C2P-CLIP: Injecting Category Common Prompt in CLIP to Enhance Generalization in Deepfake Detection

Chuangchuang Tan[*1,2], **Renshuai Tao**[*1,2], **Huan Liu**[1,2], **Guanghua Gu**[4,5], **Baoyuan Wu**[6], **Yao Zhao**[1,2,3†], **Yunchao Wei**[1,2,3]

[1]Institute of Information Science, Beijing Jiaotong University
[2]Visual Intellgence +X International Cooperation Joint Laboratory of MOE
[3]Pengcheng Laboratory, Shenzhen, China
[4]School of Information Science and Engineering, Yanshan University
[5]Hebei Key Laboratory of Information Transmission and Signal Processing
[6]School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China
tanchuangchuang@bjtu.edu.cn, yzhao@bjtu.edu.cn

Abstract

This work focuses on AIGC detection to develop universal detectors capable of identifying various types of forgery images. Recent studies have found large pre-trained models, such as CLIP, are effective for generalizable deepfake detection along with linear classifiers. However, two critical issues remain unresolved: 1) understanding why CLIP features are effective on deepfake detection through a linear classifier; and 2) exploring the detection potential of CLIP. In this study, we delve into the underlying mechanisms of CLIP's detection capabilities by decoding its detection features into text and performing word frequency analysis. Our finding indicates that CLIP detects deepfakes by recognizing similar concepts. Building on this insight, we introduce Category Common Prompt CLIP, called C2P-CLIP, which integrates the category common prompt into the text encoder to inject category-related concepts into the image encoder, thereby enhancing detection performance. Our method achieves a 12.4% improvement in detection accuracy compared to the original CLIP.

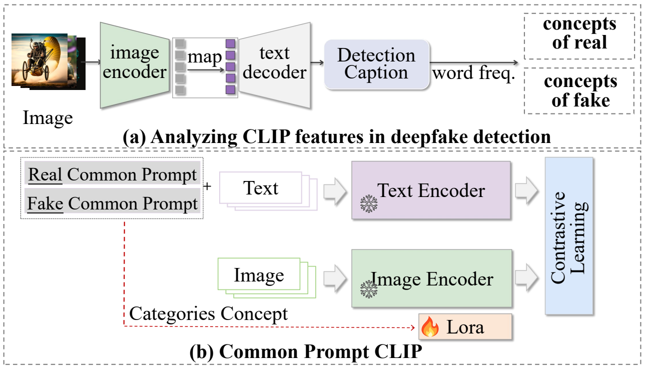


Figure 1: Category Common Prompt CLIP. (a) To investigate the mechanism by which CLIP detects deepfakes, we decode the detection feature into text. Here, the detection features refer to the image features transformed by the linear classifier parameters. Our analysis of these texts reveals that the detection capability arises from the matching of similar concepts. (b) Building on this insight, we propose a method to enhance the generalization capability of image encoders by introducing a category common prompt. This approach injects manually specified category concepts into the image encoder, aiming to improve its detection performance.

Introduction

With the development of various image generation techniques, such as Generative Adversarial Networks (GANs) (Goodfellow et al. 2014; Karras et al. 2018, 2019), diffusion models (Ho et al. 2020; Rombach et al. 2022; Zhang et al. 2023; Yin et al. 2023), etc., distinguishing fake images from real ones has become increasingly challenging for the human eye. As the threshold for image forgery decreases, its misuse will have negative impacts on aspects like the economy and politics. The development of universal systems for detecting deepfake has become imperative.

Recently, several forgery detection methods (Li et al. 2021; Tao et al. 2024) have been developed to identify deepfake images, with a particular focus on detecting face forgery (Yan et al. 2023, 2024). Despite these advancements, existing methods often struggle with unseen deepfake sources, leading to inadequate generalization performance. To address this issue, various approaches (Tan et al. 2023; Ojha et al. 2023; Tan et al. 2024c) have been proposed to enhance generalization ability. Some studies (Tan et al. 2023, 2024c) focus on developing artifact representations that capture low-level forged traces, such as frequency information, gradients, and neighborhood pixel relationships. Other approaches (Qian et al. 2020; Liu et al. 2024; Tan et al. 2024b) aim to design more effective detectors to improve overall detection performance.

Recently, large pre-trained models like CLIP (Radford et al. 2021) have been widely used in various computer vision tasks (Huang et al. 2023; Jiao et al. 2023; Hui et al. 2024). Several studies have extended CLIP's capabilities to-

