clear indicators of surprise. Forgery detection models can use these prominent features to enhance detection accuracy.

Background Attribute. The background in images/videos also influences the performance of forgery detection models. Figures 4 (f) and 5 (f) indicate that forgery detection models find it easier to detect forged images with the countryside attribute and harder to detect those with the home attribute. Background complexity may be a direct factor. Countryside backgrounds generally have lower complexity, featuring large natural landscapes such as fields, trees, and skies. These elements are relatively simple and have fewer variations, making it easier for forgery techniques to generate these backgrounds without introducing complex artifacts. Consequently, detection models can more easily identify forged elements in these simple backgrounds. By contrast, home backgrounds typically include many details and complex objects such as furniture, appliances, and decorations. Detection models need to process more details and variations, making it harder to detect forgeries.

Gender Attribute. The accuracy of forgery detection models is often lower for female samples ((g) in Figure 4 and 5). Similar to children in age attribute, female facial features are generally finer and smoother, lacking prominent wrinkles and rough skin texture. These fine features may make it harder for detection models to capture forgery artifacts. Additionally, women tend to wear makeup in greater numbers than men. Cosmetics can enhance or conceal certain facial features, and introduce artificial details such as eyeliner and lipstick. These changes can also make it more challenging for forgery detection models to distinguish between real and forged images, as the makeup may mask subtle forgery artifacts that the model relies on for detection.

Glasses Attribute. Based on Figure 4 (h) and Figure 5 (h), forgery detection models perform similarly when detecting forged data with and without the glasses attribute. This can be attributed to glasses' simple and fixed geometric features (such as frames and lenses). When

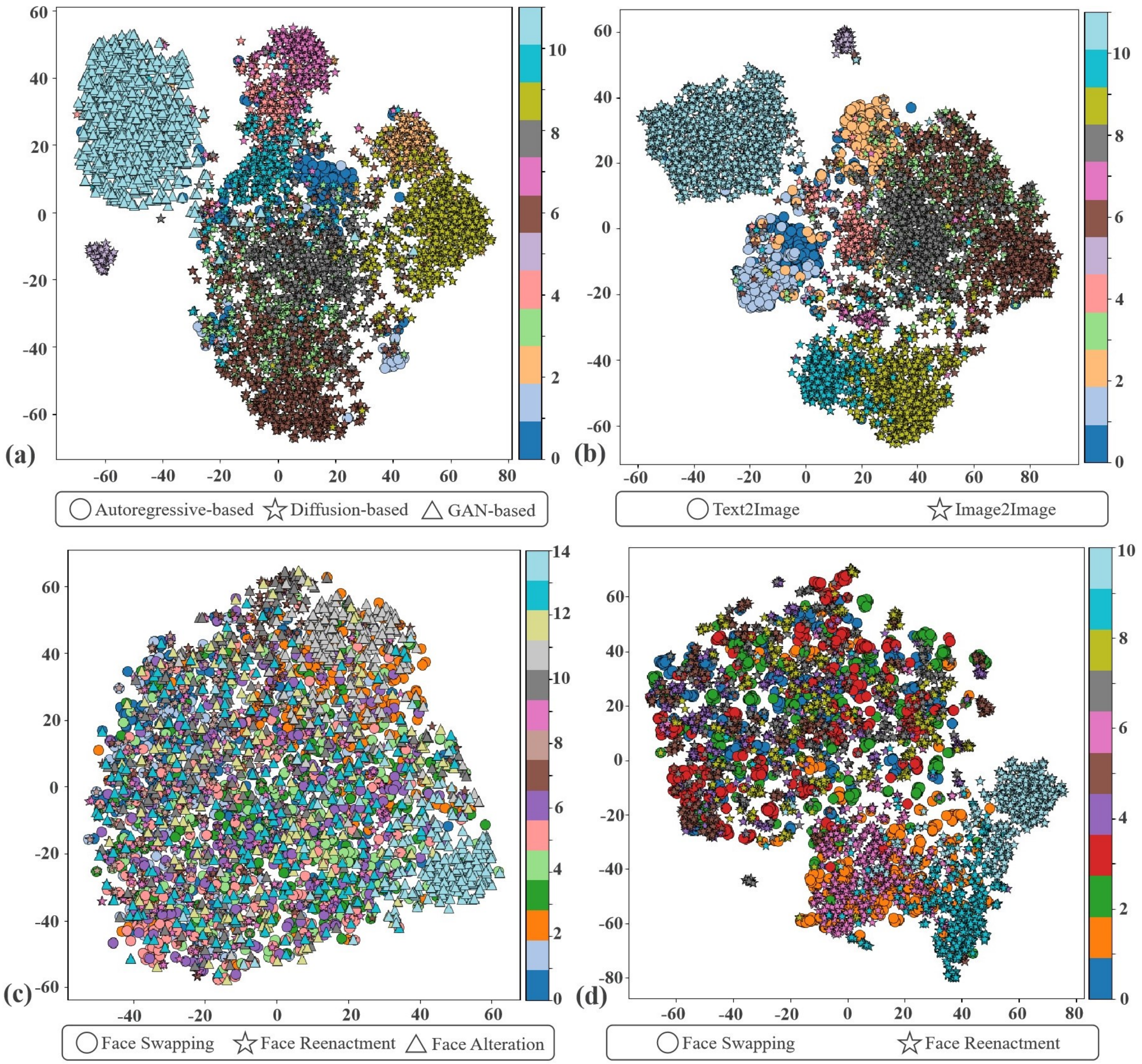

Figure 3: The forgery feature visualization for different forgery techniques on image-level (a-c) and video-level (d) datasets with t-SNE [58]. (a) different generation frameworks, (b) different input modalities, (c) and (d) different generation manners.

generating faces with glasses, forgery techniques can maintain the stability of these geometric features well, resulting in forged images of similar quality to those without glasses.

