FSFM: A Generalizable Face Security Foundation Model via Self-Supervised Facial Representation Learning

Supplementary Material

Gaojian Wang[1,2], Feng Lin[1,2*], Tong Wu[1,2], Zhenguang Liu[1,2], Zhongjie Ba[1,2], Kui Ren[1,2]
[1]State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
{wolo, flin, cocotwu, liuzhenguang, zhongjieba, kuiren}@zju.edu.cn
Project Page: https://fsfm-3c.github.io

A. Overview

This supplementary material provides additional insights, details, and results to support our FSFM framework comprehensively, structured as follows:

- Facial Masking Strategies in Masked Image Modeling (MIM) (Sec. B): We delve into the impact of different facial masking strategies on naive MAE, including quantitative and qualitative analysis of attention differences.
- Instance Discrimination (ID) in FSFM (Sec. C): We highlight the connections and distinctions between our method and prior works that also integrate ID (or Siamese encoder architecture) into MIM (or degraded inputs).
- Implementation Details (Sec. D): Detailed descriptions of hyperparameters, pretraining, and finetuning settings.
- Additional Results and Comparisons (Sec. E): Extended experiments comparing FSFM against other models like ViT-based FAS and the base vision-language pertaining (VLP) model, CLIP.
- Ablations and Visualizations (Sec. F): Supplementary evidence validating FSFM's ability to learn robust and transferable facial representations.
- Limitations (Sec. G) of our work.

B. Revealing the secrets of facial masking strategies in MIM

In the main paper, we explore various facial masking strategies for masked image modeling (MIM) in FSFM, with additional visualizations provided in Fig. 1, and validate the effectiveness of CRFR-P masking through ablation studies on downstream face security tasks. However, a critical question remains: how do different facial masking strategies affect the MIM-pretrained model or its learned representations of real faces?

To address this, we quantitatively and qualitatively analyze the properties of attention maps. Given that most MIM-pretrained models, including ours, are built on the Vision Transformer (ViT) architecture [17], where the main component, the attention mechanism [46], is naturally interpretable [50]. Here, we adopt the naive MAE (the MIM network in FSFM) and follow its settings [23] with ViT-B/16 as the encoder and a 75% masking ratio. We conduct self-supervised pretraining on real face images from FF++_o [40] (the default dataset for our ablations). We alter only the masking strategy, encompassing simple random masking, Fasking-I, FRP, CRFR-R, and CRFR-P, and examine the following aspects of attention heads in the pretrained models: 1) mean attention distance to measure the flow of local and global facial information; 2) Kullback-Leibler (KL) divergence to investigate the diversity of attention; 3) visualized attention maps to uncover key regions of focus.

B.1. Local or global?

To explore whether the pretrained model attends to faces locally or globally, we calculate the mean attention distance [17] in each attention head across all blocks/layers, as shown in Fig. 2 (*Top*). The model (MAE ViT-B/16 encoder) pretrained with simple random masking tends to focus on local information in the lower blocks and shifts toward global attention in the deeper blocks, similar to the supervised model [17]. Fasking-I primarily aggregates global information as the visible patches predominantly consist of broad backgrounds and skin. FRP also causes large mean attention distances, but these are slightly smaller than those of Fasking-I, mainly because visible patches in FRP are more evenly distributed across all facial regions. CRFR-R fully masks a facial region before applying random masking, which encourages attention to different regions, con-

[*] Corresponding author

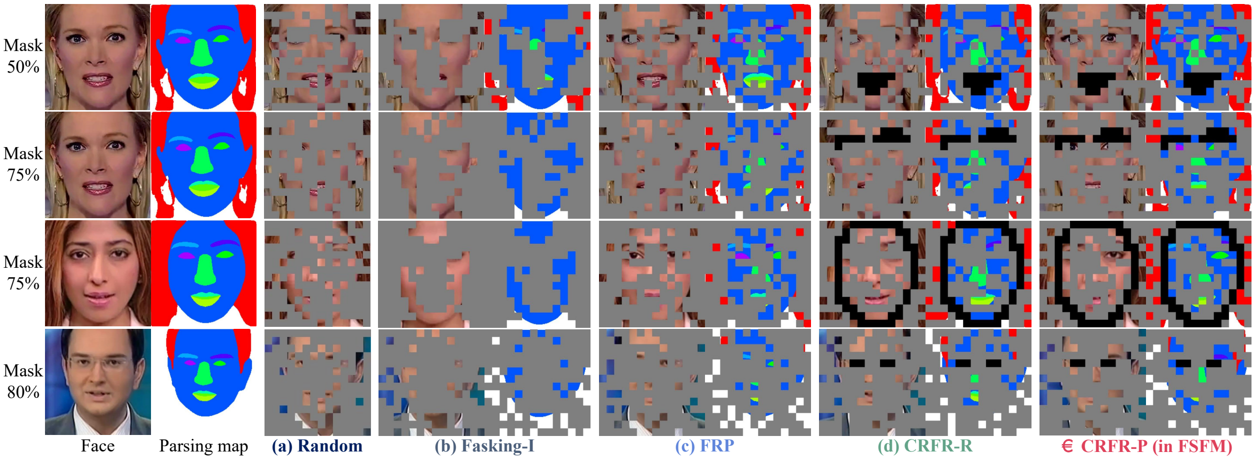


Figure 1. Additional visualizations of different facial masking strategies. (a) Random masking [23]. (b) Fasking-I adapted from [3]. (c) FRP: Facial Region Proportional masking. (d) CRFR-R: Covering a Random Facial Region followed by Random masking. (e) CRFR-P: Covering a Random Facial Region followed by Proportional masking.