[*]These authors contributed equally.
[†]Corresponding author

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

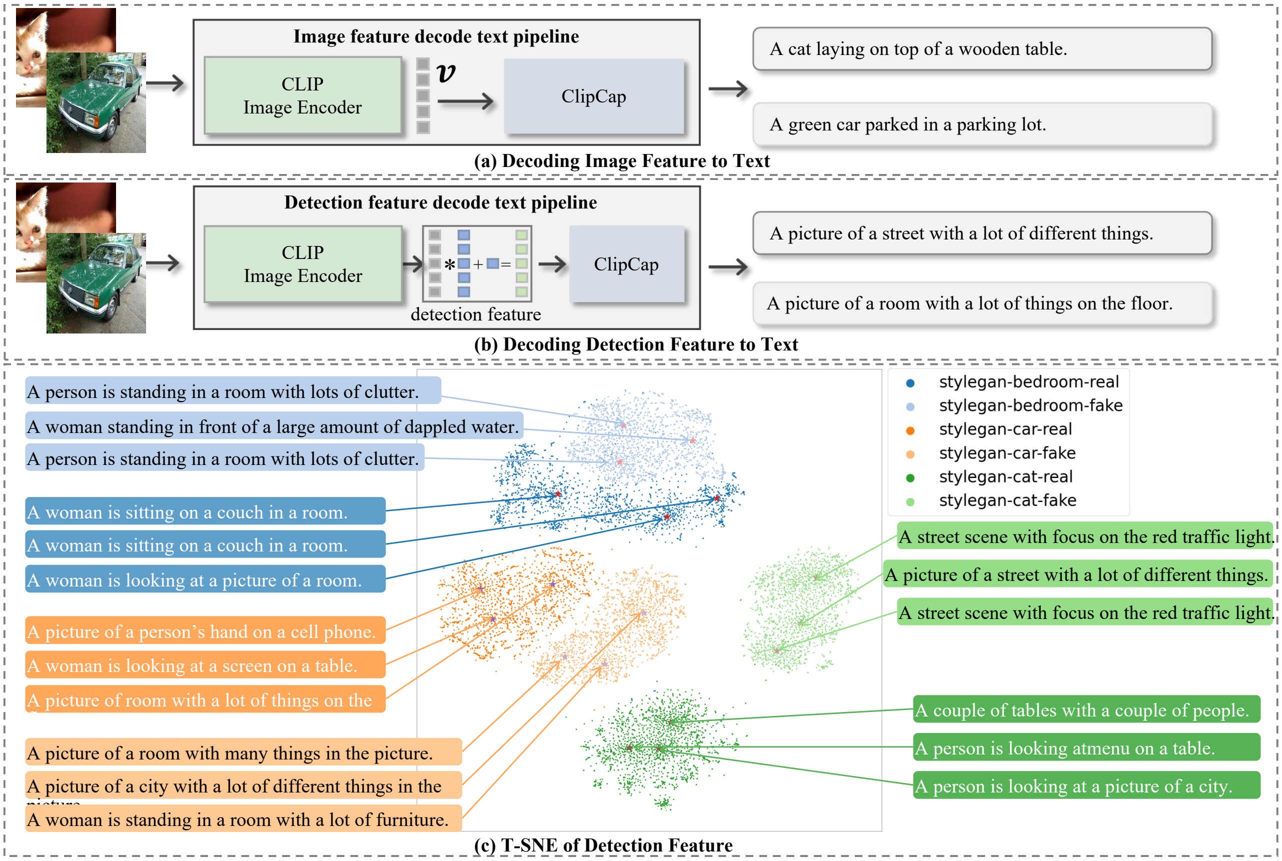


Image feature decode text pipeline
CLIP
Image Encoder
v
ClipCap
A cat laying on top of a wooden table.
A green car parked in a parking lot.
Detection feature decode text pipeline
CLIP
Image Encoder
∗ + =
detection feature
ClipCap
A picture of a street with a lot of different things.
A picture of a room with a lot of things on the floor.
A person is standing in a room with lots of clutter.
A woman standing in front of a large amount of dappled water.
A person is standing in a room with lots of clutter.
A woman is sitting on a couch in a room.
A woman is sitting on a couch in a room.
A woman is looking at a picture of a room.
A picture of a person's hand on a cell phone.
A woman is looking at a screen on a table.
A picture of room with a lot of things on the
A picture of a room with many things in the picture.
A picture of a city with a lot of different things in the
A woman is standing in a room with a lot of furniture.
● stylegan-bedroom-real
● stylegan-bedroom-fake
● stylegan-car-real
● stylegan-car-fake
● stylegan-cat-real
● stylegan-cat-fake
A street scene with focus on the red traffic light.
A picture of a street with a lot of different things.
A street scene with focus on the red traffic light.
A couple of tables with a couple of people.
A person is looking atmenu on a table.
A person is looking at a picture of a city.

(a) Decoding Image Feature to Text

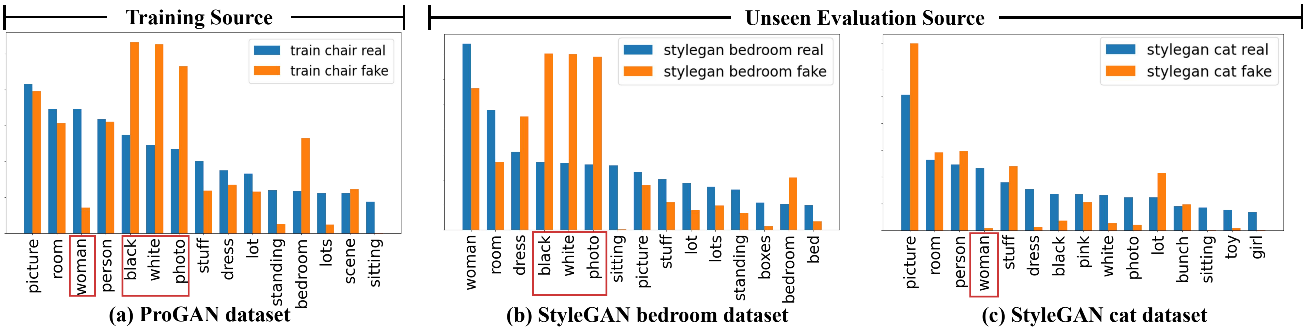
(b) Decoding Detection Feature to Text

(c) T-SNE of Detection Feature

Figure 2: Analyzing CLIP Features in Deepfake Detection. (a) Decoding Image Feature to Text. We employ ClipCap (Mokady, Hertz, and Bermano 2021) to decode the image feature v to text. (b) Decoding Detection Feature to Text. To discern the specific information within the image features that contribute to classification, we decode the detection features into text. The detection features are defined as the combination of image features v and linear classifier fc parameters: $v * fc.weight + fc.bias$. Notice the linear mapping between image features and detection features. Notably, the decoded text bears no direct relevance to the original image content. (c) T-SNE visualization of Detection Feature. We use T-SNE to visualize the detection features from the StyleGAN dataset and decode the textual representations of the three clustering centers within each subset.

wards achieving generalizable deepfake detection. For instance, UniFd (Ojha et al. 2023) utilizes image features extracted by CLIP for linear classification to detect deepfakes. Additionally, FatFormer (Liu et al. 2024) improves detection performance by incorporating frequency analysis and employing the text encoder as an adaptor within the frozen CLIP vision model. Despite these advancements, two critical issues remain unresolved: 1) elucidating why CLIP, trained with contrastive learning, is capable of achieving generalizable deepfake detection through a linear classifier, and 2) exploring the detection potential of CLIP.

