RESEARCH-ARTICLE

GM-DF: Generalized Multi-Scenario Deepfake Detection

YINGXIN LAI, Great Bay University, Dongguan, Guangdong, China

HONGYANG WANG, Shijiazhuang Tiedao University, Shijiazhuang, Hebei, China

JING YANG, Great Bay University, Dongguan, Guangdong, China

XIANGUI KANG, Sun Yat-Sen University, Guangzhou, Guangdong, China

BIN LI, Shenzhen University, Shenzhen, Guangdong, China

LINLIN SHEN, Shenzhen University, Shenzhen, Guangdong, China

View all

Open Access Support provided by:

Shijiazhuang Tiedao University

Shenzhen University

Great Bay University

Sun Yat-Sen University

Published: 27 October 2025

Citation in BibTeX format

MM '25: The 33rd ACM International Conference on Multimedia
October 27 - 31, 2025
Dublin, Ireland

Conference Sponsors:
SIGMM

https://doi.org/10.1145/3746027.3755386
ISBN: 9798400720352

GM-DF: Generalized Multi-Scenario Deepfake Detection

Yingxin Lai
Great Bay University
Xiamen University
Dongguan, China
laiyingxin2@gmail.com

Hongyang Wang
Shijiazhuang Tiedao University
BeiHe, China
1202310098@student.stdu.edu.cn

Jing Yang
Great Bay University
Dongguan, China
imyangjing0@gmail.com

Xiangui Kang
Sun Yat-Sen University
Shenzhen, China
isskxg@mail.sysu.edu.cn

Bin Li
Shenzhen University
Shenzhen, China
libin@szu.edu.cn

Linlin Shen
Shenzhen University
Shenzhen, China
llshen@szu.edu.cn

Zitong Yu*
Great Bay University
National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
Guangdong Provincial Key Laboratory of Intelligent Information Processing &
Shenzhen Key Laboratory of Media Security, Shenzhen University
Dongguan Key Laboratory for Intelligence and Information Technology
Dongguan, China
yuzitong@gbu.edu.cn

Abstract

Recent advances in face forgery detection have shown strong in-domain performance but often fail to generalize to out-of-distribution data, especially when confronted with unseen manipulation techniques or domain shifts (e.g., lighting conditions, camera noise). We propose a novel Mixture-of-Experts framework, termed GM-DF, that decouples domain-specific and domain-invariant features to tackle cross-domain face forgery detection. Our method builds upon a foundation model (CLIP) and incorporates three key modules: (1) Dataset–Embedding Generator that leverages lightweight expert layers and database-aware feature normalization to adaptively modulate features at a per-domain level, capturing idiosyncratic cues without overfitting; (2) Multi-Dataset Representation mechanism that fuses these expert embeddings using scaled dot-product attention and integrates a mask image modeling (MIM) task to amplify local forgery artifacts; (3) Meta-Domain-Embedding Optimizer, inspired by MAML, which alternates between domain-specific (inner-loop) and domain-invariant (outer-loop) updates to facilitate rapid adaptation on new domains. Additionally, inspired by [13] we introduce second-order feature propagation in the intermediate layers of CLIP to enhance fine-grained artifact cues and propose domain-class disentangled prompts to flexibly encode multi-domain text representations. Together, these strategies enable GM-DF to learn robust, shared forgery cues while preserving essential domain nuances. Our extensive experiments on multiple cross-domain benchmarks demonstrate that GM-DF significantly outperforms state-of-the-art approaches in both detection accuracy and domain transferability, reducing reliance on superficial artifacts and improving generalization to unseen forgeries. Importantly, our design requires minimal overhead beyond standard CLIP, making GM-DF both effective and computationally efficient for real-world face forgery detection.

*Corresponding author

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755386

CCS Concepts

• **Face systems → Image Safety**; **Visual content-based safety**.

Keywords

Deepfake Detection; MoE

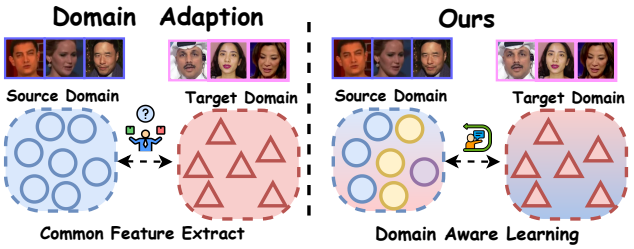

Figure 1: A comparison between our method and existing deepfake detection baselines reveals key differences.

ACM Reference Format:
Yingxin Lai, Hongyang Wang, Jing Yang, Xiangui Kang, Bin Li, Linlin Shen, and Zitong Yu. 2025. GM-DF: Generalized Multi-Scenario Deepfake Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, Dublin, Ireland, 10 pages. https://doi.org/10.1145/3746027.3755386

1 Introduction

Advancements in deep learning have fueled the development of sophisticated face forgery techniques, such as Deepfake, FaceSwap, Face2Face, NeuralTextures, and others. These methods enable the creation of highly realistic forged face images, posing significant threats to political and personal reputations and giving rise to pressing social challenges. As a result, developing robust forgery detection methods to mitigate these risks has become imperative.

Early face forgery detection methods used CNNs for binary classification. To address their limitations, subsequent work explored cues like noise supervision [39, 46], frequency domain information [12, 47], and reconstruction-based modeling [3, 5]. While these approaches achieve high intra-dataset accuracy, they often fail to generalize across different domains. The relentless emergence of diverse manipulated facial datasets has heightened the urgent need to improve generalization capabilities to address real-world application demands. However, previous approaches have predominantly concentrated on learning domain-invariant features, overlooking domain-specific characteristics and naively combining datasets with varying forgery techniques and heterogeneous distributions. This strategy gives rise to two critical challenges: **1) Heterogeneous Distribution Inconsistency:** In the high-dimensional feature space of model training, incompatible statistical differences across domains lead to heterogeneous distribution inconsistency. This misalignment hampers the ability of a unified discriminant function to effectively align multiple distributions, amplifying domain divergence and undermining the model's generalization across multi-source settings. Such an outcome fundamentally contradicts the objective of achieving comprehensive alignment within a single learning framework.**2) Domain-Specific Feature Entanglement :** This challenge arises from the model's inability to mitigate biases toward new distributions within the Structural Risk Minimization (SRM) framework, resulting in the coupling of distinct forgery features at the representation level. In the absence of techniques like orthogonal decomposition, disentangling shared and domain-specific features becomes exceedingly difficult. This entanglement substantially exacerbates error accumulation in representation learning, degrading the model's capacity to accurately differentiate authentic and forged samples across diverse domains.