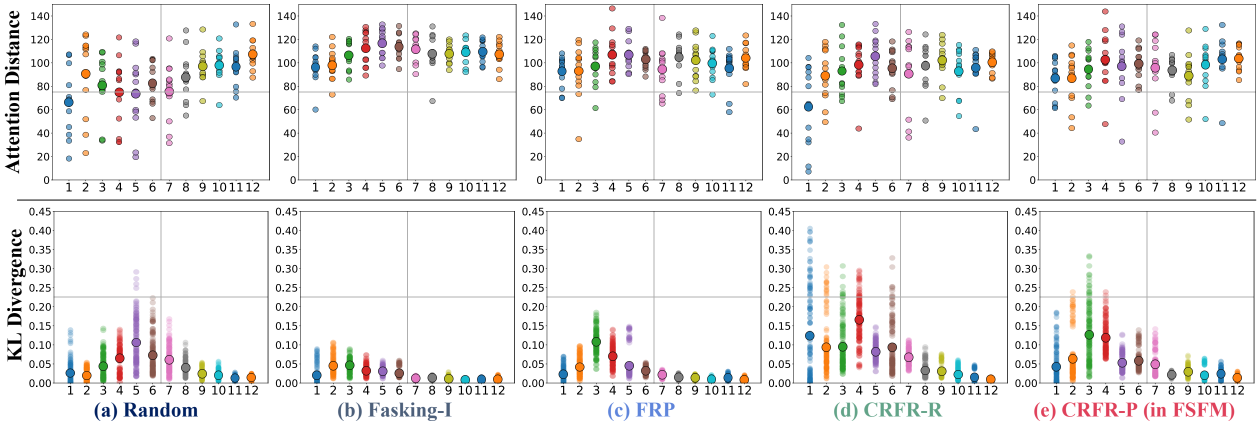


Figure 2. The mean attention distance (*Top*) and Kullback-Leibler divergence (*Bottom*) of each attention head (small dot) across all blocks (*x-axis*) in the ViT-B/16 encoder pretrained by MAE [23] with (a) Random, (b) Fasking-I, (c) FRP, (d) CRFR-R, and (e) CRFR-P masking strategies, along with the average one (large dot) for each block.

sequently resulting in more global attention in the middle (3rd to 8th) blocks compared to the simple random masking counterparts. Compared with CRFR-R, CRFR-P masks the remaining regions proportionally instead of randomly, making the 1st block more global w.r.t. the more unmasked regions. Compared with FRP, CRFR-P fully masks a region before applying proportional masking, which exposes more patches within other regions at the same masking ratio, thus leading to more local attention than FRP.

Overall, the model pretrained with CRFR-P exhibits well-distributed attention distances across all blocks, indicating a synergistic effect of FRP and CRFR-R, enabling appropriate attention to both local and global information.

B.2. Similar or different?

To assess whether the pretrained model focuses on similar or different tokens, we compute the Kullback-Leibler (KL) divergence between attention maps of each head across all blocks, following [50], as shown in Fig. 2 (*Bottom*). As the visible patches are mostly background and skin, the model pretrained with Fasking-I aggregates similar tokens, leading to low KL divergence across all attention heads. Interestingly, we find that proportional masking reduces diversity among attention heads, likely due to its homogeneous presentation of visible tokens, *i.e.*, derived from each facial region. In contrast, covering a random facial region increases attention diversity, as evidenced by higher KL divergence in CRFR-R versus the simple random masking counterparts and CRFR-P versus the FRP counterparts. This suggests that the model is compelled to look at different regions after fully masking a facial region.

Overall, the model pretrained with FRP lacks diversity across attention heads, while CRFR-R shows excessive diversity. Similar to the phenomenon observed in mean atten-

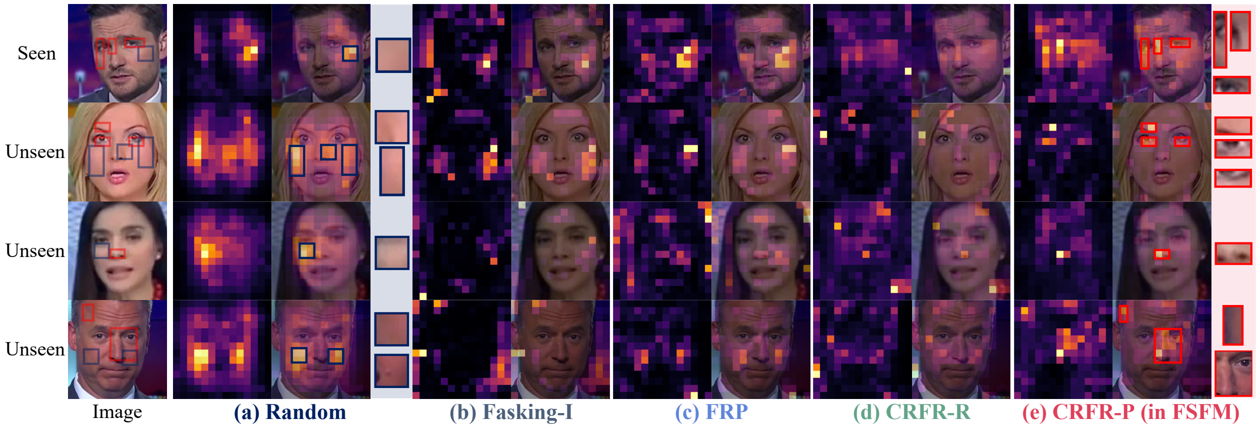
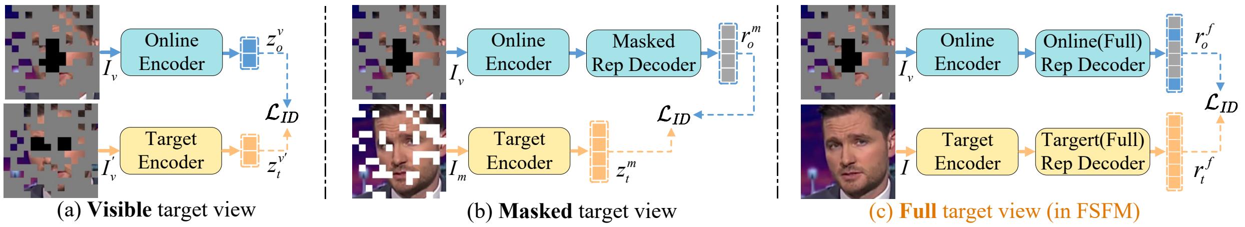