Why can CLIP features be used for general deepfake detection with linear classifiers? In this work, we attempt to discuss this question by decoding CLIP features into text using ClipCap, as shown in Figure 2(a)(b). Specifically, in UniFd, a linear classifier denoted as fc is employed on CLIP's image features, represented as v, to detect forged images. Inspired by this, we define the image features v after transformation by linear classifier fc as detection features. To elucidate the mechanism behind CLIP's detection capabilities, we perform a textual decoding of both image and detection features. As depicted in Figure 2(a), a cat image is processed through the CLIP visual model to extract image features, subsequently decoded into text by ClipCap: 'A cat laying on top of a wooden table'. Conversely, decoding detection features into text: 'A picture of a street with a lot of different things'. We further visualize the detection features of the StyleGAN dataset using T-SNE, as depicted in Figure 2(c). We apply k-means clustering to each subset to identify three cluster centers, and the decoded texts for these cluster centers are also shown in Figure 2(c). This observation suggests that semantic information, such as the presence of a cat in the image, is no longer captured in the detection features. This feature does not include the concept of true or false when expressed in natural language. Furthermore, we conduct a word frequency analysis on the text decoded from the detection features of both the training set and the unseen source, with the results illustrated in Figure 3. It can be observed that some words with frequency differences between



(a) ProGAN dataset **(b) StyleGAN bedroom dataset** **(c) StyleGAN cat dataset**

Figure 3: Word Frequency Analysis on Various Sources. We conduct a word frequency analysis on the text decoded from detection features of both the training set (ProGAN) and the unseen test source (StyleGAN). The top 15 words are shown in the graph. The analysis reveals significant differences in word frequencies between the training and test sets. Notably, certain words present in the test set also appear in the training set. For instance, the word 'women' shows substantial frequency variation between (a) and (c). This observation supports the conclusion that CLIP achieves generalizable forgery detection by matching similar concepts or groups of concepts.

real and fake images in the training set also appear in the test set. We infer that CLIP performs forgery detection by identifying and matching similar concepts.

How to Improve the Detection Performance of CLIP? Building on the aforementioned insight, we propose a novel and effective approach named Category Common Prompt CLIP (C2P-CLIP) to enhance CLIP's detection performance. This method utilizes Category Common Prompts to inject category concepts into the image encoder through the text encoder, thereby enhancing its ability to distinguish between real and fake items. Specifically, our approach begins with the generation of text captions using ClipCap. We then assign consistent category-specific text prompts to captions from the same category, referred to as category common prompts. For instance, captions of fake samples are paired with the prompt "Deepfake", while captions of real samples are paired with "Camera". Subsequently, we retrain CLIP using these new image-text pairs, effectively embedding category-common concepts into the image encoder.

To thoroughly evaluate the generalizability of C2P-CLIP, we perform extensive simulations using a comprehensive image database generated by 20 different models [1]. Despite not introducing additional parameters during testing, C2P-CLIP significantly outperforms the original CLIP and achieves state-of-the-art performance.

Related Work

Face Forgery Detection

Face forgery detection has been a prominent area of research due to the rise of face edit and generation. Rossler *et al.*(Rossler et al. 2019) employs the Xception network trained on image data to effectively detect manipulated facial images. In addition, several studies (Masi et al. 2020; Qian et al. 2020) have explored the use of frequency artifacts to improve the robustness of deepfake detection methodologies. To enhance the generalization capability of detection systems, particularly when confronted with unseen sources. A range of methodologies (Wang et al. 2021; Chen et al. 2022; He et al. 2021; Shiohara et al. 2022) have been developed to diversify training data, employing strategies such as adversarial training, image reconstruction, and blending techniques. The UCF (Yan et al. 2023) employs a multi-task learning strategy to extract common forgery features, thereby enhancing the model's generalization capability. Similarly, the LSDA approach (Yan et al. 2024) constructs and simulates variations within and across forgery features in the latent space, expanding the forgery feature space and enabling the learning of a more generalizable decision boundary.

AIGC Detection

With the advancement of generative technologies, the scope of forged content has expanded beyond facial forgeries to encompass a wide range of scenes. Consequently, recent research (Chen et al. 2024b; Niu et al. 2021) has increasingly focused on AIGC detection, which presents unique challenges compared to face forgery detection due to its broader variety of Deepfake types, demanding higher generalization capabilities. In this context, CNN-Spot (Wang et al. 2020) leverages data augmentation techniques to enhance generalization in detection tasks. Several approaches have been proposed to capture low-level artifact representations in AIGC detection, including frequency-based features (Jeong et al. 2022), gradients (Tan et al. 2023), and neighboring pixel relationships (Tan et al. 2024c), random-mapping feature (Tan et al. 2024a). In addition to low-level features, large pre-trained models have been employed to capture high-level forging traces for AIGC detection (Chen et al. 2024a). UniFD (Ojha et al. 2023), for example, directly utilizes image features from the CLIP model for linear classification. FatFormer (Liu et al. 2024) integrates frequency analysis with a text encoder as an adapter to the frozen CLIP vision model, thereby enhancing detection performance.

[1] ProGAN, StyleGAN, BigGAN, CycleGAN, StarGAN, GauGAN, Deepfake, SITD, SAN, CRN, IMLE, Guided, LDM, Glide, DALL·E, Midjourney, SDv1.4, SDv1.5, Wukong, VQDM

