BIG-MOE: BYPASS ISOLATED GATING MOE FOR GENERALIZED MULTIMODAL FACE ANTI-SPOOFING

Yingjie Ma[1,2], Zitong Yu[2,3], Xun Lin[2], Weicheng Xie[1,4], Linlin Shen[1,3,4*]*

[1]College of Computer Science and Software Engineering, Shenzhen University [2]Great Bay University
[3]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
[4]Guangdong Provincial Key Laboratory of Intelligent Information Processing

ABSTRACT

In the domain of facial recognition security, multimodal Face Anti-Spoofing (FAS) is essential for countering presentation attacks. However, existing technologies encounter challenges due to modality biases and imbalances, as well as domain shifts. Our research introduces a Mixture of Experts (MoE) model to address these issues effectively. We identified three limitations in traditional MoE approaches to multimodal FAS: (1) Coarse-grained experts' inability to capture nuanced spoofing indicators; (2) Gated networks' susceptibility to input noise affecting decision-making; (3) MoE's sensitivity to prompt tokens leading to overfitting with conventional learning methods. To mitigate these, we propose the Bypass Isolated Gating MoE (BIG-MoE) framework, featuring: (1) Fine-grained experts for enhanced detection of subtle spoofing cues; (2) An isolation gating mechanism to counteract input noise; (3) A novel differential convolutional prompt bypass enriching the gating network with critical local features, thereby improving perceptual capabilities. Extensive experiments on four benchmark datasets demonstrate significant generalization performance improvement in multimodal FAS task. The code is released at https://github.com/murInJ/BIG-MoE.

Index Terms— Face Anti-Spoofing, Multimodal, Prompt Learning, Mixture of Experts

1. INTRODUCTION

Face Recognition (FR) technology, celebrated for its efficiency and accuracy in applications such as security surveillance and mobile payments, now confronts escalating security threats from sophisticated face rendering attacks. Traditional FR systems struggle to discern these attacks, which include printed photos, video playback, and 3D masks, underscoring the urgent need for robust security measures.

To counter these threats, the research community has turned to Face Anti-Spoofing (FAS) techniques, which differentiate between genuine and spoofed faces [3]. Multimodal FAS methods [4, 5, 6, 7, 2, 8, 9], integrating information from RGB images, depth maps, and infrared images, have

_____
* Corresponding authors

Fig. 1: **Existing MoE prompt learning paradigm vs. Ours.** Both (a) conventional MoE prompt learning and (b) the parameter-efficient expert retrieval [1] approaches input prompt and feature tokens into the gating network or product key gating (PK Gate) network to generate scores for expert selection and subsequent processing with different gating mechanisms and types of experts. (c) Our Isolated Gating Mechanism (IGM) concatenates prompt and feature tokens for gating network scoring, and then processes feature tokens exclusively, isolating expert network input to enhance noise resilience and processing precision.

shown promise in capturing comprehensive physical and behavioral facial features [10]. However, the integration of multimodal data is challenging, often hampered by intermodal feature bias and imbalance, and subtle multimodal spoof clues are easily drowned by domain shifts caused by sensor/environment discrepancy [11, 12].

The Mixture of Experts (MoE) model, adept at handling complex data distributions, decomposes a large network into specialized smaller networks, reducing computational load through sparse activation and enhancing model generalization [13, 14, 15, 16]. This architecture excels in multi-task and multi-modal learning scenarios, especially with high dimensional and heterogeneous data [17]. MoE has also shown excellent results for sparse representations in FAS tasks [18, 19, 20]. Building on this, our research integrates fine-grained experts [1] into the MoE framework for multimodal FAS tasks, improving the capture of detailed data features crucial for FAS performance [15, 16]. To counteract the vulnerability to input noise, we propose an Isolation Gating Mechanism, depicted in Fig. 1, which processes input vectors to robustly

Fig. 2: **BIG-MoE Framework Overview**: The diagram succinctly captures the essential process and components of our approach: (a) **Prompt Generation**: This step outlines the creation and integration of initial prompts. (b) **CPB**: Describes the Convolutional Prompt Bypass, focusing on its enhancement of feature extraction via Central Difference Convolution (CDC) [2] and multimodal prompt integration. (c) **IGMA**: Highlights the Isolated Gating Mechanism Adapter's role in gating and its interaction with CPB across layers, promoting information exchange for enhanced model performance and robustness.

understand feature relationships without direct fitting.