Dressing Style Attribute. It can be found from Figure 4 (i) and Figure 5 (i) that forgery detection models perform similarly when detecting forged data with the casual wear attribute and the formal wear attribute. This may due to their similar complexity. Although casual and formal wear differ in style, the complexity of details in both types of clothing is relatively similar. Formal wear may include more details (such as ties and buttons), but these details do not significantly affect the quality of forged images. Casual wear may have more varied styles, but its complexity is comparable to formal wear.

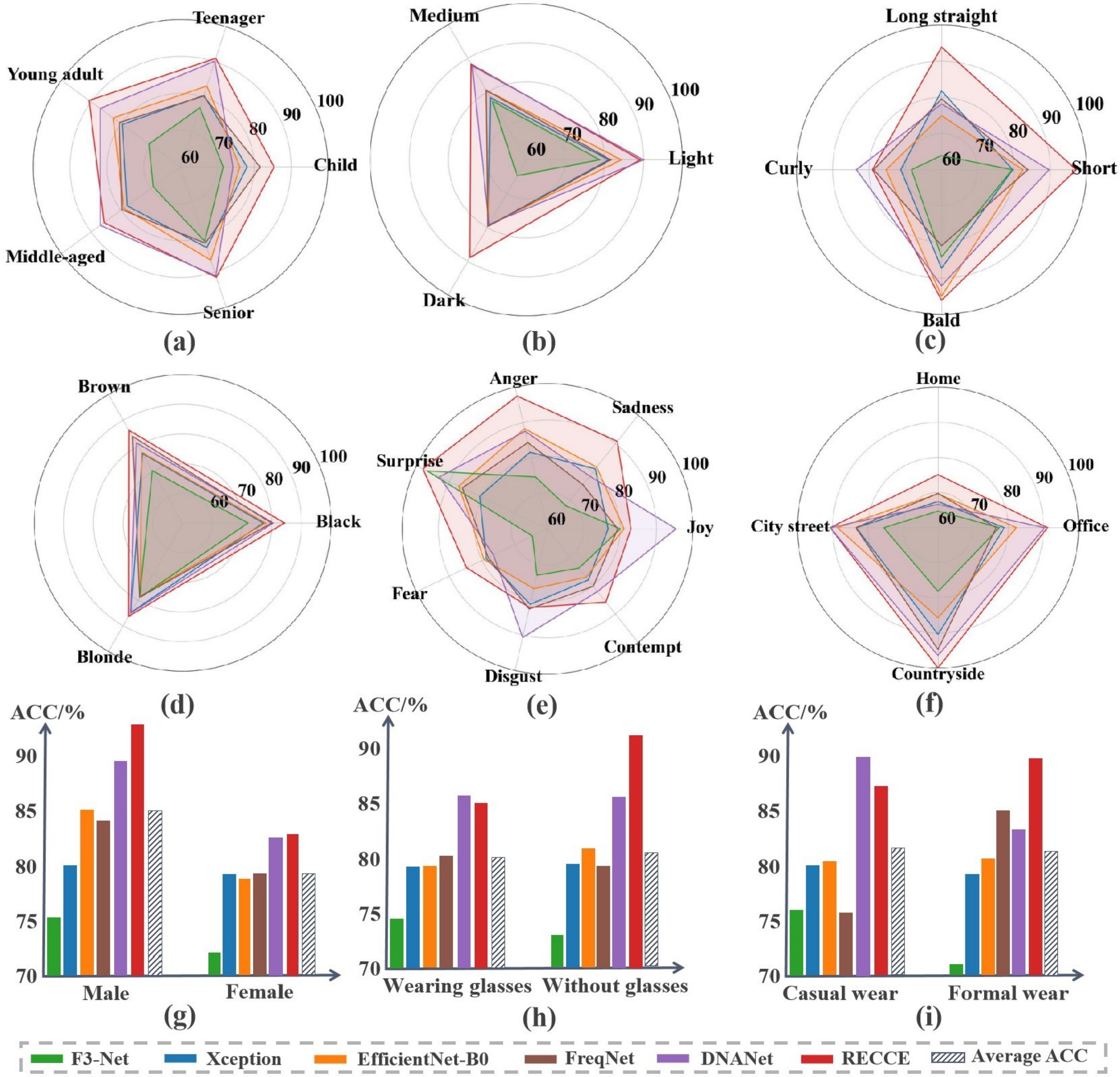

Figure 4: Comparative evaluation of various forgery detection techniques on image-level samples from different attribute perspectives, including (a) age attribute, (b) skin tone attribute, (c) hair style attribute,(d) hair color attribute, (e) expression attribute, (f) background attribute, (g) gender attribute, (h) glasses attribute, and (i) dressing style attribute.

9 Challenges and Future Work

In light of the rapid advancements in face generation techniques, the progress of face forgery detection techniques has significantly lagged behind. Extensive experimentation and analysis reveal several deficiencies in the current forgery detection methods, including inadequate identification accuracy, limited generalization capabilities, and restricted scope for detecting various types of forgery. This section provides a comprehensive overview of the existing challenges in face forgery detection and offers potential valuable directions for future research.