To tackle these challenges, we propose GM-DF, a novel framework that integrates a shared backbone with a two-stage strategy to address domain-specific artifacts and unify forgery labels across diverse datasets. In the first stage, GM-DF employs soft partitioning and domain-aware modeling, enhancing the backbone with domain-specific layers and context-sensitive prompts. This preserves critical cues—like texture anomalies or manipulation artifacts—adapting to unique dataset traits while avoiding the pitfalls of a generic model. In the second stage, we align the global label embedding space using multi-prompt embeddings and ranking constraints, overcoming inconsistent labeling across datasets. This fosters a coherent semantic space and boosts knowledge transfer. Alternating between domain specialization and label unification, also prevents negative transfer, balancing adaptability. In addition, we also observed that the existing protocol is only from the perspective of a single data set. Although a data set has contained multiple forgery methods, the protocol for training and testing of multi-data sets is still lacking. In order to fill the research gap, we proposed a new protocol. Our contributions are as follows:

- We are the first to systematically address negative transfer and label inconsistency in multidata set deepfake detection, establishing a new protocol with five mainstream datasets to fairly evaluate generalisation capacity.
- We propose a unified framework featuring a domain adaptation strategy with soft partitioning and domain-aware modelling to minimise negative transfer, and a cross-dataset label alignment method using multiprompt embeddings and ranking constraints to unify inconsistent labels.
- Extensive experiments demonstrate that our method, GM-DF, achieves state-of-the-art performance on both traditional protocols and the proposed multidataset protocol, significantly benefiting from multidataset training.

2 Related Work

2.1 Face Forgery Detection

Due to significant security and privacy threats, face forgery detection has garnered considerable research interest. While early methods [15, 50] treated it as a binary classification problem and achieved excellent intra-dataset performance, they often suffer from overfitting and poor generalization. Consequently, research has shifted towards improving cross-domain robustness. Various approaches have been explored to capture forgery artifacts. Some works leverage frequency-domain cues [37, 43, 47] or spatial artifacts like blending boundaries [20, 30, 35]. Others have developed advanced learning strategies, such as contrastive learning to model inconsistencies [73] and specialized data augmentation to improve generalizability [4, 42]. More recently, vision foundation models have been utilized for robust forgery localization [26, 41, 72]. Inspired by advancements in vision-language models [28, 36], a recent trend focuses on interpretability. For instance, SIDA [22] enhances detection with semantic-driven explanations, thereby improving model transparency.

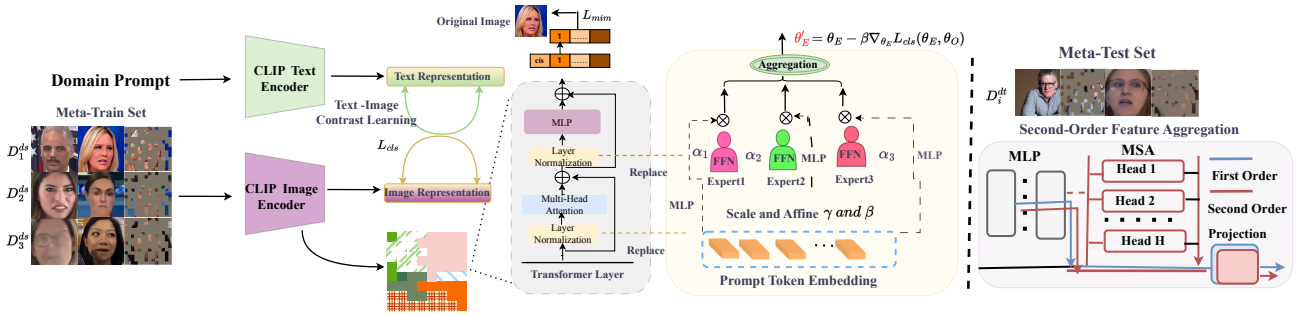

Figure 2: The framework of the proposed method. It integrates meta-learning modeling with image-text contrastive learning. It comprises three pivotal components: Dataset-Embedding Generator (DEG) and a Multi-Dataset Representation (MDP), as well as a Meta-Domain-Embedding Optimizer(MDEO). Firstly, the DEG incorporates a Dataset Information Layer (DIL) and a dynamic text feature affine aimed at mapping discriminative features unique to each domain, and the second part MDP is the face mask image modeling (MIM) reconstruction module, which provides additional detail information for the global features of CLIP. To consider the subtle difference between real and fake, we propose to use the second-order statistical features to constrain the feature distribution. In this process, MDEO was used to optimize two learned features.

2.2 Joint Training on Multiple Datasets

In traditional vision tasks [34, 68, 70, 75?] like object detection and semantic segmentation, naively combining data from multiple sources leads to poor generalization due to inconsistent label spaces and domain gaps. To address this, researchers have explored universal model training [7, 14]. For instance, some works unify detection heads and align features across domains using attention and instance relabeling [7]. Others train a universal detector by leveraging diverse supervisory signals to bypass explicit domain modeling [63], or use pseudo-labeling to show that a unified detector can surpass specialist models [74].While multi-domain training has been investigated for generic vision tasks, its application to face forgery detection remains largely unexplored. This area presents unique challenges, as forgery domains vary significantly in environments, media quality, and attack types. The resulting data imbalance and domain bias make developing a universal forgery detector particularly difficult.

3 Method

3.1 Dataset–Embedding Generator

We use CLIP as the foundation model for fine-tuning due to its strong generalization ability [52, 71]. This module follows the Mixture of Experts (MoE) [40] network structure to build a mixture of expert layers to learn domain-invariant features; unlike the domain-specific module we propose based on this, we use N independent experts. Each residual block consists of a Dataset Information Layer and a Multilayer Perceptron (MLP), Since the domain-specific embedding is much smaller than the normal backbone, it can be used if there is a low additional computational cost and restrain the trends of the overfit, experts from various domains carried out the process to extract domain-invariant and domain-specific features as follows:

$$F_(x) = F_\theta(x) + \Delta F_\theta^n(x) \tag{1}$$

Here, $F_\theta(x)$ represents the original function that is shared by all source domains to learn the common domain-invariant features. ΔF_θ^n adaptively extracts discriminative and unique domain-specific characteristics. Although existing work demonstrates that activations in different transformer blocks contribute to the stability of training [25, 52, 71], the diversity across individual domains is often compromised in multi-domain settings. To address this domain discrepancy in multi-dataset learning, we propose a database-aware feature normalization mechanism. This approach dynamically adjusts the feature statistics through learnable domain embeddings, enabling the model to effectively handle the characteristics specific to the dataset. Building upon the foundation of Layer normalization, we introduce domain-specific modulation parameters that are conditioned on the properties of the data set.

$$\mathrm{DNorm}(x, c_d) = \frac{x-\mu}{\sigma} \odot (\gamma_0 + \mathbf{W}\gamma c_d) + (\beta_0 + \mathbf{W}\beta c_d) \tag{2}$$

Here, $c_d \in \mathbb{R}^k$ represents the domain embedding for domain d, while $\mathbf{W}\gamma, \mathbf{W}\beta \in \mathbb{R}^{c\times k}$ are learnable projection matrices. The base parameters $\gamma_0, \beta_0 \in \mathbb{R}^c$ maintain the scaling and shifting of the domain-invariant characteristics, ensuring generalizability between domains.