Furthermore, we explore the integration of prompt learning [21] within the MoE model. To maximize prompt learning's benefits while avoiding feature confusion, we introduce a Convolutional Prompt Bypass (CPB) that interacts with the gated network to capture local deception features without direct feature propagation. Our contributions are threefold:

- We proposed the BIG-MoE, a novel multimodal FAS architecture that pioneers the application of MoE with fine-grained experts. This pioneering approach allows for more effective extraction of subtle cues and integration of multimodal features.

- The BIG-MoE framework features an Isolated Gating Mechanism to shield the model against input noise and includes a convolutional prompt bypass, which fortifies the gating network with essential cues, thereby enhancing the model's robustness against overfitting and noise.

- Extensive experiments demonstrate the reliability and superior performance of BIG-MoE for generalized multimodal FAS.

2. METHODOLOGY

As shown in Fig. 2, our proposed Bypass Isolated Gating MoE (BIG-MoE) framework is fundamentally composed of a pre-trained Vision Transformer (ViT), coupled with a sophisticated prompt generation module, the Convolutional Prompt Bypass (CPB), and the Isolated Gating Mechanism Adapter (IGMA). Input data is transformed into visual prompt tokens by the prompt generation module, which are then enhanced by the CPB module. Concurrently, the input is processed through the ViT Encoder and IGMA, with the latter leveraging the

CPB's visual prompts to augment gating perception. The aggregated outputs from both modules are fed into a classifier, and the predictions are refined by cross-entropy loss during backpropagation.

2.1. Isolated Gating Mechanism Adapter

Traditional MoE architectures are constrained by routing overhead in fine-grained expert partitioning. The PEER [1] architecture, however, employs the Product Key Retrieval (PKR) technique to efficiently identify and retrieve top-k experts from a large pool for a d-dimensional input vector x, using low-dimensional sub-keys to construct a key set and inner product calculations, thus reducing computational load while preserving accuracy. The gating network, parameterized by Θ, refines the expert outputs, incorporating a noise term R_{noise}, to yield the final gating decision $G(x; \Theta)$.

The sensitivity of the gating network to input noise escalates with an increasing number of fine-grained experts, which, despite the introduction of training noise, fails to optimize performance or fully exploit multimodal processing. We attribute this to the gating network's constrained feature perception due to low-dimensional sub-keys, impacting noise robustness. To address this, we introduce an IGM that distinguishes the expert-processed vector x_e from the gating vector x_g, enabling a more nuanced nonlinear transformation to reduce noise impact and enhance system performance efficiently. This refined process is formalized as follows:

$$F_{\text{MoE}}(x; \Theta, \{W_i\}_{i=1}^N) = \sum_{i=1}^{N} G(x; \Theta)_i F_i(x; W_i) \quad (1)$$

$$G(x; \Theta) = \text{softmax}(\text{TopK}(f_g(\text{PKR}(x; \Theta) + R_{\text{noise}}, k)))_i \quad (2)$$

Table 1: Cross-dataset testing results under the fixed-modal scenarios (Protocol 1) among CASIA-CeFA (C) [22], PADISI (P) [23], CASIA-SURF (S) [24], and WMCA (W) [25]. DG, MM, and FM are short for domain-generalized, multi-modal, and flexible-modal, respectively. Best results are marked in **bold**.