Figure 3. Visualization of the self-attention map averaged across all heads from the last block of the ViT-B/16 encoder pretrained by MAE [23] with (a) Random, (b) Fasking-I, (c) FRP, (d) CRFR-R, and (e) CRFR-P masking strategies. The rectangles in (a) Random and (e) CRFR-P highlight the regions of interest (ROI) for comparison. All faces, except for the first row, are from the test set and were unseen during pretraining.



(a) **Visible** target view (b) **Masked** target view (c) **Full** target view (in FSFM)

Figure 4. Comparison of typical target views and designs adapted for FSFM, derived from self-supervised pretraining methods that integrate both MIM and ID (or Siamese encoder architecture). (a) Visible patches from a different mask [2, 30, 47, 54]. (b) Masked patches from the same mask [9]. (c) Full patches without masking [1, 27, 45, 51, 55].

tion distance, CRFR-P strikes a balance in KL divergence across different heads, which also seems to act as a co-effect of FRP and CRFR-R counterparts, implicating appropriate attention to different key tokens.

B.3. Key regions?

To uncover which regions of real faces are critical to the pretrained model, we visualize the mean attention map of the last block and overlay it onto the input face in Fig. 3, as the pretext decoder or downstream head follows the last block. We can observe significant attention differences in salient regions across different masking strategies. At first glance, the attention regions from simple random masking appear to cover the entire face. However, it predominantly highlights the skin, which can be easily recovered from visible neighboring patches, while ignoring more challenging key regions. This suggests that the model solves face reconstruction through shortcuts instead of learning meaningful features. Similarly, the attention in Fasking-I is distributed across skin and background regions, as expected. While FRP and CRFR-R activate more attention areas, they still struggle to focus on key facial regions. In contrast, CRFR-P highlights attention across key regions like the eyes, eyebrows, and nose, indicating that the pretrained model tackles the challenge head-on: focusing on meaningful region-level features beyond just low-level pixels of real faces.

In summary, the CRFR-P masking strategy effectively directs attention to critical facial regions with appropriate range and diversity for both intra-region consistency and inter-region coherency, enabling the pretrained facial model to avoid trivial solutions (shortcuts) and capture the intrinsic properties of real faces. Furthermore, we hope this section provides new insights into fundamental face representation.

C. Connection and analysis of ID in FSFM

We illustrate the relation and distinction between our method and previous works that integrate ID (or Siamese encoder architecture) into MIM (or degraded input). While these hybrid approaches have demonstrated effectiveness in natural vision and face analysis, our empirical studies reveal that face security tasks necessitate more precise and reliable semantic alignment from the ID network. In response, we distinguish key designs such as target view & network structure, data augmentation, and loss function, which support local-to-global correspondences in FSFM.

C.1. Target view and network design

From the perspective of the input view, the online/student branch typically processes visible patches from the masked image, while the target/teacher branch varies across methods. Thus, we incorporate different target views and design paradigms into FSFM, as shown in Fig. 4. (a) Visible patches from a different mask [2, 30, 47, 54]: the online and target encoders produce latent features z_o^v and z_t^v for subsequent contrast learning. (b) Masked patches from the same mask [9]: to align the representation r_o^m of masked patches with the encoded target z_t^m, a masked representation (rep) decoder predicts r_o^m from the visible tokens z_o^v output by the online encoder. This decoder computes cross-attention between masked tokens (as Q) and full tokens (as K and V), following the latent regressor in CAE [9] and resembling the prompting decoder in [24]. (c) Full patches without masking [1, 27, 45, 51, 55]: some methods [1, 51, 55] are decoder-free designs that match visible online features z_o^v with full target features z_t^f to compute $\mathcal{L}_{ID}(z_o^v, z_t^f)$. Unlike CMAE [27], which introduces a feature/rep decoder after the online encoder, *i.e.*, $\mathcal{L}_{ID}(r_o^f, z_t^f)$, we add an additional target rep decoder to compute $\mathcal{L}_{ID}(r_o^f, r_t^f)$ in a disentangled representation space. *i.e.*, Siamese rep decoders. This design further reduces the gap in distribution between low-level pixel features and high-level semantic representations.

Based on our ablations (in the main paper) of downstream face security tasks, FSFM performs better when using full patches as the target view alongside Siamese rep decoders. By predicting the entire face representation from visible parts, the ID network aligns global and local views of the same face. In light of this, FSFM structures the encoded space with semantically complete facial representations through "local-to-global" correspondences [5], which endows the encoder with strong facial discriminability.

C.2. Data augmentation

Most ID methods [5–8, 20, 22] rely on strong data augmentations, including spatial and color transforms, to prevent trivial solutions. For MIM, applying strong augmentations such as color enhancements is suboptimal [23], as masking corruption itself effectively regularizes the pretext task. This is further evident in methods [27, 30, 51] that combine MIM and ID, where only simple augmentations—random size cropping or flipping—are applied to the masked input of the online branch, while strong or simple augmentations are used for the full (unmasked) target view.