To enable domain-specific adaptation, we apply this feature normalization in a cascaded manner through transformer blocks. This process allows for fine-tuned domain-specific feature adjustments while retaining the global structure learned across all domains.

$$\begin{aligned} x_{\text{att}} &= \mathrm{LayerNorm}(\mathrm{MHA}(x) + x) \\ x_{\text{ffn}} &= \mathrm{DNorm}(\mathrm{MLP}(x_{\text{att}}), c_d) + x_{\text{att}} \end{aligned} \tag{3}$$

In this structure, MHA denotes multi-head attention, and MLP represents the feedforward network. The inclusion of domain-specific modulation enables the model to adapt its feature representations in accordance with dataset-specific characteristics, improving its flexibility and robustness in multi-domain learning.

3.2 Multi-Dataset Representation

After obtaining the domain embedding and expert views, we calculate the scaled dot-product attention and mark it as the expert

views, which is formulated as [11]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{4}$$

where Q denotes the query, K denotes the key, $\sum$ denotes the value of the input embedding, and the scale factor of d_k is the key of dimension. Here we compute the attention score containing the task information Setting Q and K as the

$$Q = K = \text{Concat}(\Delta F_{\theta_1}(x), \Delta F_{\theta_2}(x), \ldots, \Delta F_{\theta_N}(x)) \in \mathbb{R}^{1\times N}, \tag{5}$$

where Concat denotes the operation that stacks vectors into a matrix. V is a matrix stacked by expert views. We summarize the expert views to obtain the aggregated expert view of the task.

Image-text pairs can learn semantic feature representations of face forgeries about specifics but may not be able to capture the details. Inspired by a previous study on forgery face reconstruction properties [3, 19, 21, 66, 69] and to improve the representation of facial details, we added a mask image modeling (MIM) task [17] that masks a number of patches of the input image and predicts their visual tokens. Commonly used, typical low-level visual tasks mask the image to capture low-level details and offer semantic information. With the learned representations, the reconstruction difference of real and fake faces significantly differs in distribution.

Given an input image X, we begin by dividing it into N patches denoted as $\{x_1, x_2, x_3, \ldots, x_n\}$, where n represents the total number of patches. Subsequently, we adopt a stochastic masking approach, referred to as [60] to apply masks to a subset of patches M. This process results in a modified image X', expressed as $X' = \{x_1, x'_{m2}, x'_{m3}, \ldots, x_n\}$. Here, x'_{m2} means that the second one is replaced by a mask. Next, we feed the masked images into a shared Transformer architecture, which yields a set of hidden vectors $\{h'_{\text{cls}}, h'_1, h'_2, \ldots, h'_N\}$. Using the knowledge encapsulated in these hidden vectors, we proceed to predict the masked regions $\{x'_{m_i} \mid m_i \in M\}$ and simultaneously perform direct pixel-level predictions.

To optimize memory consumption, a Gumbel-Softmax variational autoencoder [24] is used. Each image block is encoded into one of the T possible values, and a classification layer operates within the hidden vector space to indirectly predict the mask indices. The loss function is given as:

$$\mathcal{L}_{\text{mim}} = -\sum_{k\in M} \log p\left(q_k^{\phi}(x)|x'\right). \tag{6}$$

Here, $p(q_k^{\phi}(x)|\tilde{x})$ represents the classification score for classifying the k-th hidden vector belonging to the visual token $q_k^{\phi}(x)$, where q_{ϕ} is a categorical distribution.

We also propose a CLIP-based framework for cross-domain face forgery detection, addressing the limitations of existing methods in local artifact sensitivity and cross-domain adaptability. Traditional CLIP models, biased towards global semantics, struggle to capture micro-forgery traces (e.g., edge discontinuities, texture anomalies), while fixed text prompts lack semantic flexibility in multi-domain scenarios. Our framework introduces three tightly-coupled innovations: cross-layer second-order feature propagation for visual artifact enhancement, domain-class disentangled dynamic prompts for adaptive semantic space construction, and gradient-coupled contrastive learning for cross-modal fine-grained alignment.

To amplify subtle manipulation traces, we propose **second-order feature aggregation** in ViT's intermediate layers (8-10). For spatial token $\mathbf{z}_i^l$ at layer l, we model cross-layer feature propagation through subsequent attention blocks:

$$\Delta_{\text{so}}^l(\mathbf{x}) = \sum_{h=1}^{H}\sum_{i=0}^{K} a_{h,i}^{l'}(\mathbf{x})\mathbf{W}_{VO}^{l',h}\,\text{MLP}^l(\mathbf{z}_i^l) \tag{7}$$

where $a_{h,i}^{l'} \in [0,1]$ denotes cross-layer attention weights from the class token to spatial tokens, and $\mathbf{W}_{VO}^{l',h} \in \mathbb{R}^{d\times d}$ represents value-output projection matrices. The aggregated features are fused through lightweight adapters:

$$\mathbf{v}(\mathbf{x}) = \text{Proj}\left(\mathbf{z}_{\text{cls}}^L + \sum_{l=8}^{10}\alpha_l \cdot \text{Norm}(\Delta_{\text{so}}^l(\mathbf{x}))\right) \tag{8}$$

This creates dedicated pathways for amplifying high-frequency forgery patterns while maintaining 99% of CLIP's original parameters. To handle multi-domain scenarios, we decompose text prompts into learnable domain-class embeddings:

$$\mathcal{T}(d, c) = \text{"A } [\mathbf{e}_{\text{dom}}^d]\ [\mathbf{e}_{\text{cls}}^c] \text{ face showing forgery clues"} \tag{9}$$

where $\mathbf{e}_{\text{dom}}^d \in \mathbb{R}^{512}$ encodes domain-specific semantics (e.g., frequency artifacts for Deepfakes, geometric distortions for FaceSwap), and $\mathbf{e}_{\text{cls}}^c \in \{\mathbf{e}_{\text{real}}, \mathbf{e}_{\text{fake}}\}$ discriminates authenticity. A wildcard embedding $\mathbf{e}_{\text{dom}}^0$ enables zero-shot generalization to unseen domains through contrastive knowledge transfer. To handle domain shift issue, we propose the adaptive temperature scaling:

$$\tau_d = \tau_{\text{base}} + \varphi\|\mathbf{e}_{\text{dom}}^d\|_2 \tag{10}$$

The unified contrastive loss enforces triple alignment:

$$\mathcal{L}_{\text{sis}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(s(\mathbf{v}_i, \mathbf{u}_{d_i,c_i})/\tau_{d_i})}{\sum_{j=0}^{D}\sum_{k\in\{\text{real,fake}\}}\exp(s(\mathbf{v}_i, \mathbf{u}_{j,k})/\tau_{d_i}) + \epsilon} \tag{11}$$

This framework achieves: (1) Class separation via $\mathbf{e}_{\text{real}}/\mathbf{e}_{\text{fake}}$ divergence, (2) Domain invariance through $\mathbf{e}_{\text{dom}}^0$ regularization, and (3) Cross-modal consistency between visual artifacts and textual descriptions, with fewer than 1% additional parameters compared to vanilla CLIP. Based on the feature prior, it is instantiated as they calculate the distance between two distributions with mean and covariance matrices. Smaller distances represent that the source domains are closer to the target domain distribution.