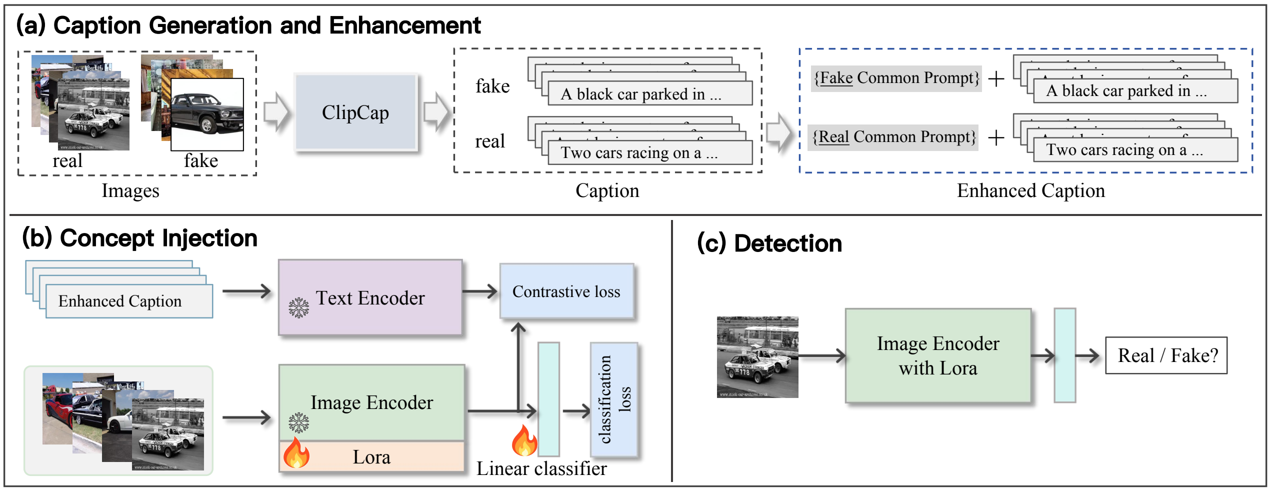


Figure 4: Architecture of C2P-CLIP for Generalizable Deepfake Detection. (a) Caption Generation and Enhancement. We obtain the caption of images using ClipCap, and leverage category common prompts to enhance those text. In this study, we adopt (Trump, Biden), (Deepfake, Camera) as the category common prompts. (b) Concept Injection (Training stage). We use the text-image pair to train the Lora layers and classifier by contrastive loss and classification loss. (c) Detection (Testing stage). Only image encoder and classifier are utilized to perform detection.

Methodology

In this section, we introduce our C2P-CLIP, a universal approach designed for generalizable deepfake detection. The overall architecture is illustrated in Figure 4. Our method comprises three two components: Caption Generation and Enhancement and Concept Injection.

Caption Generation and Enhancement

The ability of CLIP features to perform forgery detection through linear classification is a fascinating phenomenon. As discussed in Section 2, we attribute this capability to CLIP's mechanism of seeking similar concepts. Building on this insight, we hypothesize that injecting explicit classification concepts into the image encoder can significantly enhance detection performance. To this end, the category common prompts are designed to enhance the captions associated with the images. These enhanced captions are then utilized in contrastive learning to embed the classification concepts into the image encoder. Specifically, during the training stage, we append a consistent prompt, such as "Camera" or "Biden", to all captions of real images, and a different prompt, such as "Deepfake" or "Trump", to the captions of fake images.

Let us consider a training dataset X containing both real and fake images, defined as follows:

$$X = \{x_j, y_j\}_{j=1}^{N}, \quad y \in \{0, 1\}, \tag{1}$$

where $y = 1$ indicates that the image is fake, and $y = 0$ indicates that the image is real. For each image in the training set, we obtain its corresponding caption using the ClipCap model. The set of captions associated with the training images is denoted as C, defined as:

$$C = \{c_j, y_j\}_{j=1}^{N}, \quad y \in \{0, 1\}. \tag{2}$$

Next, we utilize the category common prompts $P = \{P_{real}, P_{fake}\}$ to enhance the captions as follows:

$$\begin{aligned} \widetilde{C} &= \{\widetilde{c_j}\}_{j=1}^{N}, \\ \widetilde{c_j} &= \begin{cases} (P_{real}, c_j), & \text{if } y = 0, \\ (P_{fake}, c_j), & \text{if } y = 1, \end{cases} \end{aligned} \tag{3}$$

where $P = \{P_{real}, P_{fake}\}$ are typically assigned as pairs of words different from the image caption, such as $(P_{real} = Camera, P_{fake} = Deepfake)$ or $(P_{real} = Biden, P_{fake} = Trump)$. These category common prompts, when appended to the original captions, enhance the textual context, enabling the model to better distinguish between real and fake images. The enhanced captions $\widetilde{C}$ are then used in contrastive learning to transmit the information from the category common prompts into the image encoder.

Concept Injection (Training Stage)

To embed the deepfake detection concept from the enhanced captions into the image encoder, we employ contrastive learning during the training stage. In this phase, the image- and text encoder are kept frozen, while the Lora layers are applied to the image encoder. The goal is to transfer the categorical concepts embedded in the enhanced captions into the image encoder, thereby improving its deepfake detection capabilities. To achieve this, we use two key losses: contrastive loss and classification loss. Together, these losses guide the image encoder to effectively learn and apply the deepfake detection concepts during training, enhancing its generalization performance across various unseen sources.

Specifically, we first compute text features u and image features v as follows::

$$\begin{aligned} u_j &= encoder_{text}(\widetilde{c_j}), \\ v_j &= encoder_{img}^{lora}(x_j). \end{aligned} \tag{4}$$

Methods	GAN						Deep fakes	Low level		Perc. loss		ADM	LDM			Glide			Dalle	mAP
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/cfg	100 steps	100 27	50 27	100 10		
CNN-Spot	100.0	93.5	84.5	99.5	89.5	98.2	89.0	73.8	59.5	98.2	98.4	73.7	70.6	71.0	70.5	80.7	84.9	82.1	70.6	83.6
Patchfor	80.9	72.8	71.7	85.8	66.0	69.3	76.6	76.2	76.3	74.5	68.5	75.0	87.1	86.7	86.4	85.4	83.7	78.4	75.7	77.7
Co-occ	99.7	81.0	50.6	98.6	53.1	68.0	59.1	69.0	60.4	73.1	87.2	70.2	91.2	89.0	92.4	89.3	88.4	82.8	81.0	78.1
Freq-spec	55.4	100.0	75.1	55.1	66.1	100.0	45.2	47.5	57.1	53.6	51.0	57.7	77.7	77.3	76.5	68.6	64.6	61.9	67.8	66.2
F3Net	100.0	84.3	69.9	99.7	56.7	100.0	78.8	52.9	46.7	63.4	64.4	70.5	73.8	81.7	74.6	89.8	91.0	90.9	71.8	76.9
UniFD	100.0	98.1	94.5	86.7	99.3	99.5	91.7	78.5	67.5	83.1	91.1	79.2	95.8	79.8	95.9	93.9	95.1	94.6	88.5	90.1
LGrad	100.0	94.0	90.7	99.9	79.4	100.0	67.9	59.4	51.4	63.5	69.6	87.1	99.0	99.2	99.2	93.2	95.1	94.9	97.2	86.4
FreqNet	99.9	99.6	96.1	99.9	99.7	98.6	99.9	94.4	74.6	80.1	75.7	96.3	96.1	100.0	62.3	99.8	99.8	96.4	77.8	92.0
NPR	100.0	99.5	94.5	99.9	88.8	100.0	84.4	98.0	100.0	50.2	50.2	98.3	99.9	99.9	99.9	99.9	99.9	99.9	99.3	92.8
FatFormer	100.0	100.0	100.0	99.8	100.0	100.0	98.0	97.9	81.2	99.8	99.9	92.0	99.8	99.1	99.9	99.1	99.4	99.2	99.8	98.2
Ours-P1	100.0	99.9	99.9	100.0	100.0	100.0	97.3	99.9	76.0	99.8	99.9	92.2	100.0	99.8	100.0	99.3	99.3	99.4	99.9	98.0
Ours-P2	100.0	100.0	100.0	99.5	100.0	100.0	98.6	98.9	84.6	99.9	100.0	94.1	100.0	99.8	100.0	99.7	99.8	99.8	99.9	**98.7**