In contrast, our FSFM behaves well without any data augmentation in either the online or target branches. This may stem from the semantic integrity preserved in unaugmented inputs, which benefits the learning of global face identity [47], especially in face security tasks where forgery and spoofing cues may be implicit anywhere. Additionally, the proposed CRFR-P masking strategy inherently introduces spatial variance tailored to facial structures, rendering simple augmentations (crop and flip) unnecessary. Consequently, FSFM processes only a single view per face image.

Base Model	Pretrain or Init	Intra-FF++ (c23)		Intra-FF++ (c40)	
		F-AUC	V-AUC	F-AUC	V-AUC
ViT-B [17]	Scratch	70.06	72.98	70.36	74.33
ViT-B [17]	Sup(IN)	97.30	98.80	88.65	93.22
MAEE [23] ViT-B	SSL(LN)	**98.64**	**99.61**	91.35	94.46
DINO [5] ViT-B	SSL(LN)	98.35	99.39	89.93	94.31
MCF [47] ViT-B	SSL(LFc)	97.84	99.19	89.81	94.20
FSFM ViT-B (Ours)	SSL(VF2)	97.74	99.03	**92.08**	**95.41**

Table 1. Intra-dataset evaluation of deepfake detection (DfD) on FF++ [40]. All base models are finetuned and tested on the c23 and c40 versions, respectively. **Best results**, second-best.

C.3. Loss function for ID

We consider two main types of loss functions for ID: contrastive [6, 8] and non-contrastive [7, 20]. Contrastive loss pulls positive views from the same sample together and pushes negative views from different samples apart. We use the widely adopted InfoNCE [38] as the contrastive loss. Non-contrastive loss maximizes the similarity of positive representations only. We use mean squared error (MSE) in BYOL [20] and negative cosine similarity (NCS) in SimSiam [7] as non-contrastive loss, respectively, but in an asymmetric formulation, as detailed in the main paper.

In FSFM, we observe that NCS outperforms the contrastive loss InfoNCE, even though methods combining MIM and ID [27, 30, 47, 51] favor the latter. We speculate this is because, in large-scale pretraining on real faces, the inter-image contrast introduced by negative sample pairs——pushing one real face away from others—does not help our model learn representations beneficial for face security tasks. Thus, we adopt asymmetric NCS by default to learn intra-face correspondences by matching the online anchor view with the target view of the same sample.

D. More implementation details

D.1. Pretraining settings

We set the mask ratio r to 0.75, similar to the baseline [23], as our ablation shows that this high ratio is also favorable for our FSFM. We do not use any data augmentation (not even crop and flip used in [23]) and only normalize the input faces during pretraining. We empirically set the loss weights λ_{fr} and λ_{cl} to 0.007 and 0.1, respectively. The projection and prediction heads are 2-layer MLPs following BYOL [20], with batch normalization (BN) replaced by layer normalization (LN) for our ViT-based architecture. The EMA momentum coefficient for updating the target branch starts from 0.996 and increases with a cosine scheduler, following BYOL [20]. We pretrain our model from scratch for 400 epochs on 4 NVIDIA RTX A6000

Method	Pretrain or Init	DG FAS Technique					OCI→M		OMI→C		OCM→I		ICM→O		Avg.	
		DM	AL	CL	ML	PL	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑
Base model																
ViT-B [17]*	Scratch						15.25	90.07	36.98	63.94	10.75	96.09	37.50	66.03	25.12	79.03
ViT-B [17]*	Sup(IN)						5.00	97.29	11.51	95.61	10.60	94.03	17.35	88.35	11.12	93.82
MAE [23] ViT-B*	SSL(LN)						8.17	97.37	27.91	77.90	18.02	91.54	25.36	80.76	19.86	86.89
DINO [5] ViT-B*	SSL(LN)						7.92	97.09	21.28	87.79	21.35	84.48	17.44	89.75	17.00	89.78
MCF [47] ViT-B*	SSL(LFc)						6.33	98.29	21.40	86.72	13.13	95.15	16.76	89.59	14.41	92.44
FSFM ViT-B (Ours)*	SSL(VF2)						6.58	97.43	**4.30**	**99.10**	14.63	90.96	10.02	**96.23**	8.88	95.93
ViT-based specialized method (Venue)																
ViTranZFAS [18] (IJCB'21)*	Init(IN)						10.95	95.05	14.33	92.10	16.64	85.07	15.67	89.59	14.40	90.45
TransFAS [48] (TBIOM'22)*	Init(IN)	✓					7.08	96.69	9.81	96.13	10.12	95.53	15.52	91.10	10.63	94.86
TTN-S [49] (TIFS'22)*	Init(IN)	✓					9.58	95.79	9.81	95.07	14.15	94.06	12.64	94.20	11.55	94.78
DiVT-V [32] (WACV'23)*	Init(IN)			✓	✓		10.00	96.64	14.67	93.08	**5.71**	97.73	18.06	90.21	12.11	94.42
TTDG-V [56] (CVPR'24)*	Init(IN)			✓			**4.16**	**98.48**	7.59	98.18	9.62	**98.18**	**10.00**	96.15	**7.84**	**97.75**

Abbreviation: Sup-Supervised SSL-Self-Supervied Init-weight initialization IN-ImageNet DG-Domain Generalization

DG FAS Technique: DM-Depth Maps AL-Adversarial Learning CL-Contrastive Learning (or triplet, similarity loss) ML-Meta Learning PL-Prototype Learning

Table 2. Cross-domain evaluation on face anti-spoofing (FAS) using visual-only ViT-based models* without including the supplementary data in [26]. For a fair comparison, the results of specialized methods are cited from the original papers. **Best results**, second-best.