Method	Type	CPS→W		CPW→S		CSW→P		PSW→C	
		HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑
SSDG [26] (ECCV2022)	DG	26.09	82.03	28.50	75.91	41.82	60.56	40.48	62.31
SSAN [27] (CVPR2022)	DG	17.73	91.69	27.94	79.04	34.49	68.85	36.43	69.29
IADG [28] (CVPR2023)	DG	27.02	86.50	23.04	83.11	32.06	73.83	39.24	63.68
ViTAF [29] (ECCV2022)	DG	20.58	85.82	29.16	77.80	30.75	73.03	39.75	63.44
MM-CDCN [2] (CVPR2020)	MM	38.92	65.39	42.93	59.79	41.38	61.51	48.14	53.71
CMFL [4] (CVPR2021)	MM	18.22	88.82	31.20	75.66	26.68	80.85	36.93	66.82
ViT+AMA [9] (IJCV2024)	FM	17.56	88.74	27.50	80.00	21.18	85.51	47.48	55.56
VP-FAS [8] (arXiv 2023)	FM	16.26	91.22	24.42	81.07	21.76	85.46	39.35	66.55
MMDG [12] (CVPR2024)	MM	**12.79**	**93.83**	15.32	92.86	**18.95**	**88.64**	29.93	76.52
ViT [30]	Baseline	20.88	84.77	44.05	57.94	33.58	71.80	42.15	56.45
BIG-MoE	Ours	17.4	90.87	**10.96**	**94.35**	19.62	84.72	**7.71**	**97.72**

Table 2: Cross-dataset testing results under the limited source domain scenarios (Protocol 3) among CeFA-CeFA (C) [22], PADISI USC (P) [23], CASIA-SURF (S) [24], and WMCA (W) [25]. Best results are marked in **bold**.

Method	CW→PS		PS→CW	
	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑
SSDG [26] (ECCV2022)	25.34	80.17	46.98	54.29
SSAN [27] (CVPR2022)	26.55	80.06	39.10	67.19
IADG [28] (CVPR2023)	22.82	83.85	39.70	63.46
ViTAF [29] (ECCV2022)	29.64	77.36	39.93	61.31
MM-CDCN [2] (CVPR2020)	29.28	76.88	47.00	51.94
CMFL [4] (CVPR2021)	31.86	72.75	39.43	63.17
ViT+AMA [9] (IJCV2024)	29.25	76.89	38.06	67.64
VP-FAS [8] (arXiv 2023)	25.90	81.79	44.37	60.83
MMDG [12] (CVPR2024)	**20.12**	**88.24**	36.60	70.35
ViT [30] (Baseline)	42.66	57.80	42.75	60.41
BIG-MoE (Ours)	22.35	83.50	**14.11**	**95.13**

$$F(x; \Theta) = f(f_e(x; W_e), W) \tag{3}$$

The refined formulation shows that the input vector x is processed by f_g for gating and retrieval, followed by further processing of f_e based on gating results to produce the final output $F(x; \Theta)$. This approach integrates fine-grained experts with IGM, resulting in the IGMA, as illustrated in Fig. 2(c).

2.2. Convolutional Prompt Bypass

Previous research has shown that routing selection in MoE models is sensitive to prompt tokens [16], which can introduce noise and limit the effectiveness of Prompt Learning when applied to MoE. To address this, we developed the CPB for the IGMA, utilizing Central Difference Convolution (CDC) [2] to enhance the extraction of local spoofing cues.

The CPB process initiates by concatenating multimodal inputs along the channel dimension to create clue prompts P_c. A 30% probability masks entire modal images, setting them to zero, which is integrated into the prompts P_m as supplemental data. Static task-related prompts P_t are concurrently acquired. These prompts are merged to form a comprehensive input prompt P. Each layer's prompt P_i is combined with the perceptive vector x_g, forming an integrated perceptive vector input to the gating network. This fusion enhances perceptual stability, particularly with composite features exhibiting substantial representational variance.

The PKR method is employed to partition the perceptive vector into two sub-spaces, avoiding interference from prompt semantics and enhancing perception stability. The combined perceptive vector x_g is processed through the Attention mechanism, generating a new prompt P_{i+1} for the next layer, as described by the formula:

$$P_{i+1} = P_i + \text{Attn}(\text{Cat}(P_i, x_c)) \tag{4}$$

Here, the Efficient Channel Attention (ECA) module within the CPB enriches IGMA with supplemental perceptive information, blending insights across layers to reduce gating sensitivity and bolster the model's performance and stability.