Table 1: Cross-model Average Precision (AP) Performance on the UniversalFakeDetect Dataset. We copy the results of CNN-Spot(2020), Patchfor(2020), Co-occurence(2019), Freq-spec(2019), and UniFD(2023) from paper (Ojha et al. 2023), and obtain results of F3Net(2020), LGrad(2023), FreqNet(2024b), NPR(2024c), and FatFormer(2024) using the official pre-trained model or re-implemented model. **Bold** and underline represent the best and second-best performance, respectively. The sets P1 and P2 are defined as Trump, Biden and Deepfake, Camera, respectively.

Here, $encoder_{text}()$ denotes the text encoder, and $encoder_{img}^{lora}()$ refers to the image encoder equipped with the Lora layer. Next, we compute the contrastive learning loss as follows:

$$\begin{aligned} \mathcal{L}_{contrastive} &= (\mathcal{L}_{v->u} + \mathcal{L}_{u->v})/2, \\ \mathcal{L}_{v->u} &= -\frac{1}{N}\sum_{i}^{N} log\frac{exp(v_i^T u_i)}{\sum_{j=1}^{N} exp(v_i^T u_j)}, \\ \mathcal{L}_{u->v} &= -\frac{1}{N}\sum_{i}^{N} log\frac{exp(u_i^T v_i)}{\sum_{j=1}^{N} exp(u_i^T v_j)}. \end{aligned} \quad (5)$$

In addition, we also apply a linear classifier $Linear$ on image features v to perform the classification. The cross-entropy loss is adopted as the classification loss $\mathcal{L}_{classification}$ (Liu et al. 2025). The final loss function is obtained by the weighted sum of the above loss functions.

$$\mathcal{L} = \mathcal{L}_{contrastive} + \alpha * \mathcal{L}_{classification}, \quad (6)$$

where α are hyper-parameters for balancing two losses.

Detection (Testing Stage)

During the evaluation phase, only the image encoder and the classifier are utilized to perform the detection. The detection process is as follows:

$$p = classifier(\widetilde{encoder}_{img}(x)). \quad (7)$$

Here, $\widetilde{encoder}_{img}$ represents the image encoder integrated with the Lora parameters, which have been fine-tuned during the training stage.

Experiments

Datasets

To demonstrate the detection performance of our C2P-CLIP, we conduct the generalization evaluation on two widely used datasets following baselines (Ojha et al. 2023; Zhu et al. 2024), including UniversalFakeDetect dataset (Ojha et al. 2023) and GenImage dataset (Zhu et al. 2024).

UniversalFakeDetect Dateset This dataset uses ProGAN as the training set following (Wang et al. 2020), which includes 20 subsets of generated images. For training, we adopt the 4-class setting $(horse, chair, cat, car)$ following (Tan et al. 2024c; Liu et al. 2024). The test set comprises 19 subsets from various generative models, including: ProGAN (Karras et al. 2018), StyleGAN (Karras et al. 2019), BigGAN (Brock et al. 2018), CycleGAN (Zhu et al. 2017), StarGAN (Choi et al. 2018), GauGAN (Park et al. 2019) and Deepfake (Rossler et al. 2019), CRN (Chen and Koltun 2017), IMLE (Li, Zhang, and Malik 2019), SAN (Dai et al. 2019), SITD (Chen et al. 2018), Guided diffusion model (Dhariwal et al. 2021), LDM (Rombach et al. 2022), Glide (Nichol et al. 2021), DALLE (Ramesh et al. 2021).

Genimage dataset This dataset primarily employs the Diffusion model for image generation, including Midjourney (Holz et al. 2022), SDv1.4 (Rombach et al. 2022), SDv1.5 (Rombach et al. 2022), ADM (Dhariwal et al. 2021), GLIDE (Nichol et al. 2021), Wukong (Gu, Meng et al. 2022. 5), VQDM (Gu et al. 2022), BigGAN (Brock et al. 2018). Following the settings on GenImage, we use SDv1.4 as the training set and the remaining models as the test set.

Implementation Details We utilize the Adam optimizer with an initial learning rate of 4×10^{-4}. The batch size is set to 128. The ViT-L/14 model of CLIP is adopted as the pre-trained model following the baseline UniFD. We apply Lora layers on the q_proj, k_proj, and v_proj layers using the Parameter-Efficient Fine-Tuning (PEFT) (Mangrulkar et al. 2022) library. For the setting of the Lora layers, we configure the hyperparameters as follows: $lora\_r = 6$, $lora\_alpha = 6$, and $lora\_dropout = 0.8$. The hyperparameter α is set to 8.0. A random seed of 123 is used for reproducibility. The