Method	Pretrain or Init	Train Set	Test Set **Video-level** AUC(%)↑				Avg. ΔOurs
			CDFV2	DFDC	DFDCP	WDF	
Base model							
Xception [12]	Sup(IN)	FF++	76.39	70.62	72.24	76.11	14.0↑
EfficientNet-B4 [44]	Sup(IN)	FF++	79.81	71.85	66.95	76.42	14.1↑
ViT-B [17]	Scratch	FF++	64.08	66.73	72.62	60.36	21.9↑
ViT-B [17]	Sup(IN)	FF++	86.24	74.48	82.11	81.20	6.9↑
MAE [23] ViT-B	SSL(IN)	FF++	79.51	75.93	87.10	80.96	7.0↑
DINO [5] ViT-B	SSL(IN)	FF++	80.47	76.90	84.64	82.06	6.9↑
MCF [47] ViT-B	SSL(LFc)	FF++	80.25	73.61	82.55	79.79	8.8↑
CLIP [39] ViT-B	VLP(WIT)	FF++	78.95	73.83	82.38	78.60	9.5↑
FSFM ViT-B (Ours)	SSL(VF2)	FF++	**91.44**	**83.47**	**89.71**	**86.96**	-

Method	Pretrain or Init	Train Set	Test Set **Frame-level** AUC(%)↑				Avg. ΔOurs
			CDFV2	DFDC	DFDCP	WDF	
Base model							
Xception [12]	Sup(IN)	FF++	69.52	68.20	68.94	68.83	15.1↑
EfficientNet-B4 [44]	Sup(IN)	FF++	73.37	69.47	64.37	71.95	14.2↑
ViT-B [17]	Scratch	FF++	61.14	64.27	69.00	60.68	20.2↑
ViT-B [17]	Sup(IN)	FF++	77.43	71.09	74.07	75.86	9.4↑
MAE [23] ViT-B	SSL(IN)	FF++	72.64	72.18	79.81	73.94	9.4↑
DINO [5] ViT-B	SSL(IN)	FF++	73.88	72.78	77.31	75.08	9.2↑
MCF [47] ViT-B	SSL(LFc)	FF++	73.16	69.63	75.78	74.10	10.8↑
CLIP [39] ViT-B	VLP(WIT)	FF++	73.02	70.66	77.46	72.04	10.7↑
FSFM ViT-B (Ours)	SSL(VF2)	FF++	**85.05**	**80.20**	**85.50**	**85.26**	-

Abbreviation: Sup-Supervised SSL-Self-Supervied VLP-Vision Language pretraining Init-weight initialization

Dataset: IN/1M natural images [14] LFc/20M facial images [47] WIT/400M (natural image, text) pairs [39] VF2/3M facial images [4]

Table 3. Cross-dataset evaluation on deepfake detection (DfD), *adding CLIP [39] ViT-B/16 image encoder as a base model*. Left: video-level AUC, Right: frame-level AUC. All base models are finetuned on FF++ (c23) and tested on unseen datasets. Avg.ΔOurs denotes the average AUC difference between our FSFM and other methods. **Best results**, second-best.

GPUs. Other settings follow the defaults in MAE [23]: we use the AdamW [36] optimizer with momentum $\beta_1 = 0.9$, $\beta_2 = 0.95$; we apply the linear lr scaling rule [19] with a base learning rate of 1.5e-4; we adopt a cosine decay [37] learning rate schedule with a warmup epoch of 40; we maintain the effective batch size as 4096 = 256 (batch size per GPU) × 4 (GPUs) × 4 (accumulated gradient iterations).

D.2. Finetuning settings in downstream tasks

For finetuning the ViTs from FSFM and other pretrained models, we adopt identical settings except for weight initialization, detailed below.

Deepfake Detection We use the c23 (HQ) version of FF++ [40] with official train/val splits for finetuning. We sample 128 frames per real video (the original YouTube subset) and 32 frames per forgery video (including Deepfakes, Face2Face, FaceSwap, and NeuralTextures subsets). We follow the official test split in other unseen datasets for testing, including CDFV2 [31], DFDC [16], DFDCp [15], and WDF [57]. We sample 32 frames per testing video. We use DLIB [29] to extract faces (without alignment and parsing) and resize them to 224×224. As WDF already provides 224×224 facial images, we directly use its test set without processing. We add only one linear layer as the binary classifier after averaging all non-[CLS] token features. We set the batch size to 64, the base learning rate to 2.5e-4, and the finetuning epochs to 10 (50 for ViT-B Scratch). Other settings adhere to the MAE ImageNet finetuning recipe [23].

Face Anti-Spoofing In the main paper, we adopt the 0-shot MCIO setting (Protocol 1) in [26] and include CelebA-Spoof [52] as supplementary data for FAS finetuning. We set the batch size to 12 for each training domain. We append the MLP head after averaging all non-[CLS] token features instead of using [CLS] ones [26], to align with other face security tasks. Additionally, for a fair comparison with other visual-only ViT-based methods, we additionally follow [56] and report the best performance without including the supplementary data, as presented in Sec. E.2.

Diffusion Face Forgery Detection For the training set, we sample 32 frames from each real video (the original YouTube subset) and each forgery video (the Deepfakes subset) from FF++ (c23) [40]. For validation and testing sets, we follow the splits provided by the DiFF benchmark [10]. We use DLIB [29] to extract faces (without alignment and parsing) and resize them to 224×224. We add one linear layer as the binary classifier after averaging all non-[CLS] token features. We set the batch size to 256, the base learning rate to 5e-4, and the finetuning epochs to 50. Other settings adhere to the MAE ImageNet finetuning recipe [23].