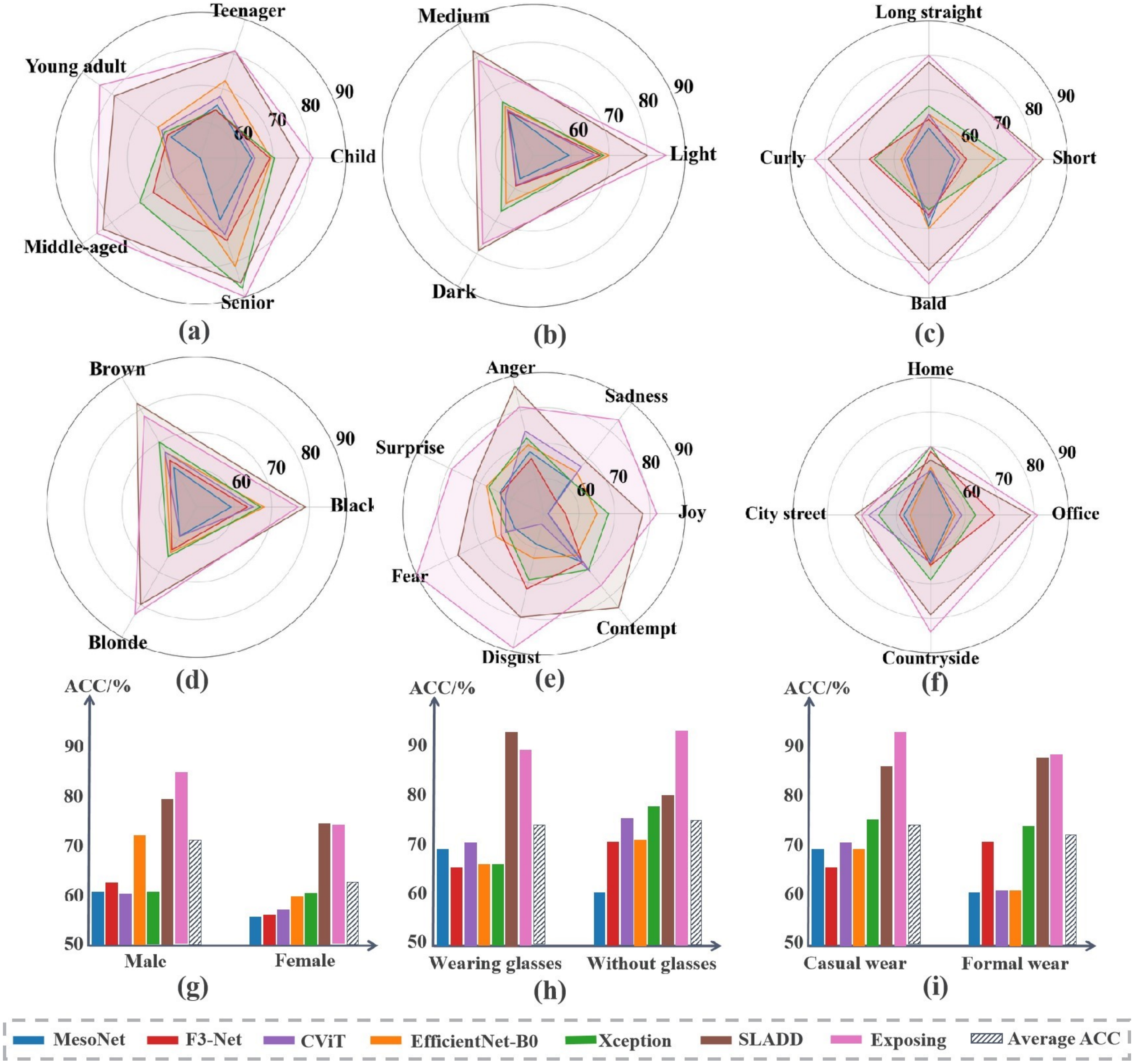

Figure 5: Comparative evaluation of various forgery detection techniques on video-level samples from different attribute perspectives, including (a) age attribute, (b) skin tone attribute, (c) hair style attribute,(d) hair color attribute, (e) expression attribute, (f) background attribute, (g) gender attribute, (h) glasses attribute, and (i) dressing style attribute.

9.1 Challenges

- **Difficulty in Handling Complex Scenarios.** The diversity of complex scenarios increases the difficulty of face forgery detection tasks. Real-world face forgery detection can be affected by environmental factors such as changes in lighting conditions, which can alter shadows and highlights on the face, making it appear darker or brighter. Changes in camera angles can distort facial shapes and features, making the face look twisted or misaligned. Additionally, variations in background complexity can blur the edges of the face or blend it with the background, making it appear unclear or disproportionate. These factors can impact the authenticity and reliability of detection results, increasing the difficulty of recognizing and detecting forgeries.

- **Poor Generalization Performance.** Although current detection models perform well on individual face forgery datasets, their generalization across different datasets remains inadequate. In real-world scenarios, the type of face forgery method used is often unknown, making it difficult to determine the specific type of forgery. Therefore, using pre-trained face forgery detection models for real-world tasks may result in unreliable detection outcomes.
- **Oversimplified Forgery Detection Tasks.** Current face forgery detection tasks focus primarily on binary classification of whether the content is forged, which is relatively crude. In real-world scenarios, there is often a need for tracing the source of the forgery, which is crucial for determining responsibility and uncovering the truth. In face video forgery tasks, attackers often target only a few video frames or audio segments to alter the video content. However, forgery detection models that focus on video-level forgery detection can easily overlook the characteristics of forged segments, significantly increasing the likelihood of detection errors.