Methods	GAN						Deep fakes	Low level		Perc. loss		ADM	LDM			Glide			Dalle	mAcc
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/cfg	100 steps	100 27	50 27	100 10		
CNN-Spot	100.0	85.2	70.2	85.7	79.0	91.7	53.5	66.7	48.7	86.3	86.3	60.1	54.0	55.0	54.1	60.8	63.8	65.7	55.6	69.6
Patchfor	75.0	69.0	68.5	79.2	64.2	63.9	75.5	75.1	75.3	72.3	55.3	67.4	76.5	76.1	75.8	74.8	73.3	68.5	67.9	71.2
Co-occ	97.7	63.2	53.8	92.5	51.1	54.7	57.1	63.1	55.9	65.7	65.8	60.5	70.7	70.6	71.0	70.6	69.6	69.9	67.6	66.9
Freq-spec	49.9	99.9	50.5	49.9	50.3	99.7	50.1	50.0	48.0	50.6	50.1	50.9	50.4	50.4	50.3	51.7	51.4	50.4	50.0	55.5
F3Net	99.4	76.4	65.3	92.6	58.1	100.0	63.5	54.2	47.3	51.5	51.5	69.2	68.2	75.4	68.8	81.7	83.25	83.1	66.3	71.3
UniFD	100.0	98.5	94.5	82.0	99.5	97.0	66.6	63.0	57.5	59.5	72.0	70.0	94.2	73.8	94.4	79.1	79.9	78.1	86.8	81.4
LGrad	99.8	85.4	82.9	94.8	72.5	99.6	58.0	62.5	50.0	50.7	50.8	77.5	94.2	95.9	94.8	87.4	90.7	89.6	88.4	80.3
FreqNet	97.9	95.8	90.5	97.6	90.2	93.4	97.4	88.9	59.0	71.9	67.4	86.7	84.6	99.6	65.6	85.7	97.4	88.2	59.1	85.1
NPR	99.8	95.0	87.6	96.2	86.6	99.8	76.9	66.9	98.6	50.0	50.0	84.6	97.7	98.0	98.2	96.3	97.2	97.4	87.2	87.6
FatFormer	99.9	99.3	99.5	97.2	99.4	99.8	93.2	81.1	68.0	69.5	69.5	76.0	98.6	94.9	98.7	94.4	94.7	94.2	98.8	90.9
Ours-P1	99.7	90.7	95.3	99.4	95.3	96.6	89.9	98.3	64.6	90.7	90.7	77.8	99.1	98.1	99.0	94.7	94.2	94.4	98.8	93.0
Ours-P2	100.0	97.3	99.1	96.4	99.2	99.6	93.8	95.6	64.4	93.3	93.3	69.1	99.3	97.3	99.3	95.3	95.3	96.1	98.6	**93.8**

Table 2: Cross-model Accuracy (Acc) Performance on the UniversalFakeDetect Dataset.

Method	Ref	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	mAcc
ResNet-50(2016)	CVPR2016	54.9	99.9	99.7	53.5	61.9	98.2	56.6	52.0	72.1
DeiT-S(2021)	ICML2021	55.6	99.9	99.8	49.8	58.1	98.9	56.9	53.5	71.6
Swin-T(2021)	ICCV2021	62.1	99.9	99.8	49.8	67.6	99.1	62.3	57.6	74.8
CNNSpot(2020)	CVPR2020	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec(2019)	WIFS2019	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
F3Net(2020)	ECCV2020	50.1	99.9	99.9	49.9	50.0	99.9	49.9	49.9	68.7
GramNet(2020)	CVPR2020	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9
UnivFD(2023)	CVPR2023	93.9	96.4	96.2	71.9	85.4	94.3	81.6	90.5	88.8
NPR (2024c)	CVPR2024	81.0	98.2	97.9	76.9	89.8	96.9	84.1	84.2	88.6
FreqNet (2024b)	AAAI2024	89.6	98.8	98.6	66.8	86.5	97.3	75.8	81.4	86.8
FatFormer(2024)	CVPR2024	92.7	100.0	99.9	75.9	88.0	99.9	98.8	55.8	88.9
Ours	Trump,Biden	82.2	95.1	95.5	95.1	98.9	98.7	93.8	98.3	94.7
Ours	Deepfake,Camera	88.2	90.9	97.9	96.4	99.0	98.8	96.5	98.7	**95.8**

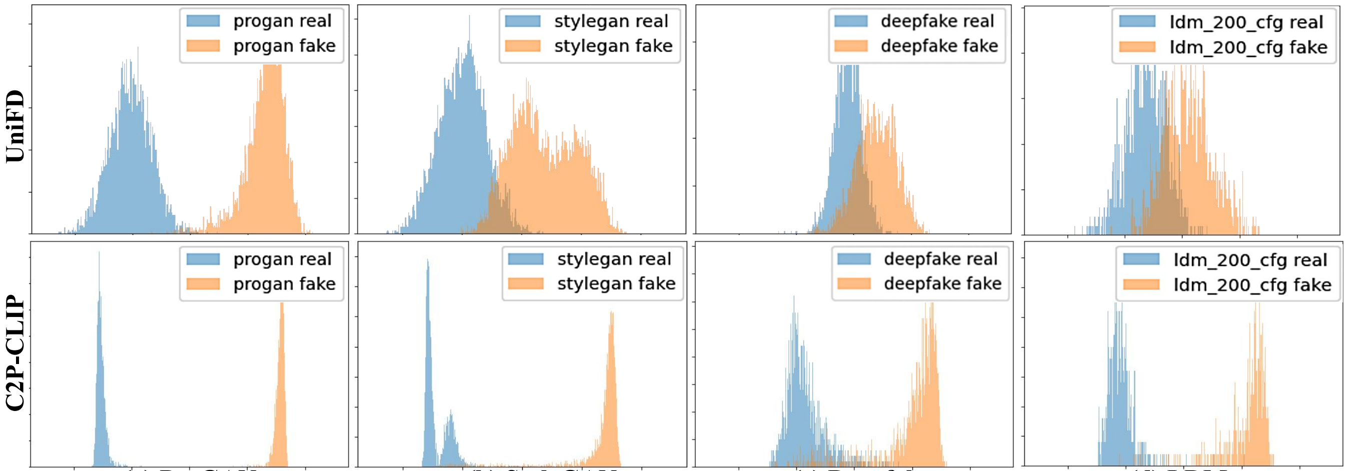
Table 3: Cross-model Accuracy (Acc) Performance on the Genimage Dataset. The SDv1.4 is employed as the training set following (Zhu et al. 2024). We copy the results of ResNet-50, DeiT-S, Swin-T, CNNSpot, Spec, F3Net, GramNet from paper (Zhu et al. 2024), and obtain the results of UnivFD, FreqNet, NPR, and FatFormer using re-implement model.

proposed method is implemented using Pytorch on 4 Nvidia GeForce RTX 4090 GPUs. Following the baselines (Ojha et al. 2023; Liu et al. 2024), we use mean average precision (mAP) and mean accuracy (mAcc) as evaluation metrics.