Method	Pretrain or Init	OCI→M		OMI→C		OCM→I		ICM→O		Avg.	
		HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑
Base model											
ViT-B [17]	Scratch	15.37	90.73	35.37	68.23	14.75	94.18	31.65	71.55	24.28	81.17
ViT-B [17, 26]	Sup(IN)	**3.52**	98.74	**2.42**	**99.52**	8.45	96.91	11.86	94.62	6.56	97.44
MAE [23] ViT-B	SSL(LN)	10.32	94.87	15.91	89.96	15.54	91.13	16.51	90.29	14.57	91.56
DINO [5] ViT-B	SSL(LN)	6.73	97.15	13.44	93.90	14.27	93.56	15.55	90.99	12.50	93.90
MCF [47] ViT-B	SSL(LFc)	4.00	98.84	8.46	96.90	8.02	97.39	10.70	95.64	7.80	97.19
CLIP [39] ViT-B	VLP(WIT)	4.29	98.76	5.00	98.89	7.14	97.92	**6.09**	**98.12**	5.63	98.42
FSFM ViT-B (Ours)	SSL(VF2)	3.78	**99.15**	3.16	99.41	**4.63**	**99.03**	7.68	97.11	**4.81**	**98.68**

Abbreviation: Sup-Supervised SSL-Self-Supervied VLP-Vision Language pretraining Init-weight initialization
Dataset: IN-ImageNet1K [14] LFc-LAION FACE cropped [47] WIT-WebImageText [39] VF2-VGGFace2 [4]

Table 4. Cross-domain evaluation on face anti-spoofing (FAS), *adding CLIP [39] ViT-B/16 image encoder as a base model*. **Best**, second.

Method	Pretrain or Init	Test Subset (AUC%↑)					Avg. w/o FF++
		FF++	T2I	I2I	FS	FE	
ViT-B [17]	Scratch	92.02	62.19	69.99	60.87	67.30	65.09
ViT-B [17]	Sup(IN)	99.15	33.38	35.83	52.20	55.42	44.21
MAE [23] ViT-B	SSL(IN)	99.25	33.01	32.88	47.77	58.70	43.09
DINO [5] ViT-B	SSL(IN)	99.30	33.85	36.02	60.37	63.18	48.35
MCF [47] ViT-B	SSL(LFc)	**99.39**	39.09	38.67	34.35	56.02	42.03
CLIP [39] ViT-B	VLP(WIT)	99.33	**69.63**	66.25	65.23	57.07	64.54
FSFM ViT-B	SSL(FF++_o)	99.31	61.74	**71.91**	**71.31**	**78.98**	**70.99**

Table 5. Cross-dataset evaluation on DiFF benchmark [11]. *adding CLIP [39] ViT-B/16 image encoder as a base model*. All base models are finetuned only on the FF++_DeepFake (c23) [40]. **Best results**, second-best.

E. Additional experimental results

E.1. Comparison of intra-dataset DfD on FF++

While our primary objective focuses on cross-domain generalization for real-world applicability, we provide an intra-dataset evaluation of deepfake detection (DfD) on FaceForensics++ (FF++) [40], as presented in Tab. 1. The metrics on the FF++_c23 show that FSFM maintains comparable intra-set performance while significantly improving cross-dataset generalization. Moreover, when evaluated on the more challenging high-compression (c40) version, FSFM outperforms all baseline vision models, further demonstrating its robustness.

E.2. Comparison with ViT-based FAS

In a fair comparison with visual-only ViT-based face anti-spoofing (FAS) methods, our method also significantly outperforms all base models, as shown in Tab. 2. FSFM surpasses most counterparts and ranks second in average metrics. TTDG-V [56], which introduces test-time domain generalization and explicit optimization goals for FAS, performs better than ours in two out of four target domains (OCI→M and OCM→I). While optimizing for a specific downstream task is beyond the scope of this study, incorporating special auxiliary supervision or domain generalization (DG) techniques into our pretrained model may further improve its generalization ability for face presentation attack detection.

E.3. Comparison with CLIP

Another line of representation learning, vision-language pretraining (VLP), particularly contrastive language-image pretraining (CLIP) [39], has shown remarkable zero-shot and generalization capabilities across diverse downstream tasks. Recent studies have successfully tailored CLIP to specific face security tasks, including deepfake detection [28, 41, 43], face anti-spoofing [21, 25, 34, 35, 42], and diffusion forgery detection [13, 33, 53]. These text-aided methods differ fundamentally from our FSFM, which is vision-only, self-supervised, and task-agnostic. Moreover, VLP demands extensive (image, text) data pairs along with significant computing resources for the additional text encoder. Despite these, we include CLIP as a base vision-language model (VLM) for comparison.

Specifically, we borrow the CLIP image encoder, also a ViT-B but pretrained on the WIT dataset with 400M image& text pairs, and finetune it on downstream face security tasks under the same settings as other base models. We supplement the corresponding results on deepfake detection, face anti-spoofing, and diffusion face forgery detection in Tab. 3, Tab. 4, and Tab. 5, respectively. We can observe that CLIP ViT-B transfers better than other base vision models on FAS and DiFF tasks, benefiting from the extensive data scale of multi-modal supervision. However, directly applying CLIP ViT-B to DfD exhibits inferior generalization. In contrast, our proposed FSFM consistently outperforms CLIP ViT-B across downstream face security tasks.

F. More ablations and visualizations

F.1. Ablation studies

This subsection presents additional ablations. Unless otherwise stated, the default settings follow the main paper.