9.2 Future Work

- **Objective Quantification of Evaluation Benchmarks.** With the increasingly complex and realistic content forgery scenarios brought about by the development of AIGC technologies, current evaluation benchmarks rely on specific model performance metrics, which can be limiting. In real-world scenarios, designing evaluation benchmarks that can accurately quantify the multi-angle forgery detection capabilities and even the adaptability of models is a crucial direction for future exploration.
- **Dynamic Updating of Benchmark Data.** When designing evaluation benchmarks, it is essential to consider the existence of diverse face forgery types. Regularly updating benchmark datasets to include the latest forgery techniques can help the benchmarks stay close to the complex real-world scenarios. Integrating user feedback data can provide new ideas for dynamically updating benchmark datasets. Additionally, as deep forgery technologies continue to evolve, establishing a dynamic labeling mechanism to address new deep forgery techniques and generative models is becoming increasingly important.
- **Building General Forgery Detection Scenarios.** Although we have constructed a general face deep forgery detection dataset that includes both task-oriented based and prompt-guided generation based face forgery techniques, incorporating both image and video modalities, the audio aspect remains a gap. Furthermore, given the relatively unexplored state of detecting face forgeries generated by diffusion methods, designing general forgery detection techniques based on the inherent differences between real and forged videos, as well as the local feature similarities and model inference paths, is a critical issue that needs to be addressed in the coming years.
- **Emphasis on Robustness of Forgery Detection Models.** The robustness of forgery detection models is key to maintaining stability and reliability in real-world scenarios with complex and variable content. Introducing adversarial samples during training and testing can enhance the robustness of models. However, while adding noise and adversarial samples can improve robustness to some extent, it can also lead to a loss in detection performance. Exploring the inherent characteristics of real samples to identify differences between forged and real samples and developing detection methods that can handle any face forgery product while ensuring detection accuracy is a primary research direction for the future.
- **Self-Evolving Forgery Detection Frameworks.** Forgery techniques and forgery detection techniques are mutually aligned and promote each other. Forgery technologies generally advance faster than forgery detection technologies, leading to significant harm from forged face products to human society. Current forgery detection models and methods rely mainly on researchers analyzing the flaws and weaknesses of forgery technologies and designing corresponding solutions. Developing self-evolving frameworks using adversarial learning mechanisms and reinforcement learning models to drive the autonomous evolution of forgery detection models, thereby improving the ability to quickly respond to various forgery products, is a key research direction for the future.

10 Potential Negative Social Impacts

The creation and use of deepfake datasets, while beneficial for advancing technology, can lead to several negative societal impacts:

- **Misuse of Forgery Methods.** In order to restore the complex forgery scenes in the real scene as much as possible, the forgery methods in the data set are realistic. These forgery methods can be misused to create misleading or harmful content, eroding public trust in media and making it difficult to distinguish between real and fake information.
- **Ethical Concerns.** Due to the transparency of the data set, a large number of face samples in the data set may provide fake resources for illegal personnel. Widespread exposure to deepfakes can lead to public skepticism and paranoia about the authenticity of all digital content.

To mitigate these impacts, we are contemplating controlled access for users and are committed to the dynamic evolution of DeepFaceGen to ensure it remains robust against emerging threats.

11 Summary of Key Findings and Discussed Topics

Findings:

(1) **Finding 1:** Prompt-guided forgery samples exhibit clear high-dimensional distribution boundaries with task-oriented forgery samples.
(2) **Finding 2:** Frequency-based methods struggle to capture the forgery fingerprints of prompt-guided data.

(3) **Finding 3:** Noise-based methods primarily detect fingerprints specific to particular forgery frameworks, but these fingerprints lack generalizability across different frameworks.

(4) **Finding 4:** Directly transferring face AIGC detection to non-face AIGC remains a challenging task.

(5) **Finding 5:** Models trained on similar generative frameworks can learn distinguishing features for non-face data.

(6) **Finding 6:** Detection performance shows no significant difference between prompt-guided data generated from text-based and image-based inputs.

(7) **Finding 7:** Mid-frequency features are more effective for detecting prompt-guided data, while high-frequency features perform better for task-oriented data (see *Appendix 5*).

(8) **Finding 8:** Texture features improve the effectiveness of face forgery detection (see *Appendix 5*).

(9) **Finding 9:** Task-oriented forgery features show no significant differences between images and videos (see *Appendix 7*).

(10) **Finding 10:** Unique findings on different facial attributes (see *Appendix 8*).

Topics:

(1) **Leveraging Temporal Information in Image Forgery Detection.** Reconstruction methods in GAN and diffusion treat generative models as auxiliary tools, overlooking the temporal dynamics inherent in the generation process. Exploiting these temporal variations can significantly improve image-based detection performance.

(2) **Prompt-guided Video Forgery Detection.** Prompt-guided forgery data remains largely unexplored in video detection. DeepFaceGen bridges this gap by providing essential data support. Techniques like noise reconstruction and contrastive learning between prompts and outputs from image forgery detection offer promising insights for detecting prompt-guided video forgeries.

(3) **Exploring Graph Structures.**Graph-based methods are underutilized in forgery detection, despite their success in spatiotemporal modeling, satellite navigation, and social network analysis. Human face videos can be represented as dynamic graphs.

(4) **Vision-Language Models (VLMs).** Application of VLMs in face forgery detection remains limited to component-based utilization. Leveraging VLMs (reinforcement learning, provenance attribution, and data distillation) represents a critical direction for future breakthroughs.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (2018), 1–7. https://api.semanticscholar.org/CorpusID:52157475

[2] Zhongjie Ba, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren. 2024. Exposing the Deception: Uncovering More Forgery Clues for Deepfake Detection. arXiv:2403.01786 [cs.CV]

[3] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. 2018. Recycle-GAN: Unsupervised Video Retargeting. In *ECCV*.

[4] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. 2022. End-to-End Reconstruction-Classification Learning for Face Forgery Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4103–4112. doi:10.1109/CVPR52688.2022.00408

[5] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. 2022. End-to-End Reconstruction-Classification Learning for Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4113–4122.

[6] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. 2024. DRCT: Diffusion Reconstruction Contrastive Training towards Universal Detection of Diffusion Generated Images. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. https://openreview.net/forum?id=oRLwyayrh1

[7] Chen Chen, Dong Wang, and Thomas Fang Zheng. 2023. CN-CVS: A Mandarin Audio-Visual Dataset for Large Vocabulary Continuous Visual to Speech Synthesis. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49357.2023.10095796

[8] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. 2022. Self-supervised Learning of Adversarial Examples: Towards Good Generalizations for DeepFake Detections. In *CVPR*.