Quantitative Analysis

Evaluation on UniversalFakeDetect The results of average precision (AP) and accuracy (Acc) are shown in Table 1 and 2, respectively. Our method is trained using ProGAN with 4 training settings, achieving detection results of 93.8% mAcc and 98.7% mAP on the 19 test subsets. The baseline UniFD directly uses the original CLIP for deepfake detection. In contrast, our method injects category concepts into the visual encoder. Compared to UniFD, our method improves mAcc by 12.4% and mAP by 8.5%. This indicates that the proposed category common prompts effectively enhance CLIP's deepfake detection capability. Furthermore, compared to the latest state-of-the-art method, FatFormer, our method improves accuracy by 2.9%. FatFormer employs frequency blocks and text encoders as adapters to fine-tune the visual encoder, which adds extra inference parameters. In contrast, our method only uses text encoders and Lora parameters during training, without adding parameters during testing, thus achieving a simple but effective improvement in CLIP's detection performance. Additionally, we test the detection performance with different categories common prompts. Using prompts (Trump, Biden) and (Deepfake, Camera), we achieved detection accuracies of 93.8% and 93.0%, respectively. This demonstrates that our method can improve performance without relying on specific prompts.

Evaluation on Genimage We show the results on the Genimage dataset in Table 3. The results of accuracy (Acc) are presented. The Genimage dataset includes 7 diffusion models and one GAN model, focusing primarily on the detection performance of methods on recent diffusion models. Due to the varying image sizes in the Genimage dataset, images smaller than 224 pixels are duplicated and then cropped to 224 pixels. We utilize the same setting to re-implement UnivFD, FreqNet, FatFormer, and NPR. When using SDv1.4 as the training set, our method achieves a 95.8% accuracy rate on the test set. Compared with the baseline UniFD, our method improves accuracy by 7.0%. Furthermore, compared to the state-of-the-art methods, the proposed method improved accuracy by 6.9%. This demon-



(a) ProGAN **(b) StyleGAN** **(c) Deepfake** **(d) LDM**

Figure 5: Logit distributions of extracted forgery features. We compare the baseline UniFD and our C2P-CLIP. A total of four testing GANs and diffusion models are considered, including ProGAN, StyleGAN, Deepfake, LDM, and DALLE.

C2 Prompt	Contrastive learning	Lora	mAcc.
			79.7
		✓	88.6
	✓	✓	89.7
✓	✓	✓	93.8

Table 4: Ablation Study. Three components are considered, including Lora, Contrastive Learning, and Category Common Prompt (C2 Prompt).

strates that our method performs well with diffusion models for training. Additionally, we also train on the Genimage dataset using different categories common prompts. When using prompts (Trump, Biden) and (Deepfake, Camera), the accuracy rates reached 94.7% and 95.8%, respectively.

Qualitative Analysis

To further assess the generalization capability of our method, we visualize the logit distributions of both UniFD and our approach, as illustrated in Figure5. This visualization sheds light on the degree of separation between 'real' and 'fake' images during the evaluation phase, thus indicating the effectiveness of our method in generalizing across various forgery representations. Our analysis shows the considerable overlap between 'real' and 'fake' regions when the baseline UniFD encounters previously unseen GANs or diffusion models, leading to the misclassification of forgeries as 'real'. In contrast, our approach maintains clear differentiation between 'real' and 'fake' categories, even when confronted with unseen sources. These findings underscore the robustness of our method in enhancing the separation between 'real' and 'fake' classes and demonstrate its superior generalization across a diverse array of image sources.

Ablation Study

We perform ablation experiments on the UniversalFakeDetect dataset. Three components are considered, including Lora, Contrastive Learning, and Category Common Prompt. Since contrastive learning only makes sense when lora is enabled, we have established four scenarios. When all three components are disabled, it is equivalent to the UniFD composed of a frozen CLIP and a learnable MLP. The results presented in Table4 demonstrate a consistent improvement through the incremental addition of three modules, increasing from 79.7% to 93.8%. Specifically, enabling lora compared to UniFD raises mAcc. from 79.7% to 88.6%. Subsequently, incorporating contrastive learning involves using a frozen text encoder to supervise the image encoder, resulting in a 1.1% improvement. Furthermore, employing a simple category common prompt enhances the text to inject manually specified category concepts into the image encoder, leading to an additional 4.1% improvement.

Conclusion

In this study, we first sought to understand why CLIP features are effective for deepfake detection through a linear classifier. To investigate what information in the image features contributes to the detection, we decode the detection features into text and conduct a word frequency analysis. To the best of our knowledge, this is the first time this issue has been approached from this perspective. Our findings indicate that CLIP performs classification by matching similar concepts rather than discerning true and false. Based on this conclusion, we propose category common prompts to finetune the image encoder by manually constructing category concepts combined with contrastive learning. This approach led to an improvement in detection performance. However, a limitation of our method is that, while it uses word frequency to analyze detection features, it lacks a comprehensive analysis of the entire caption, resulting in incomplete information. We aim to address this issue in future work.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (No.2021ZD0112100), National Natural Science Foundation of China (No.U24B20179, No.62120106009,No.92470203,No.U23A20314), National Natural Science Foundation of China under Grants 62072394, in part by the Natural Science Foundation of Hebei province under Grant F2024203049, Beijing Natural Science Foundation (No.L242022,No.L242021), and the Fundamental Research Funds for the Central Universities (2024XKRC082).

References

Brock, A.; et al. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Chai, L.; et al. 2020. What makes fake images detectable? understanding properties that generalize. In *European Conference on Computer Vision*, 103–120. Springer.

Chen, C.; Chen, Q.; Xu, J.; and Koltun, V. 2018. Learning to see in the dark. In *Proceedings of the IEEE Conference on CVPR*, 3291–3300.

Chen, L.; et al. 2022. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF Conference on CVPR*, 18710–18719.

Chen, Q.; and Koltun, V. 2017. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE ICCV*, 1511–1520.