3.3 Meta-Domain-Embedding Optimizer

In this subsection, we propose a metadomain embedding optimizer based on the MAML [11] paradigm to focus on the generic and personality characteristic abilities of learning domain-specific and domain-common characteristics. Here, we define each domain as a single task t. In the training process, we sample batches of multidomain data, consisting of meta-train set D_i^{ds} and meta-test set D_i^{dt}, here for simplicity we assume that the full model is described as a function $f(\cdot)$, which receives an image x as input and y as output. The loss function optimized per meta-train domain task during the training uses cross-entropy loss defined as

$$L_{cls}(f(\theta_E, \theta_O)) = \sum_{(x_j, y_j)\in D_{d_i}} \big[y_j \log f(x_j) + (1-y_j)\log(1-f(x_j))\big]. \tag{12}$$

In this process, referred to as the inner-loop updatation, we just update the learnable token parameter in the meta-train and freeze all other feature extraction. θ_E represents the meta-MoE's expert, while θ_O represents the base model's parameters. After generating the initial domain embeddings θ_i and evaluating the losses obtained in the data batch, the updated domain

embeddings are obtained by calculating the gradient of the losses L_{cls} and performing gradient descent updates.

$$\theta_E^{'} \leftarrow \theta_E - \beta \frac{\sigma L_{cls}\left(f\left(\theta_E, \theta_O\right)\right)}{\sigma \theta_E}, \tag{13}$$

where β is the learning rate of gradient descent. In the subsequent step, the model meta-parameters $\theta_E^{'}$ undergo optimization to enhance the performance of the meta-test set D_i^{ds} to obtain the loss L_{cls} and the prediction for domain i.

Similarly, during the meta-test phase, the meta-test sample D_i^{dt} is utilized to update the network. The features are aggregated using the aggregation model after passing through the expert layer. Furthermore, the loss of consistency L_{sis} of the features is used to minimize the distance between the source domain and the target domain with reconstructed facial features to aid in fine-grained forgery feature learning. The overall model loss is stated as follows

$$L_{total} = L_{sis} + L_{cls} + L_{mim}. \tag{14}$$

Then we can optimize the generator $f(\cdot)$ by the gradient:

$$\theta_O^{'} \leftarrow \theta_O - \delta \frac{\sigma L_{total}\left(f\left(\theta_E, \theta_O\right)\right)}{\sigma \theta_O}. \tag{15}$$

In summary, θ_E is updated during the meta-train process to learn the private characteristics of each domain and has higher flexibility due to the dynamic prompt vector. θ_O is updated during the meta-test process to capture generic forged clues, which helps the model acquire complementary information and be used for multi-domain training.

4 Experiments

4.1 Experimental Settings

Datasets. The performance of the proposed method is rigorously evaluated across multiple benchmark datasets for deepfake detection, including FaceForensics++ (FF++) [51], Celeb-DF (V2) [32], DFDC [10], and WildDeepfake [76]. These data sets are selected for their diversity and widespread adoption within the deepfake detection community, enabling a robust assessment under both established protocols and the newly introduced benchmark.

Implementation Details We use ViT-B/16 [49] as the backbone model. We used RetinaFace [9] to detect facial areas and scaled the image of the face to 224 × 224 with a patch size of 16. We trained the model using the Adam optimizer with the learning rate set to 3e-6. The batch size during training was 32, and 40 training epochs were performed. During the dataset merging phase, multiple datasets are randomly shuffled and then consolidated into a new dataset.

Protocols. Although some existing studies have proposed different forgery methods for training of a single dataset. No pilot study is available for training multiple datasets with diverse forgery patterns and large-scale characteristics in the real world. In addition to the training perspective, only a single forgery test domain is usually used in the evaluation of algorithm performance, which leads to biased comparisons of state-of-the-art methods. To address the above-mentioned issues, we provide a novel data arrangement and training/testing strategy to benchmark the fair evaluations. Specifically, the data sets 5 (each data set is regarded as an individual domain), that is, FF++ [51], WDF [76], Celeb-DF [32], DFDC [10], and DFF [54] are merged into a large set D, which can be further divided into training sets $\{D_{\text{FF++}}, D_{\text{WDF}}, D_{\text{Celeb-DF}}, D_{\text{DFF}}\}$ and test set $\{D_{\text{DFDC}}, D_{\text{i}}\}$. $i = \{\text{FF++}, \text{WDF}, \text{Celeb-DF}, \text{DFF}\}$ which denotes the subsets removed from the training set for the testing set. Specifically, considering the cost of training time, the large-scale DFDC [10] is only used for testing. We randomly select subsets of data $n \leq 3$ for training. $n = 1$ for the traditional single-domain training protocol while $n = 3$ denotes the newly established multi-domain protocols: $\{D_{\text{FF++}} \cup D_{\text{WDF}} \cup D_{\text{Celeb-DF}}\}$, $\{D_{\text{FF++}} \cup D_{\text{WDF}} \cup D_{\text{DFF}}\}$, $\{D_{\text{FF++}} \cup D_{\text{FF++}} \cup D_{\text{DFF}}\}$ and $\{D_{\text{WDF}} \cup D_{\text{Celeb-DF}} \cup D_{\text{DFF}}\}$ for training, respectively. More details can be found in the supplementary material.

Evaluation metrics. Three common metrics,Accuracy (ACC (%)), Area Under ROC Curve (AUC (%)), and Equal Error Rate (EER (%)) are adopted. EER is defined as the error rate when the false acceptance rate (FAR) is equal to the false rejection rate (FRR), and can be expressed as $EER = \frac{FRR+FAR}{2}$. However, in the realm of evaluating face forgery detection, we are more concerned with keeping the FAR at a relatively low level to ensure that it will not be easy to authenticate a forged face. For this purpose, we introduce the a priori probability of positive examples when calculating the EER. The impact on the system of a positive sample misclassified as a negative example is much greater than the impact of a negative sample as a positive example. To counteract this effect, we introduced the P_{real} parameter. In addition, we found that the original EER did not take into account the effect of testing in multiple domains. Calculating the EER directly on each dataset and then averaging the values may be affected by extreme values, and we judged performance against multiple domains by taking the maximum instead of simply averaging them. This ensures that our evaluation is more accurate and robust.

4.2 Results on Traditional Protocols

Single-Source Training and Testing: For a comprehensive comparison, we compare our method with other state-of-the-art methods under both frame-level in single-source training and multi-source training settings. The results in Table 1 establish our method as a state-of-the-art solution in deepfake detection, offering both high performance and exceptional generalizability. By outperforming the previous best method, LSDA, by 7% in average AUC and achieving a remarkable 10.7% gain on DFDC, our approach not only pushes the boundaries of academic research but also provides a practical tool for real-world applications. This work sets a new standard for cross-dataset evaluation, paving the way for more robust and reliable deepfake detection systems.