[9] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. SimSwap: An Efficient Framework For High Fidelity Face Swapping. *Proceedings of the 28th ACM International Conference on Multimedia* (2020). https://api.semanticscholar.org/CorpusID:222278682

[10] Yunjey Choi, Min-Je Choi, Mun Su Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2017. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)* (2017), 8789–8797. https://api.semanticscholar.org/CorpusID:9417016

[11] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2019. StarGAN v2: Diverse Image Synthesis for Multiple Domains. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 8185–8194. https://api.semanticscholar.org/CorpusID:208617800

[12] François Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1800–1807. doi:10.1109/CVPR.2017.195

[13] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2023. On The Detection of Synthetic Images Generated by Diffusion Models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49357.2023.10095167

[14] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. CogView: Mastering Text-to-Image Generation via Transformers. *arXiv preprint arXiv:2105.13290* (2021).

[15] S. Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. 2022. Explaining Deepfake Detection by Analysing Image Matching. In *European Conference on Computer Vision*. https://api.semanticscholar.org/CorpusID:250698762

[16] Faceswap. 2020. Faceswap: Deepfakes software for all. https://github.com/deepfakes/faceswap.

[17] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors. doi:10.48550/ARXIV.2203.13131

[18] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. 2021. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4358–4367. doi:10.1109/CVPR46437.2021.00434

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=nZeVKeeFYf9

[20] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7132–7141. doi:10.1109/CVPR.2018.00745

[21] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-Free Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 852–863. https://proceedings.neurips.cc/paper_files/paper/2021/file/076ccd93ad68be51f23707988e934906-Paper.pdf

[22] Tero Karras, Samuli Laine, and Timo Aila. 2018. A Style-Based Generator Architecture for Generative Adversarial Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 4396–4405. https://api.semanticscholar.org/CorpusID:54482423

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2019. Analyzing and Improving the Image Quality of StyleGAN. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 8107–8116. https://api.semanticscholar.org/CorpusID:209202273

[24] Daejin Kim, Mohammad Azam Khan, and Jaegul Choo. 2021. Not just Compete, but Collaborate: Local Image-to-Image Translation via Cooperative Mask Prediction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6505–6514. doi:10.1109/CVPR46437.2021.00644

[25] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2019. FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping. *ArXiv* abs/1912.13457 (2019). https://api.semanticscholar.org/CorpusID:209515957

[26] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face X-Ray for More General Face Forgery Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5000–5009. doi:10.1109/CVPR42600.2020.00505

[27] Xinyang Li, Shengchuan Zhang, Jie Hu, Liujuan Cao, Xiaopeng Hong, Xudong Mao, Feiyue Huang, Yongjian Wu, and Rongrong Ji. 2021. Image-to-image Translation via Hierarchical Style Disentanglement. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8635–8644. doi:10.1109/CVPR46437.2021.00853

[28] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[29] Ruipeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. 2023. Exposing the Fake: Effective Diffusion-Generated Images Detection. *ArXiv* abs/2307.06272 (2023). https://api.semanticscholar.org/CorpusID:259837077

[30] Musadaq Mansoor, Mohammad Nauman, Hafeez Ur Rehman, and Alfredo Benso. 2022. Gene Ontology GAN (GOGAN): a novel architecture for protein function prediction. *Soft Computing* 26, 16 (August 2022), 7653–7667. doi:10.1007/s00500-021-06707-z

[31] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. 2020. Two-branch Recurrent Network for Isolating Deepfakes in Videos. *ArXiv* abs/2008.03412 (2020). https://api.semanticscholar.org/CorpusID:221090663

[32] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. 2018. RSGAN: face swapping and editing using face and hair representation in latent spaces. *ACM SIGGRAPH 2018 Posters* (2018). https://api.semanticscholar.org/CorpusID:4929075

[33] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 16784–16804. https://proceedings.mlr.press/v162/nichol22a.html

[34] Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. FSGAN: Subject Agnostic Face Swapping and Reenactment. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 7183–7192. doi:10.1109/ICCV.2019.00728

[35] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards Universal Fake Image Detectors that Generalize Across Generative Models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 24480–24489. https://api.semanticscholar.org/CorpusID:257038440

[36] Open AI. 2023. DALL·E. https://openai.com/index/dall-e-3.

[37] Open AI. 2024. Sora. https://openai.com/index/sora.

[38] René Alejandro Rejón Pia and Chenglong Ma. 2023. Classification Algorithm for Skin Color (CASCo): A new tool to measure skin color in social science research. *Social Science Quarterly* 104 (2023), 168.

[39] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. GANimation: Anatomically-Aware Facial Animation from a Single Image. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 835–851.

[40] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII* (<conf-loc content-type="InPerson">Glasgow, United Kingdom</conf-loc>). Springer-Verlag, Berlin, Heidelberg, 86–103. doi:10.1007/978-3-030-58610-2_6

[41] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues. arXiv:2007.09355 [cs.CV]

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[43] Lukas Ruff, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Robert A. Vandermeulen, Alexander Binder, Emmanuel Müller, and M. Kloft. 2018. Deep One-Class Classification. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:49312162

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487 [cs.CV]

[45] Enrique Sanchez and Michel F. Valstar. 2018. Triple consistency loss for pairing distributions in GAN-based face synthesis. *ArXiv* abs/1811.03492 (2018). https://api.semanticscholar.org/CorpusID:53211512

[46] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4510–4520. doi:10.1109/CVPR.2018.00474

[47] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. 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* (<conf-loc>, <city>Copenhagen</city>, <country>Denmark</country>, </conf-loc>) *(CCS '23)*. Association for Computing Machinery, New York, NY, USA, 3418–3432. doi:10.1145/3576915.3616588

[48] Shaoanlu. 2017. Faceswap-GAN. https://github.com/shaoanlu/faceswap-GAN. CP/OL, accessed 2021-10-15.

[49] Kaede Shiohara and Toshihiko Yamasaki. 2022. Detecting Deepfakes with Self-Blended Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18720–18729.