Effect of Masking Ratio r We also examine the impact of different masking ratios for CRFR-P masking on our pretraining framework. As shown in Tab. 6, FSFM achieves the best overall performance with a 0.75 masking ratio. Adopting lower masking ratios leads to trivial reconstruction and alignment tasks due to more available information. Conversely, using a higher masking ratio makes pretext tasks too challenging to learn sufficient facial representations for

Component	Setting	Deepfake Detection		Face Anti-spoofing	
		F-AUC↑	V-AUC↑	HTER↓	AUC↑
Masking ratio r	0.35	73.92	79.56	15.84	90.02
	0.50	74.31	79.48	21.57	84.09
	0.65	74.92	80.13	17.76	87.37
	0.75	**76.39**	**82.31**	**17.44**	**88.26**
	0.85	75.40	80.83	19.19	86.13
Pretraining model	Size				
Model size (parameters)	ViT-S/16(22M)	74.80	80.20	19.32	89.13
	ViT-B/16(86M)	76.39	82.31	17.44	88.26
	ViT-L/16(303M)	**77.43**	**83.15**	**16.23**	**93.13**

Table 6. Ablations on deepfake detection (DfD) and face anti-spoofing (FAS) with average metrics. The model is pretrained on FF++_o [40]. Default settings are shaded in gray.

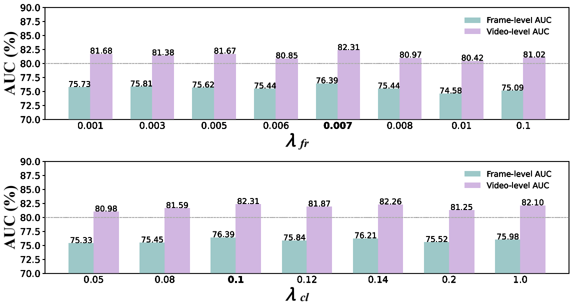


Figure 5. Ablations of loss weights on deepfake detection (DfD) with average metrics. The model is pretrained on FF++_o [40]. Default: $\lambda_{fr} = 0.007$, $\lambda_{cl} = 0.1$.

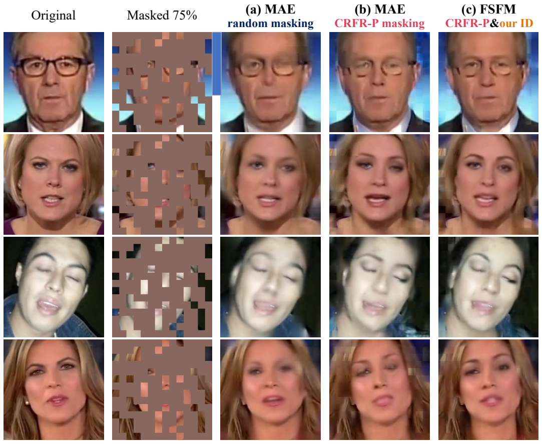


Figure 6. **Reconstruction Visualization** of real face images with a masking ratio of 75%, using MIM models pretrained from: (a) a naive MAE with simple random masking [23], (b) a naive MAE with our CRFR-P masking, and (c) our FSFM. All models were pretrained on the train and validation sets of FF++_o [40] without adversarial learning, for 400 epochs. Images are from the test set.

downstream face security tasks. Accordingly, we select a 75% masking ratio as the default setting.

Effect of Model Scaling Tab. 6 also shows that FSFM benefits from larger model sizes when pretrained on FF++_o. The transfer performance on downstream face security tasks improves as the model scales up. Due to limited computing resources, we were unable to pretrain larger models on more face images, *e.g.*, pertaining ViT-L/16(303M) on the full VGGFace2 dataset, but we aspire to explore this in future work and update model zoos accordingly.

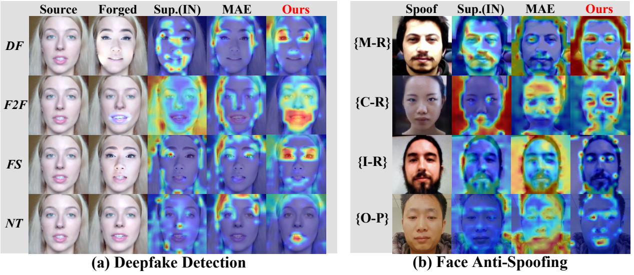


Figure 7. **CAM Visualization.** (a) DfD on various manipulations from FF++ [40]. (b) FAS on the MCIO protocol. FSFM highlights forgery artifacts and spoof clues. Images are from the test set.

Effect of Loss Weight To explore the impact of the reconstruction and distillation losses, we vary various loss weights, *i.e.* λ_{fr} and λ_{cl}. Results in Fig. 5 show that the configuration ($\lambda_{fr} = 0.007$, $\lambda_{cl} = 0.1$) performs better on challenging cross-dataset DfD.

F.2. Visualizations

Reconstruction To demonstrate the superiority of the facial representations pretrained with FSFM, we further follow MAE [23] to visualize reconstruction results, as shown in Fig. 6. We can see that FSFM demonstrates better reconstruction quality concerning intra-region consistency (preserving fine-grained textures within facial regions), inter-region coherency (maintaining spatial relationships across regions), and local-to-global correspondence (aligning local appearance with global facial looking).

CAM We provide additional CAM visualizations in Fig. 7, which are consistent with the observations in the main paper, further substantiating the effectiveness of our method.

G. Limitations

Despite the promising results demonstrated by FSFM across various face security tasks, our work has several limitations that warrant further exploration: **Pretraining Dataset Bias** FSFM is pretrained on large-scale facial images, and its performance can be affected by the quantity, diversity, and quality of the pretraining data. Pretraining on specific datasets like VGGFace2 [4] may inherit their biases (*e.g.*, race, ethnicity, and age), potentially reducing fairness. **Absence of Multi-modal Learning** Since our work focuses on general visual face security, the current framework processes only image or frame data for downstream forgery&spoofing image&video detection, ignoring potential complementary signals (*e.g.*, audio inconsistencies in deepfakes or physiological cues in spoofing), which could further enhance capabilities.