Multle-Source Training and Testing: Table 2 presents the performance of our method (Ours, GM-DF) in cross-manipulation evaluation, where the test datasets include GID-DF (C23 and C40) and GID-F2F (C23 and C40). The evaluation metrics are ACC (%) and AUC (%). Our method achieves the highest scores across all datasets and compression levels, significantly outperforming all baseline methods. For instance, on GID-DF (C23), the ACC reaches 92.52% and AUC is 96.79%, which are 3.89% and 1.55% higher than the second-best LSDA, respectively. On GID-F2F (C23), the ACC is 73.18% and AUC is 87.92%, outperforming LSDA's 70.21% and 84.79%. Even at higher compression levels (C40), such as GID-DF (C40) with ACC = 80.03% and AUC = 87.62%, and GID-F2F (C40) with ACC = 71.41% and AUC = 76.94%, our method still maintains a leading performance. This consistent superiority across different manipulation types and video qualities highlights the robustness and adaptability of our method, making it exceptionally well-suited for real-world deepfake detection challenges.

CLIP-Based Methods: Table 3 shows the results of our method in the cross-dataset evaluation for video-level AUC, where the training data is from FF++ and the test datasets include CDF-v2, DFDC, and DFD. Our method achieves the highest AUC scores across all datasets: 0.948 on CDF-v2, 0.841 on DFDC, and 0.963 on DFD, clearly outperforming all other methods. For instance, on CDF-v2, our method is 2.5% higher than the second-best FCG (0.923); on DFDC, it exceeds FCG's 0.812 by 2.9%; and on DFD, it surpasses VLFFD's 0.948 by 1.5%. Compared to existing methods like VLFFD, FFAA, and RepDFD, our method consistently shows an edge, particularly on the challenging DFDC dataset. The AUC of 0.963 on DFD highlights the remarkable ability of our method in detecting complex deepfakes, demonstrating that it makes a significant step forward in leveraging the potential of CLIP for robust and generalizable deepfake detection.

Table 1: Cross-dataset evaluation results (Frame-level AUC). All methods are trained on FF++ and evaluated on other datasets. The best results are indicated in bold, and the second-best results are underlined.

Method	Venue	CDF-v1	CDF-v2	DFDC	DFDCP	DFD	Avg.
Xception [51]	ICCV'19	0.779	0.737	0.708	0.737	0.816	0.755
EfficientB4 [58]	ICML'19	0.791	0.749	0.696	0.728	0.815	0.756
F3Net [47]	AAAI'20	0.777	0.735	0.702	0.735	0.798	0.749
X-ray [31]	CVPR'20	0.709	0.679	0.633	0.694	0.766	0.696
FFD [8]	CVPR'20	0.784	0.744	0.703	0.743	0.802	0.755
SPSL [37]	CVPR'21	0.815	0.765	0.704	0.741	0.812	0.767
SRM [39]	CVPR'21	0.793	0.755	0.700	0.741	0.812	0.760
Recce [3]	CVPR'22	0.768	0.732	0.713	0.734	0.812	0.752
SBI [53]	CVPR'22	-	0.813	-	0.799	0.774	-
UCF [66]	ICCV'23	0.779	0.753	0.719	0.759	0.807	0.763
ED [2]	AAAI'24	0.818	0.864	0.721	0.851	-	-
LSDA [64]	CVPR'24	0.867	0.830	0.736	0.815	0.880	0.826
CFM [38]	TIFS'24	-	0.828	-	0.758	0.915	-
Ours	-	**0.893**	**0.892**	**0.847**	**0.882**	**0.928**	**0.882**

Table 2: Cross-Manipulation Evaluation: ACC (%) and AUC (%) for Multi-Source Training and Testing.

Method	Venue	GID-DF (C23)		GID-DF (C40)		GID-F2F (C23)		GID-F2F (C40)	
		ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)
EfficientNet [59]	PMLR'19	82.40	91.11	67.60	75.30	63.32	80.10	61.41	67.40
ForensicTransfer [6]	Arxiv'18	72.01	-	68.20	-	64.50	-	55.00	-
Multi-task [44]	BTAS'19	70.30	-	66.76	-	58.74	-	56.50	-
F^3-Net [48]	ECCV'20	83.57	94.95	77.50	85.77	61.07	81.20	64.64	73.70
MLGD [27]	AAAI'18	84.21	91.82	67.15	73.12	63.46	77.10	58.12	61.70
LTW [56]	AAAI'21	85.60	92.70	69.15	75.60	65.60	80.20	65.70	72.40
DCL [57]	AAAI'22	87.70	94.90	75.90	83.82	68.40	82.93	67.85	75.07
M2TR [62]	ICML'22	81.07	94.91	74.29	84.85	55.71	76.99	66.43	71.70
Implicit [18]	CVPR'23	88.21	95.03	76.90	84.55	69.36	84.37	67.99	74.80
LSDA [64]	CVPR'24	88.63	95.24	77.45	85.23	70.21	84.79	68.11	75.35
Ours	-	**92.52**	**96.79**	**80.03**	**87.62**	**73.18**	**87.92**	**71.41**	**76.94**

Table 3: Cross-dataset evaluation results (Video-level AUC) with CLIP based method.

Method	Venue	CDF-v2	DFDC	DFD
Vanilla CLIP [49]	ICML'21	0.777	0.742	0.834
VLFFD [55]	arXiv'23	0.848	-	0.948
FFAA [23]	arXiv'24	-	0.740	0.920
FCG [16]	arXiv'24	0.923	0.812	-
RepDFD [33]	arXiv'24	0.899	0.810	-
CLIPping [25]	ICMR'24	-	0.719	0.866
Ours	-	**0.948**	**0.841**	**0.963**

4.3 Results on Multi-Dataset Protocols

While many researcher [50, 65, 67] propose diverse forgery methods from multiple sources, our protocol takes a more high-dimensional approach from the dataset perspective, offering meaningful contributions to our research community. The strength of our research community is that it embraces a diversity of viewpoints, which fosters the exchange and collision of ideas.

To investigate the feasibility of co-training from multiple datasets to improve the cross-dataset performance, we use multiple datasets from different sources for training. We also add the recently proposed methods that use stable diffusion-generated face data for exploration, which we believe are more in line with the real-world nature of forgery methods. Table 4 lists the results for four three-domain combinations (e.g., FF++ & Celeb-DF & DFF, FF++ & Celeb-DF & WDF, etc.) on all remaining unseen test sets.Information and visualization of these datasets can be found in Fig.**??**, they each have their own mark. As the training set diversifies across multiple domains, ours model's detection performance on these unseen domains—especially on DFDC, a large-scale and more realistic protorcol—improves notably. The key reason is that, during multi-domain training, the

Table 4: The results on the Multi-Domain Deepfake Detection Benchmarks based on M_{EER}(%) and AUC (%).