[50] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. Animating Arbitrary Objects via Deep Motion Transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2372–2381. doi:10.1109/CVPR.2019.00248

[51] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

[52] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 2024. Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Learning. arXiv:2403.07240 [cs.CV]

[53] 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 (CVPR)*. 28130–28139.

[54] Mingxing Tan and Quoc V. Le. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946 [cs.LG]

[55] Ming Tao, Hao Tang, Fei Wu, Xiaoyuan Jing, Bing-Kun Bao, and Changsheng Xu. 2022. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16494–16504. doi:10.1109/CVPR52688.2022.01602

[56] Soumya Tripathy, Juho Kannala, and Esa Rahtu. 2020. FACEGAN: Facial Attribute Controllable rEenactment GAN. arXiv:2011.04439 [cs.CV]

[57] ultralytics. 2020. YoloV5. https://github.com/ultralytics/yolov5.

[58] Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605. https://api.semanticscholar.org/CorpusID:5855042

[59] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. 2022. M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval* (Newark, NJ, USA) *(ICMR '22)*. Association for Computing Machinery, New York, NY, USA, 615–623. doi:10.1145/3512527.3531415

[60] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. 2020. ImaGINator: Conditional Spatio-Temporal GAN for Video Generation. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1149–1158. doi:10.1109/WACV45572.2020.9093492

[61] Zhendong Wang, Jianmin Bao, Wen gang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. DIRE for Diffusion-Generated Image Detection. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 22388–22398. https://api.semanticscholar.org/CorpusID:257557819

[62] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. 2023. AltFreezing for More General Video Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4129–4138.

[63] Deressa Wodajo and Solomon Atnafu. 2021. Deepfake Video Detection Using Convolutional Vision Transformer. arXiv:2102.11126 [cs.CV]

[64] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. 2018. ReenactGAN: Learning to Reenact Faces via Boundary Transfer. In *ECCV*.

[65] Runze Xu, Zhiming Zhou, Weinan Zhang, and Yong Yu. 2017. Face Transfer with Generative Adversarial Network. *ArXiv* abs/1710.06090 (2017). https://api.semanticscholar.org/CorpusID:32489585

[66] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. 2023. TALL: Thumbnail Layout for Deepfake Video Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22658–22668.

[67] 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:2406.19435 [cs.CV] https://arxiv.org/abs/2406.19435

[68] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. 2024. Transcending Forgery Specificity with Latent Space Augmentation for Generalizable Deepfake Detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8984–8994. doi:10.1109/CVPR52733.2024.00858

[69] Tianyun Yang, Ziyao Huang, Juan Cao, Lei Li, and Xirong Li. 2022. Deepfake Network Architecture Attribution. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*.

[70] Hanqing Zhao, Tianyi Wei, Wenbo Zhou, Weiming Zhang, Dongdong Chen, and Nenghai Yu. 2021. Multi-attentional Deepfake Detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2185–2194. doi:10.1109/CVPR46437.2021.00222

[71] Hanqing Zhao, Tianyi Wei, Wenbo Zhou, Weiming Zhang, Dongdong Chen, and Nenghai Yu. 2021. Multi-attentional Deepfake Detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2185–2194. doi:10.1109/CVPR46437.2021.00222

[72] Ya Zhao, Rui Xu, and Mingli Song. 2019. A Cascade Sequence-to-Sequence Model for Chinese Mandarin Lip Reading. *ACM* (2019).

[73] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. 2021. Exploring Temporal Coherence for More General Video Face Forgery Detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 15024–15034. https://api.semanticscholar.org/CorpusID:237091271

[74] Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. 2024. PatchCraft: Exploring Texture Patch for Efficient AI-generated Image Detection. arXiv:2311.12397 [cs.CV] https://arxiv.org/abs/2311.12397

[75] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2242–2251. doi:10.1109/ICCV.2017.244

12 Detailed Numerical Results

Table 6: The AUC scores of Evaluation Methods at video-level (D: Detection technique, F: Forgery method).

F D	Modality	Talking Head Video	FSGAN	DeepFakes	BlendFace	DSS	MMReplacement	SimSwap	FaceShifter
MesoNet	Video	0.678	0.713	0.665	0.658	0.644	0.618	0.708	0.642
EfficientNet-B0	Video	0.711	0.723	0.688	0.702	0.589	0.701	0.697	0.699
Xception	Video	0.740	0.731	0.615	0.733	0.598	0.805	0.651	0.856
F3-Net	Video	0.841	0.559	0.691	0.791	0.686	0.735	0.574	0.680
CViT	Video	0.766	0.757	0.616	0.770	0.647	0.711	0.589	0.703
SLADD	Video	0.757	0.871	0.738	0.712	0.716	0.758	0.809	0.716
Exposing	Video	0.798	0.869	0.755	0.746	0.743	0.703	0.870	0.731
TALL	Video	0.701	0.864	0.801	0.784	0.731	0.731	0.801	0.741
AltFreezing	Video	0.712	0.834	0.765	0.793	0.701	0.765	0.831	0.732
LSDA	Video	0.744	0.854	0.777	0.801	0.730	0.801	0.860	0.722

F D	Modality	ATVG-Net	Motion-cos	FOMM	AnimateDiff	AnimateLCM	Hotshot	Zeroscope	MagicTime
MesoNet	Video	0.632	0.729	0.735	0.701	0.652	0.737	0.769	0.699
EfficientNet-B0	Video	0.635	0.701	0.694	0.794	0.841	0.788	0.803	0.825
Xception	Video	0.504	0.857	0.556	0.846	0.836	0.873	0.851	0.825
F3-Net	Video	0.717	0.537	0.560	0.818	0.854	0.872	0.846	0.779
CViT	Video	0.591	0.614	0.688	0.799	0.810	0.804	0.878	0.813
SLADD	Video	0.702	0.893	0.812	0.884	0.901	0.894	0.872	0.921
Exposing	Video	0.749	0.916	0.911	0.952	0.933	0.958	0.932	0.927
TALL	Video	0.701	0.754	0.856	0.806	0.831	0.801	0.822	0.842
AltFreezing	Video	0.735	0.801	0.843	0.861	0.852	0.835	0.831	0.830
LSDA	Video	0.778	0.863	0.878	0.933	0.898	0.932	0.904	0.902

Table 7: The AUC scores of Cross-generalization Ability Verification Experiments at video-level (D: Detection technique, F: Forgery method).