Chen, Y.; Zhang, L.; Niu, Y.; Chen, P.; Tan, L.; and Zhou, J. 2024a. Guided and Fused: Efficient Frozen CLIP-ViT with Feature Guidance and Multi-Stage Feature Fusion for Generalizable Deepfake Detection. *arXiv preprint arXiv:2408.13697*.

Chen, Y.; Zhang, L.; Niu, Y.; Tan, L.; and Chen, P. 2024b. Learning on Less: Constraining Pre-trained Model Learning for Generalizable Diffusion-Generated Image Detection. *arXiv preprint arXiv:2412.00665*.

Choi, Y.; et al. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on CVPR*, 8789–8797.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on CVPR*, 11065–11074.

Dhariwal, P.; et al. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Goodfellow, I. J.; et al. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*.

Gu, J.; Meng, X.; et al. 2022. 5. Wukong, 2022. 5. Inhttps: //xihe.mindspore.cn/modelzoo/wukong.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on CVPR*, 10696–10706.

He, K.; et al. 2016. Deep residual learning for image Recognition. In *Proceedings of the IEEE Conference on CVPR*, 770–778.

He, Y.; et al. 2021. Beyond the Spectrum: Detecting Deepfakes via Re-Synthesis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2534–2541. International Joint Conferences on Artificial Intelligence Organization.

Ho, J.; et al. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Holz, D.; et al. 2022. Midjourney. Inhttps://www.midjourney.com/home/.

Huang, T.; Dong, B.; Yang, Y.; Huang, X.; Lau, R. W.; Ouyang, W.; and Zuo, W. 2023. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF ICCV*, 22157–22167.

Hui, W.; Zhu, Z.; Zheng, S.; and Zhao, Y. 2024. Endow SAM with Keen Eyes: Temporal-spatial Prompt Learning for Video Camouflaged Object Detection. In *Proceedings of the IEEE/CVF Conference on CVPR*, 19058–19067.

Jeong, Y.; et al. 2022. BiHPF: Bilateral High-Pass Filters for Robust Deepfake Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 48–57.

Jiao, S.; Wei, Y.; Wang, Y.; Zhao, Y.; and Shi, H. 2023. Learning mask-aware clip representations for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 36: 35631–35653.

Karras, T.; et al. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

Karras, T.; et al. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on CVPR*, 4401–4410.

Li, J.; et al. 2021. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF Conference on CVPR*, 6458–6467.

Li, K.; Zhang, T.; and Malik, J. 2019. Diverse image synthesis from semantic layouts via conditional imle. In *Proceedings of the IEEE/CVF ICCV*, 4220–4229.

Liu, H.; Tan, Z.; Tan, C.; Wei, Y.; Wang, J.; and Zhao, Y. 2024. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on CVPR*, 10770–10780.

Liu, M.; Bai, H.; Li, F.; Zhang, C.; Wei, Y.; Wang, M.; Chua, T.-S.; and Zhao, Y. 2025. PSVMA+: Exploring Multi-Granularity Semantic-Visual Adaption for Generalized Zero-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1): 51–66.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical Vision transformer using shifted windows. In *Proceedings of the IEEE/CVF ICCV*, 10012–10022.

Liu, Z.; et al. 2020. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF Conference on CVPR*, 8060–8069.

Mangrulkar, S.; Gugger, S.; Debut, L.; Belkada, Y.; Paul, S.; and Bossan, B. 2022. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. https://github.com/huggingface/peft.

Masi, I.; et al. 2020. Two-branch recurrent network for isolating deepfakes in videos. In *European Conference on Computer Vision*, 667–684. Springer.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nataraj, L.; Mohammed, T. M.; Chandrasekaran, S.; Flenner, A.; Bappy, J. H.; Roy-Chowdhury, A. K.; and Manjunath, B. 2019. Detecting GAN generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Niu, Y.; Tondi, B.; Zhao, Y.; Ni, R.; and Barni, M. 2021. Image splicing detection, localization and attribution via JPEG primary quantization matrix estimation and clustering. *IEEE Transactions on Information Forensics and Security*, 16: 5397–5412.

Ojha, U.; et al. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on CVPR*, 24480–24489.

Park, T.; et al. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on CVPR*, 2337–2346.

Qian, Y.; et al. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, 86–103. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language superVision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. Pmlr.

Rombach, R.; et al. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on CVPR*, 10684–10695.

Rossler, A.; et al. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF ICCV*, 1–11.

Shiohara, K.; et al. 2022. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on CVPR*, 18720–18729.

Tan, C.; Liu, P.; Tao, R.; Liu, H.; Zhao, Y.; Wu, B.; and Wei, Y. 2024a. Data-independent operator: A training-free artifact representation extractor for generalizable deepfake detection. *arXiv preprint arXiv:2403.06803*.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024b. Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Domain Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5052–5060.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024c. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on CVPR*, 28130–28139.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; and Wei, Y. 2023. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *Proceedings of the IEEE/CVF Conference on CVPR*, 12105–12114.

Tao, R.; Le, M.; Tan, C.; Liu, H.; Qin, H.; and Zhao, Y. 2024. ODDN: Addressing Unpaired Data Challenges in Open-World Deepfake Detection on Online Social Networks. *arXiv preprint arXiv:2410.18687*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357. PMLR.

Wang, C.; et al. 2021. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF Conference on CVPR*, 14923–14932.

Wang, S.-Y.; et al. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on CVPR*, 8695–8704.

Yan, Z.; Luo, Y.; Lyu, S.; Liu, Q.; and Wu, B. 2024. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on CVPR*, 8984–8994.

Yan, Z.; Zhang, Y.; Fan, Y.; and Wu, B. 2023. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF ICCV*, 22412–22423.

Yin, Y.; Xu, D.; Tan, C.; Liu, P.; Zhao, Y.; and Wei, Y. 2023. Cle diffusion: Controllable light enhancement diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8145–8156.

Zhang, X.; et al. 2019. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–6. IEEE.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*.

Zhu, J.-Y.; et al. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE ICCV*, 2223–2232.

Zhu, M.; Chen, H.; Yan, Q.; Huang, X.; Lin, G.; Li, W.; Tu, Z.; Hu, H.; Hu, J.; and Wang, Y. 2024. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36.