Method	FF++& Celeb-DF&DFF		FF++&Celeb-DF&WDF		FF++&DFF&WDF		Celeb-DF& DFF&WDF	
	M_{EER}(%)	AUC(%)	M_{EER}(%)	AUC(%)	M_{EER}(%)	AUC(%)	M_{EER}(%)	AUC(%)
MesoNet [1] (*WIFS* 2018)	45.31	53.34	44.70	69.25	46.42	57.16	40.54	54.92
Multl-task [45] (*BTAS* 2019)	36.33	57.33	36.98	66.03	39.14	74.72	34.91	62.17
F^3-Net [47] (*ECCV* 2020)	36.12	57.76	35.82	68.35	37.23	67.53	31.05	66.89
Xception [50] (*ICCV* 2021)	33.45	71.09	37.68	66.64	35.56	76.88	33.30	65.40
REECE [3] (*CVPR* 2022)	32.57	70.85	35.07	69.86	36.64	78.72	30.14	71.92
UCF [66] (*ICCV* 2023)	35.90	69.72	34.78	65.41	35.02	74.13	34.02	69.11
Implicit [18] (*CVPR* 2023)	33.09	69.54	37.66	72.12	38.91	74.10	33.18	68.31
Ours	**30.13**	**74.33**	**32.81**	**72.37**	**32.90**	**80.19**	**28.36**	**73.73**

Table 5: Ablation of each component on the protocol of FF++ & Celeb-DF & DFF to WDF.

ID	Baseline	DA	MIM	Meta-MoE	AUC	ACC
I	✓				73.45	71.07
II	✓	✓			74.36	71.65
III	✓		✓		75.11	73.33
IV	✓			✓	77.21	74.18
V	✓	✓	✓		75.71	73.42
VI	✓	✓		✓	75.12	74.39
GM-DF	✓	✓	✓	✓	**79.70**	**75.13**

model not only encounters the commonly used face-swap pipelines found in FF++ and Celeb-DF (e.g., FaceSwap) but also sees newly emerging forgery patterns in DFF or WDF (based on Stable Diffusion). Such exposure helps ours model learn more "cross-domain shared" forensic features, such as local texture inconsistencies, blurred facial boundaries, or high-frequency artifacts due to diffusion-based generation, thus reducing its reliance on domain-specific artifacts.

An interesting observation is that, although greater domain diversity typically yields better cross-dataset generalization, looking at the AUC (%) or M_EER (%) on individual domains may show fluctuations. On the one hand, multi-domain information leads to a positive transfer effect, helping the model discover more universally discriminative cues. On the other hand, certain forgery patterns may conflict with each other or dilute the learned "memory" of any single domain. Nevertheless, as shown in Table 4, our method maintains stable and superior results across all multi-domain combinations. For instance, when trained on FF++ & Celeb-DF & DFF, our M_EER reaches 30.13% and AUC hits 74.33%, outperforming the other methods by at least 2.44% in M_EER. Meanwhile, the combination Celeb-DF & DFF & WDF achieves our lowest M_EER (28.36%), further underscoring that our framework effectively "leverages strengths and avoids pitfalls" under large-scale, heterogeneous data, without losing critical deepfake cues to inter-domain discrepancies.

4.4 Ablation Study

Effectiveness of DA loss. In the **??** first and second rows **??** compared to baseline, the DA loss achieved an improvement of approximately 0. 91%, demonstrating the need for the alignment of the distributions of the two feature datasets through statistical features of higher order. We can observe a consistent improvement in performance when using the DA loss function,

Table 6: Natural language descriptions of the real and fake face used to train the model. BLIP Generate indicates that the BLIP [29] model generates descriptive information.

Prompt	Real Prompt	Fake Prompt
P1	A photo of real face	A photo of fake face
P2	This is a photo of real	This is a photo of fake
P3	{BLIP Generate} A photo of real face	{BLIP Generate} A photo of real face
P4	Real	Fake
P5	This is how a real face looks like	This is how a fake face looks like
P6	This photo contains real face	This photo contains fake face
P7	Real face is in this photo	Fake face is in this photo

Table 7: Impact of text prompts described in Table 6.

Method	FF + Celeb-DF→WDF		FF + DFF→WDF		Celeb-DF + DFF→WDF	
	AUC(%)	EER(%)	AUC(%)	EER(%)	AUC(%)	EER(%)
P1	63.08	**29.23**	**76.09**	**30.95**	**78.09**	**31.27**
P2	61.30	29.85	75.88	34.56	77.92	31.92
P3	61.37	31.81	69.93	32.03	74.25	34.87
P4	62.11	30.67	70.25	31.87	72.22	33.89
P5	**63.33**	30.35	74.40	33.30	77.18	31.46
P6	60.43	35.17	72.20	34.51	76.54	34.46
P7	61.10	33.29	72.43	34.69	78.12	33.15

which demonstrates the advantage of dataset alignment with the visual-linguistic pretraining model. **Effectiveness of MIM loss.** Comparing the first and third rows, it can be seen that the addition of the reconstruction module improves the AUC by 1.66% over the original model, indicating that the reconstructed features can effectively enhance the ability of fine-grained information extraction on the forged face.

Effectiveness of different text prompts. To validate the effect of different prompts on experimental performance, we introduced new templates into the prompts group. In Tables 6 & 7, the specific language descriptions of the real and fake face categories are shown. We scrutinize the impact of distinct text prompts on the model. In particular, varied texts exhibit commendable performance across diverse datasets with marginal differences. This substantiates the notion that text can effectively manifest dynamicized parameters in real-world contexts, thereby affirming our concept of stating dynamic affine transformations tailored to each dataset. An intriguing discovery emerges when using BLIP [29] to generate images with detailed descriptive information alongside the original combination of category images. Surprisingly, performance experiences relative degradation, potentially attributed to interference induced by category-independent prompts.

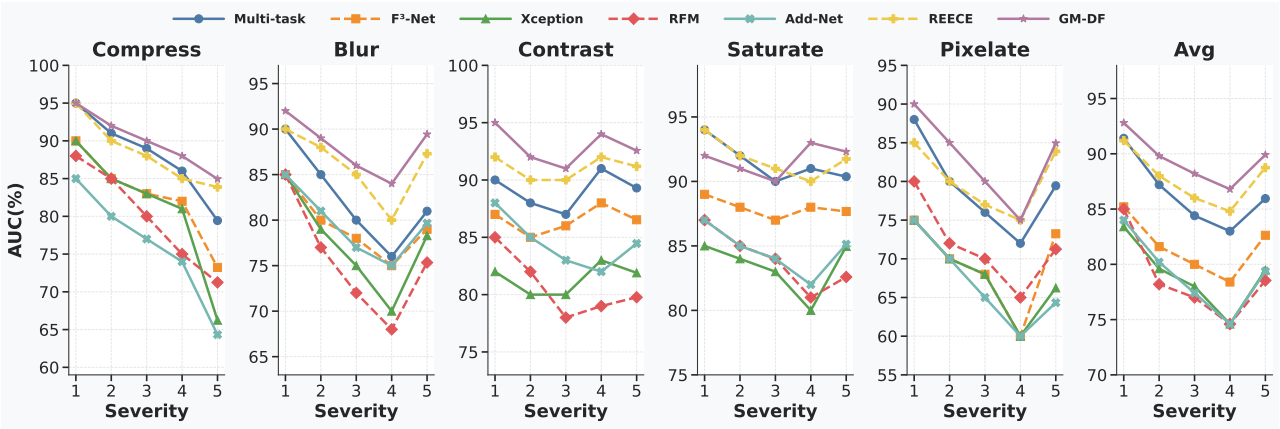

Figure 3: Robustness to various unseen corruptions. We report the Image-level AUC (%) of our methods under five different levels of seven particular types of corruption. "Average" denotes the mean across all corruptions at each severity level. Our GM-DF is more robust than previous methods for all corruptions.