F D	Year	Modality	Talking Head Video	FSGAN	DeepFakes	BlendFace	DSS	MMReplacement
MesoNet	2018	Video	0.423	0.477	0.460	0.411	0.818	0.523
EfficientNet-B0	2019	Video	0.531	0.624	0.589	0.713	0.882	0.563
Xception	2019	Video	0.711	0.594	0.608	0.681	0.893	0.642
F3-Net	2020	Video	0.682	0.742	0.519	0.633	0.873	0.682
CViT	2021	Video	0.773	0.612	0.593	0.580	0.842	0.563
SLADD	2022	Video	0.773	0.643	0.569	0.761	0.875	0.683
AltFreezing	2023	Video	0.685	0.612	0.568	0.716	0.862	0.661
Exposing	2024	Video	0.683	0.613	0.588	0.720	0.916	0.653
TALL	2024	Video	0.692	0.643	0.564	0.710	0.871	0.654
LSDA	2024	Video	0.701	0.621	0.581	0.726	0.901	0.672
F D	Year	Modality	SimSwap	FaceShifter	ATVG-Net	Motion-cos	FOMM	AnimateDiff
MesoNet	2018	Video	0.589	0.489	0.577	0.443	0.582	0.565
EfficientNet-B0	2019	Video	0.621	0.677	0.656	0.558	0.672	0.532
Xception	2019	Video	0.673	0.652	0.663	0.699	0.458	0.794
F3-Net	2020	Video	0.467	0.605	0.693	0.650	0.591	0.688
CViT	2021	Video	0.593	0.736	0.642	0.663	0.716	0.801
SLADD	2022	Video	0.549	0.712	0.751	0.643	0.685	0.769
AltFreezing	2023	Video	0.643	0.701	0.701	0.654	0.601	0.735
Exposing	2024	Video	0.684	0.782	0.653	0.710	0.593	0.854
TALL	2024	Video	0.631	0.687	0.653	0.602	0.583	0.701
LSDA	2024	Video	0.665	0.752	0.711	0.701	0.534	0.795
F D	Year	Modality	AnimateLCM	Hotshot	Zeroscope	MagicTime	-	-
MesoNet	2018	Video	0.639	0.752	0.801	0.781	-	-
EfficientNet-B0	2019	Video	0.704	0.864	0.807	0.801	-	-
Xception	2019	Video	0.763	0.759	0.857	0.821	-	-
F3-Net	2020	Video	0.746	0.656	0.761	0.732	-	-
CViT	2021	Video	0.743	0.804	0.798	0.735	-	-
SLADD	2022	Video	0.784	0.828	0.805	0.823	-	-
AltFreezing	2023	Video	0.794	0.801	0.801	0.813	-	-
Exposing	2024	Video	0.837	0.897	0.853	0.836	-	-
TALL	2024	Video	0.732	0.746	0.787	0.752	-	-
LSDA	2024	Video	0.801	0.845	0.835	0.797	-	-

Table 8: The AUC scores of Evaluation Methods at image-level (D: Detection technique, F: Forgery method).

F D	Modality	MMReplacement	FaceShifter	FSGAN	DeepFakes	BlendFace	SBS	DSS	ATVG-Net	FOMM
Xception	Image	80.09	81.31	81.63	81.54	80.31	80.32	80.71	81.02	78.51
DRCT	Image	81.06	82.01	81.10	83.95	83.44	83.11	84.36	82.33	82.89
F3-Net	Image	77.10	77.78	77.86	78.63	77.24	78.39	77.60	82.33	82.89
UnivFD	Image	85.54	85.87	85.12	85.84	85.86	85.42	86.29	85.27	84.17
RECCE	Image	83.48	85.11	85.70	86.28	84.16	85.77	85.15	85.96	82.87
FreqNet	Image	81.11	81.99	82.37	82.15	81.27	81.75	83.18	82.41	80.63
EfficientNet-B0	Image	79.15	80.45	80.14	76.51	76.55	79.51	80.04	76.15	77.02
DIRE	Image	78.44	80.24	80.01	75.85	77.54	80.64	79.65	77.54	77.65
DNADet	Image	83.01	86.01	83.45	85.01	85.01	83.02	80.21	88.67	80.54
NPR	Image	86.32	89.64	87.74	88.64	88.01	87.61	85.64	87.01	84.69

F D	Modality	Talking Head Video	StarGAN2	StyleGAN2	MaskGAN	SC-FEGAN	DiscoFaceGAN	DALL·E	DALL·E3	Wenxin
Xception	Image	77.43	82.22	78.50	78.51	78.39	77.34	75.15	86.29	86.57
DRCT	Image	81.08	84.18	81.52	80.32	82.06	83.45	81.74	88.41	84.34
F3-Net	Image	75.18	81.08	75.42	73.75	76.24	76.94	84.73	87.23	89.28
UnivFD	Image	83.29	86.88	84.63	84.07	84.65	84.19	84.25	90.17	93.13
RECCE	Image	81.99	88.04	82.82	83.33	82.10	84.21	86.62	84.90	92.60
FreqNet	Image	79.51	83.96	80.13	80.78	79.86	80.77	84.25	87.13	91.98
EfficientNet-B0	Image	79.81	79.12	78.04	80.14	76.51	71.12	87.15	88.34	93.24
DIRE	Image	75.01	79.54	78.65	78.64	76.77	73.54	89.91	88.05	92.72
DNADet	Image	82.31	83.20	84.01	85.01	82.01	83.64	87.64	89.21	90.01
NPR	Image	87.64	86.05	87.01	89.61	86.63	86.65	91.54	91.41	94.35