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pages 456–473. Springer, 2022. 3, 4

[2] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan L Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24256–24265, 2023. 3, 4

[3] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. Marlin: Masked autoencoder for facial video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1493–1504, 2023. 2

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 5, 6, 7

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 4, 5, 6

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 4

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 4

[9] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223, 2024. 3, 4

[10] Harry Cheng, Yangyang Guo, Tianyi Wang, Liqiang Nie, and Mohan Kankanhalli. Diffusion facial forgery detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 5939–5948, New York, NY, USA, 2024. Association for Computing Machinery. 5

[11] Harry Cheng, Yangyang Guo, Tianyi Wang, Liqiang Nie, and Mohan Kankanhalli. Diffusion facial forgery detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5939–5948, 2024. 6

[12] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 5

[13] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366, 2024. 6

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 6

[15] B Dolhansky. The dee pfake detection challenge (dfdc) pre view dataset. *arXiv preprint arXiv:1910.08854*, 2019. 5

[16] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 5

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1, 4, 5, 6

[18] Anjith George and Sébastien Marcel. On the effectiveness of vision transformers for zero-shot face anti-spoofing. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 5

[19] P Goyal. Accurate, large minibatch sg d: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 5

[20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 4

[21] Jiabao Guo, Huan Liu, Yizhi Luo, Xueli Hu, Hang Zou, Yuan Zhang, Hui Liu, and Bo Zhao. Style-conditional prompt token learning for generalizable face anti-spoofing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 994–1003, 2024. 6

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 2, 3, 4, 5, 6, 7

[24] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049*, 2022. 4

[25] Xueli Hu, Huan Liu, Haocheng Yuan, Zhiyang Fu, Yizhi Luo, Ning Zhang, Hang Zou, Jianwen Gan, and Yuan Zhang. Fine-grained prompt learning for face anti-spoofing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7619–7628, 2024. 6

[26] Hsin-Ping Huang, Deqing Sun, Yaojie Liu, Wen-Sheng Chu, Taihong Xiao, Jinwei Yuan, Hartwig Adam, and Ming-Hsuan Yang. Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In *European conference on computer vision*, pages 37–54. Springer, 2022. 5, 6

[27] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3, 4

[28] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1006–1015, 2024. 6

[29] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 5

[30] Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2142–2152, 2023. 3, 4

[31] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020. 5

[32] Chen-Hao Liao, Wen-Cheng Chen, Hsuan-Tung Liu, Yi-Ren Yeh, Min-Chun Hu, and Chu-Song Chen. Domain invariant vision transformer learning for face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6098–6107, 2023. 5

[33] Li Lin, Irene Amerini, Xin Wang, Shu Hu, et al. Robust clip-based detector for exposing diffusion model-generated images. *arXiv preprint arXiv:2404.12908*, 2024. 6

[34] Ajian Liu, Shuai Xue, Jianwen Gan, Jun Wan, Yanyan Liang, Jiankang Deng, Sergio Escalera, and Zhen Lei. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 222–232, 2024. 6

[35] Si-Qi Liu, Qirui Wang, and Pong C. Yuen. Bottom-up domain prompt tuning for generalized face anti-spoofing. In *Computer Vision – ECCV 2024*, pages 170–187, Cham, 2025. Springer Nature Switzerland. 6

[36] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[37] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5, 6

[40] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 1, 4, 5, 6, 7

[41] Stefan Smeu, Elisabeta Oneata, and Dan Oneata. Declip: Decoding clip representations for deepfake localization. *arXiv preprint arXiv:2409.08849*, 2024. 6

[42] Koushik Srivatsan, Muzammal Naseer, and Karthik Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19685–19696, 2023. 6

[43] Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. *arXiv preprint arXiv:2408.09647*, 2024. 6

[44] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5

[45] Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2132–2141, 2023. 3, 4

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[47] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Liang Liu, Yabiao Wang, and Chengjie Wang. Toward high quality facial representation learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5048–5058, 2023. 3, 4, 5, 6

[48] Zhuo Wang, Qiangchang Wang, Weihong Deng, and Guodong Guo. Face anti-spoofing using transformers with relation-aware mechanism. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):439–450, 2022. 5

[49] Zhuo Wang, Qiangchang Wang, Weihong Deng, and Guodong Guo. Learning multi-granularity temporal characteristics for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 17:1254–1269, 2022. 5

[50] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14475–14485, 2023. 1, 2

[51] Kun Yi, Yixiao Ge, Xiaotong Li, Shusheng Yang, Dian Li, Jianping Wu, Ying Shan, and Xiaohu Qie. Masked image modeling with denoising contrast. *arXiv preprint arXiv:2205.09616*, 2022. 3, 4

[52] Yuanhan Zhang, ZhenFei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 70–85. Springer, 2020. 5

[53] Yaning Zhang, Tianyi Wang, Zitong Yu, Zan Gao, Linlin Shen, and Shengyong Chen. Mfclip: Multi-modal fine-grained clip for generalizable diffusion face forgery detection. *arXiv preprint arXiv:2409.09724*, 2024. 6

[54] Zhiyu Zhao, Bingkun Huang, Sen Xing, Gangshan Wu, Yu Qiao, and Limin Wang. Asymmetric masked distillation for pre-training small foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18516–18526, 2024. 3, 4

[55] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 3, 4

[56] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Shouhong Ding, and Lizhuang Ma. Test-time domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–187, 2024. 5, 6

[57] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2382–2390, 2020. 5