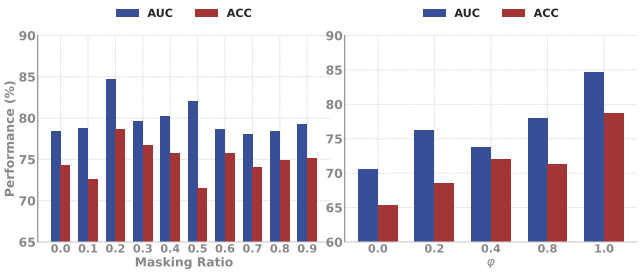

Figure 4: Ablation study of masking strategy and φ with AUC (%) of cross-dataset evaluation on DFDC.

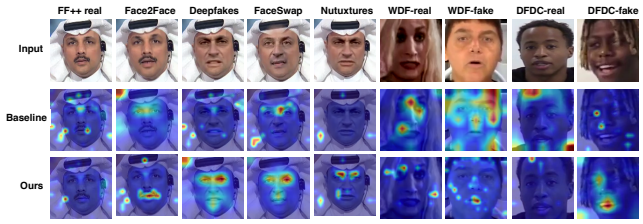

Figure 5: The model's attention is shown in a heatmap, where darker areas indicate more focus. The first column is the input image, the second shows CLIP [49] results, and the third shows our model's results.

Effectiveness of Meta-MoE. To quantify the importance of Meta-Moe module, we compare our text-based supervisory signals with meta learning and without two-stage learning. It can be seen from the fourth line that meta-MoE plays an important role in performance improvement (from 73.45% to 77.21%), which is mainly caused by learning the characteristic features of the domain. The mask-supervision method exhibits better generalization, suggesting that mask supervision alone can restrain overfitting to the training data. Moreover, unlike the only text backbone we improves steadily with more fine-grained supervision, which further confirms the scalability and versatility of multi-dataset learning.

Analysis of masking ratio. The quantitative results of the cross-dataset evaluation are shown in Figure 4. We observe that the minimum and randomized masking strategy achieves optimality under medium masking rates. Their performance is severely degraded as the masking rate is greater than 80%. Random masking strategies work best at 20% masking rate. This indicates that some important edges of the face may be damaged by using the random masking strategy.

Analysis of robustness against distortions. Considering the prevalence of image processing on the Web, we investigate the performance under several distortions proposed by [3, 45], namely image compression, Gaussian blurring, contrast dithering, saturation dithering, and pixelization. Quality-degraded images using different degradation methods are shown in Fig. 8. The results are shown in Fig. 3. We can see that our model is more robust to the listed ingressions than the existing methods. Both our method and previous methods are generally robust to compression, contrast, and saturation. However, in scenarios that blur and pixelate, the performance [3, 45, 47, 50, 61] is still much lower than the proposed method, indicating the robustness of the proposed method.

Visualization. We employed a joint training approach using three datasets FF++ [51], Celeb-DF (V2) [32], and DFF [54]. Subsequently, we conducted visual analyses on individual in-domain datasets as well as various cross-domain datasets. From Figure 5, it can be observed that directly merging datasets often leads the model to lose effective focus in challenging scenarios, such as WDF [76], where attention shifts to background regions. In contrast, our proposed multi-domain fusion model consistently concentrates on facial regions and successfully detects manipulated faces.

5 Conclusion

In this paper, we investigate the generalization capacity of deepfake detectors when trained on multi-dataset scenarios and propose a novel protocol for multiscenario training. We design a Generalized Multi-Scenario Deepfake Detection (GM-DF) framework to learn of both specific and common features across datasets. Using generic text representations to learn relationships between different datasets, we propose a novel meta-learning strategy to capture the relational information among datasets. In addition, GM-DF employs contrastive learning in image-text pairs to capture common characteristics of the data set and uses self-supervised mask relation learning to mask out partial correlations between regions during training. Extensive experiments demonstrate the superior generalization of our method.

Acknowledgments

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515140037), the Open Fund of National Engineering Laboratory for Big Data System Computing Technology (Grant No. SZU-BDSC-OF2024-02), the Guangdong Provincial Key Laboratory (Grant No. 2023B1212060076), the Guangdong Key Laboratory of Information Security Technology at Sun Yat-sen University, the Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing (Project No. 2024KCXTD047), the Science and Technology Development Fund of Macau (Project No. 0044/2024/AGJ), and the Science and Technology Foundation of Guangdong Province (Project No. 2024A0505090002). The computational resources were supported by the SongShan Lake HPC Center (SSL-HPC) at Great Bay University.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 1–7.

[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. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[3] 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*. 4113–4122.

[4] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. 2022. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18710–18719.

[5] Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Bo Zhang. 2021. Magdr: Mask-guided detection and reconstruction for defending deepfakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9014–9023.

[6] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. 2018. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510* (2018).

[7] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. 2021. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7373–7382.

[8] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. 2020. On the Detection of Digital Face Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5203–5212.

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

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*. 1126–1135.

[12] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*. PMLR, 3247–3258.

[13] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. 2024. Interpreting the second-order effects of neurons in clip. *arXiv preprint arXiv:2406.04341* (2024).

[14] Rui Gong, Dengxin Dai, Yuhua Chen, Wen Li, and Luc Van Gool. 2021. mDALU: Multi-source domain adaptation and label unification with partial datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8876–8885.

[15] David Güera and Edward J Delp. 2018. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 1–6.

[16] Yue-Hua Han, Tai-Ming Huang, Shu-Tzu Lo, Po-Han Huang, Kai-Lung Hua, and Jun-Cheng Chen. 2024. Towards More General Video-based Deepfake Detection through Facial Feature Guided Adaptation for Foundation Model. *arXiv preprint arXiv:2404.05583* (2024).

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16000–16009.

[18] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. 2023. Implicit Identity Driven Deepfake Face Swapping Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4490–4499.

[19] Shuai Huang, Yongxiong Wang, and Huan Luo. 2025. CCSUMSP: A cross-subject Chinese speech decoding framework with unified topology and multi-modal semantic pre-training. *Information Fusion* (2025), 103022.

[20] Shuai Huang, Yongxiong Wang, and Huan Luo. 2025. A dual-branch generative adversarial network with self-supervised enhancement for robust auditory attention decoding. *Engineering Applications of Artificial Intelligence* (2025), 111122.