F D	Modality	SD1	OJ	SD2	SDXL	DF-GAN	Midjourney	SDXLR	pix2pix	VD
Xception	Image	87.34	89.42	87.49	85.76	95.12	76.71	83.83	83.12	84.72
DRCT	Image	86.83	89.01	85.91	87.65	96.71	79.52	93.46	77.61	83.84
F3-Net	Image	90.95	92.72	89.11	91.43	93.45	81.65	86.40	81.21	89.52
UnivFD	Image	90.83	97.90	96.43	97.75	94.54	88.64	91.46	90.71	96.67
RECCE	Image	93.40	95.04	93.70	93.28	95.22	94.44	90.03	89.52	90.21
FreqNet	Image	90.94	93.08	90.92	92.61	97.11	85.69	89.40	88.66	96.55
EfficientNet-B0	Image	92.85	91.78	90.04	92.04	90.82	87.51	89.11	89.51	98.78
DIRE	Image	93.56	92.45	90.41	91.01	92.54	89.78	89.51	91.51	94.25
DNADet	Image	90.01	88.01	91.45	90.01	95.01	88.67	89.01	89.54	89.68
NPR	Image	92.12	92.28	92.01	91.87	98.88	91.01	90.64	91.30	93.01

Table 9: The AUC scores of Cross-generalization Ability Verification Experiments at image-level (D: Detection technique, F: Forgery method).

F D	Year	Modality	MMReplacement	FaceShifter	FSGAN	DeepFakes	BlendFace	SBS	DSS	ATVG-Net	FOMM
Xception	2019	Image	0.758	0.798	0.785	0.775	0.767	0.764	0.769	0.805	0.752
EfficientNet-B0	2019	Image	0.701	0.751	0.714	0.698	0.700	0.692	0.700	0.713	0.703
F3-Net	2020	Image	0.785	0.894	0.865	0.790	0.806	0.785	0.803	0.866	0.788
RECCE	2022	Image	0.799	0.892	0.855	0.794	0.796	0.788	0.794	0.849	0.794
DNADet	2022	Image	0.769	0.811	0.799	0.783	0.783	0.774	0.782	0.815	0.764
DIRE	2023	Image	0.700	0.743	0.670	0.696	0.710	0.690	0.716	0.716	0.706
UnivFD	2023	Image	0.801	0.886	0.855	0.800	0.826	0.807	0.816	0.886	0.801
FreqNet	2024	Image	0.768	0.882	0.848	0.780	0.771	0.773	0.780	0.857	0.768
DRCT	2024	Image	0.706	0.733	0.709	0.706	0.726	0.695	0.726	0.710	0.712
NPR	2024	Image	0.815	0.871	0.888	0.816	0.825	0.804	0.829	0.848	0.814

F D	Year	Modality	Talking Head Video	StarGAN2	StyleGAN2	MaskGAN	SC-FEGAN	DiscoFaceGAN	DALL·E	DALL·E3	Wenxin
Xception	2019	Image	0.738	0.783	0.744	0.730	0.754	0.720	0.791	0.820	0.748
EfficientNet-B0	2019	Image	0.693	0.679	0.690	0.687	0.705	0.663	0.725	0.805	0.701
F3-Net	2020	Image	0.764	0.794	0.773	0.785	0.792	0.727	0.862	0.764	0.769
RECCE	2022	Image	0.762	0.788	0.774	0.780	0.798	0.730	0.875	0.864	0.803
DNADet	2022	Image	0.750	0.788	0.756	0.745	0.766	0.729	0.791	0.857	0.757
DIRE	2023	Image	0.710	0.686	0.716	0.694	0.704	0.670	0.864	0.846	0.802
UnivFD	2023	Image	0.804	0.806	0.807	0.804	0.791	0.764	0.860	0.854	0.810
FreqNet	2024	Image	0.752	0.771	0.755	0.759	0.771	0.726	0.824	0.836	0.820
DRCT	2024	Image	0.723	0.726	0.700	0.671	0.714	0.650	0.870	0.834	0.795
NPR	2024	Image	0.817	0.825	0.836	0.826	0.816	0.781	0.874	0.850	0.804

F D	Year	Modality	SD1	OJ	SD2	SDXL	DF-GAN	Midjourney	SDXLR	pix2pix	VD
Xception	2019	Image	0.820	0.844	0.842	0.799	0.924	0.806	0.751	0.696	0.709
EfficientNet-B0	2019	Image	0.743	0.761	0.698	0.737	0.697	0.823	0.683	0.558	0.689
F3-Net	2020	Image	0.846	0.852	0.856	0.670	0.930	0.855	0.824	0.758	0.857
RECCE	2022	Image	0.871	0.866	0.870	0.755	0.937	0.793	0.815	0.714	0.862
DNADet	2022	Image	0.832	0.856	0.856	0.809	0.873	0.805	0.766	0.695	0.766
DIRE	2023	Image	0.815	0.906	0.865	0.815	0.803	0.846	0.812	0.756	0.803
UnivFD	2023	Image	0.856	0.830	0.865	0.862	0.834	0.804	0.806	0.765	0.804
FreqNet	2024	Image	0.900	0.927	0.907	0.794	0.976	0.820	0.776	0.738	0.831
DRCT	2024	Image	0.823	0.915	0.915	0.822	0.815	0.850	0.825	0.762	0.825
NPR	2024	Image	0.841	0.868	0.874	0.874	0.824	0.799	0.810	0.749	0.803