[21] Shuai Huang, Yongxiong Wang, Huan Luo, Shuwen Jia, Han Chen, Chendong Qin, Zhongcai He, and Rui Luo. 2025. SSAAD: A Multi-Scale Temporal-Frequency Graph Network for Binary Auditory Attention Detection with Self-Supervised Learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[22] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. 2025. SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model.

[23] Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, and Wenming Yang. 2024. FFAA: Multimodal Large Language Model based Explainable Open-World Face Forgery Analysis Assistant. *arXiv preprint arXiv:2408.10072* (2024).

[24] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[25] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. 2024. CLIPping the Deception: Adapting Vision-Language Models for Universal Deepfake Detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*.

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

[27] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

[30] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5001–5010.

[31] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[32] 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*. 3207–3216.

[33] Kaiqing Lin, Yuzhen Lin, Weixiang Li, Taiping Yao, and Bin Li. 2024. Standing on the Shoulders of Giants: Reprogramming Visual-Language Model for General Deepfake Detection. *arXiv preprint arXiv:2409.02664* (2024).

[34] Xun Lin, Ajian Liu, Zitong Yu, Rizhao Cai, Shuai Wang, Yi Yu, Jun Wan, Zhen Lei, Xiaochun Cao, and Alex Kot. 2025. Reliable and Balanced Transfer Learning for Generalized Multimodal Face Anti-Spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[35] Xun Lin, Wenzhong Tang, Haoran Wang, Yizhong Liu, Yakun Ju, Shuai Wang, and Zitong Yu. 2024. Exposing image splicing traces in scientific publications via uncertainty-guided refinement. *Patterns* 5, 9 (2024).

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[37] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. 2021. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 772–781.

[38] Anwei Luo, Chenqi Kong, Jiwu Huang, Yongjian Hu, Xiangui Kang, and Alex C Kot. 2023. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security* (2023).

[39] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. 2021. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16317–16326.

[40] Saeed Masoudnia and Reza Ebrahimpour. 2014. Mixture of experts: a literature survey. *Artificial Intelligence Review* 42 (2014), 275–293.

[41] Changtao Miao, Qi Chu, Weihai Li, Suichan Li, Zhentao Tan, Wanyi Zhuang, and Nenghai Yu. 2021. Learning forgery region-aware and ID-independent features for face manipulation detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4, 1 (2021), 71–84.

[42] Changtao Miao, Zichang Tan, Qi Chu, Huan Liu, Honggang Hu, and Nenghai Yu. 2023. F 2 Trans: High-Frequency Fine-Grained Transformer for Face Forgery

Detection. *IEEE Transactions on Information Forensics and Security* 18 (2023), 1039–1051.

[43] Changtao Miao, Zichang Tan, Qi Chu, Nenghai Yu, and Guodong Guo. 2022. Hierarchical frequency-assisted interactive networks for face manipulation detection. *IEEE Transactions on Information Forensics and Security* 17 (2022), 3008–3021.

[44] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. 2019. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 1–8.

[45] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. 2019. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 1–8.

[46] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2307–2311.

[47] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*. Springer, 86–103.

[48] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*. Springer, 86–103.

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

[50] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1–11.

[51] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1–11.

[52] Aditya Sanghi, Rao Fu, Vivian Liu, Karl DD Willis, Hooman Shayani, Amir H Khasahmadi, Srinath Sridhar, and Daniel Ritchie. 2023. Clip-sculptor: Zero-shot generation of high-fidelity and diverse shapes from natural language. In *CVPR*. 18339–18348.

[53] 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 (CVPR)*.

[54] Haixu Song, Shiyu Huang, Yinpeng Dong, and Wei-Wei Tu. 2023. Robustness and Generalizability of Deepfake Detection: A Study with Diffusion Models. arXiv:2309.02218 [cs.CV]

[55] Ke Sun, Shen Chen, Taiping Yao, Haozhe Yang, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. 2023. Towards general visual-linguistic face forgery detection. *arXiv preprint arXiv:2307.16545* (2023).

[56] Ke Sun, Hong Liu, Qixiang Ye, Yue Gao, Jianzhuang Liu, Ling Shao, and Rongrong Ji. 2021. Domain general face forgery detection by learning to weight. In *Proceedings of the AAAI conference on Artificial Intelligence*, Vol. 35. 2638–2646.

[57] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. 2022. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2316–2324.

[58] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR.

[59] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 6105–6114.

[60] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems* 33 (2020), 6827–6839.

[61] Chengrui Wang and Weihong Deng. 2021. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14923–14932.

[62] 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*. 615–623.

[63] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. 2019. Towards universal object detection by domain attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7289–7298.

[64] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. 2024. Transcending Forgery Specificity with Latent Space Augmentation for Generalizable Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[65] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. 2024. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495* (2024).

[66] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. 2023. UCF: Uncovering Common Features for Generalizable Deepfake Detection. *arXiv preprint arXiv:2304.13949* (2023).

[67] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. 2023. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*.

[68] Guoqing Yang, Zhiming Luo, Jianzhe Gao, Yingxin Lai, Kun Yang, Yifan He, and Shaozi Li. 2024. A Multilevel Guidance-Exploration Network and Behavior-Scene Matching Method for Human Behavior Anomaly Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 5865–5873.

[69] Ziming Yang, Jian Liang, Yuting Xu, Xiao-Yu Zhang, and Ran He. 2023. Masked Relation Learning for DeepFake Detection. *IEEE Transactions on Information Forensics and Security* (2023).

[70] Qian Yu, Zong Ke, Guofu Xiong, Yu Cheng, and Xiaojun Guo. 2024. Identifying money laundering risks in digital asset transactions based on ai algorithms. *Available at SSRN 5129145* (2024).

[71] Yaning Zhang, Tianyi Wang, Zitong Yu, Zan Gao, Linlin Shen, and Shengyong Chen. 2024. MFCLIP: Multi-modal Fine-grained CLIP for Generalizable Diffusion Face Forgery Detection. *arXiv preprint arXiv:2409.09724* (2024).

[72] Yaning Zhang, Tianyi Wang, Zitong Yu, Zan Gao, Linlin Shen, and Shengyong Chen. 2025. MFCLIP: Multi-modal fine-grained CLIP for generalizable diffusion face forgery detection. *IEEE Transactions on Information Forensics and Security* (2025).

[73] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. 2021. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 15023–15033.

[74] Xiangyun Zhao, Samuel Schulter, Gaurav Sharma, Yi-Hsuan Tsai, Manmohan Chandraker, and Ying Wu. 2020. Object detection with a unified label space from multiple datasets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 178–193.

[75] Yijie Zhu, Yibo Lyu, Zitong Yu, Rui Shao, Kaiyang Zhou, and Liqiang Nie. 2025. EmoSym: A Symbiotic Framework for Unified Emotional Understanding and Generation via Latent Reasoning. In *Proceedings of the 33nd ACM International Conference on Multimedia*.

[76] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. 2020. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*. 2382–2390.