3. EXPERIMENT

3.1. Data and Evaluation Metrics

In this study, we followed the MMDG's Protocols 1 and 3 [12], applying a Leave-One-Out (LOO) test on fixed modalities: S (SURF) [24], P (PADISI USC) [23], C (CeFA) [22], and W (WMCA) [25]. Performance was measured using Half Total Error Rate (HTER) and Area Under the Receiver Operating Characteristic Curve (AUC).

3.2. Implementation Details

All input images were standardized to 224 × 224 × 3 pixels, segmented into 14 × 14 patches, and inputted into the ViT where token hidden dimension d=768. We trained the model using the Adam optimizer, a learning rate of 5e-5, weight decay of 1e-3, over 100 epochs with a batch size of 32. The classifier was a single fully connected layer reducing the class token output from 768 to 2. The model was based on a pretrained ViT-Base on ImageNet, with the IGMA structure featuring 2 activated experts per head, in total 1600 experts, a hidden dimension of 8, and a 64-dimension CPB.

3.3. Cross-testing Results

Sufficient Source Domains Scenario. The results in Table 1 highlight our model's state-of-the-art performance (3 out of 4) across several sub-protocols. Specifically, our model with HTER dropping to 7.71%, a 34.44% decrease from the baseline, and AUC rising to 97.72%, a 41.27% increase from the baseline in 'PSW→C' setting. These improvements are a testament to the BIG-MoE architecture's superiority in handling generalized multimodal FAS tasks, indicating the excellent generalization capacity of our model across unseen scenarios.

Fig. 3: Ablation study on expert numbers and activations. (a) HTER with Varying Numbers of Activated Experts. (b) HTER with Different Total Expert Counts. The ablation study investigates the impact of expert count and activation on model performance, providing insights into the optimal configuration for expert utilization in the model.

Limited Source Domains Scenario. The results of 'PS→CW' in Table 2 also demonstrate our model's superior generalization performance under limited source domains, demonstrating enhanced multimodal generalization over 'ViT (Baseline)'. With the AUC of 95.13% and the HTER of 14.11%, our model leads in state-of-the-art generalization performance, highlighting the model's outstanding ability to generalize across limited source domains scenarios

3.4. Ablation Study

To validate the rationality and effectiveness of BIG-MoE, we conducted meticulous ablation experiments. These aimed to evaluate the impact of prompt settings on model performance, comparing BIG-MoE with a Vision Transformer, a coarse-grained MoE (ST MoE), and a fine-grained MoE (PEER) to highlight the advantages of our CPB and IGMA. Experiments were conducted using the CPW→S configuration, testing prompt settings with P_s alone, P_s with P_d, and the full setup. Results showed that all prompt configurations improved performance, substantiating the rationality and effectiveness of our approach and demonstrating BIG-MoE's potential to enhance model capabilities, providing insights for future work.

Impact of Experts' Granularity. Fig. 3 indicates that while moderate increases in IGMA granularity enhance the Adapter's performance, overly fine granularity can lead to a decline in effectiveness. This suggests a critical trade-off: granularity must be judiciously adjusted to maximize system performance, underscoring the need for a balanced granularity strategy in model optimization.

Effectiveness of IGMA. The data 'w/ IGMA+CPB (With P_t)' in Table 3 indicates that, with the help of perceptual cues, IGMA sees a 15.03% HTER reduction and a 12.94% AUC increase over the resluts of PEER [1]. The integration of fine-grained experts with cues in the IGMA framework maximizes performance, surpassing the benefits of prompts alone. The framework is indispensable for achieving optimal results.

Effectiveness of CPB. The results from 'w/ IGMA' to 'BIG-MoE' in Table 3 delineate the significant performance enhancement attributable to each prompt element, thereby validating our design rationale. These findings not only

Table 3: Ablation results on the proposed BIG-MoE.

Method	CPW→ S	
	HTER(%)↓	AUC(%)↑
ViT [30] (Baseline)	20.88	84.77
w/ ST MoE [15]	14.31	88.69
w/ PEER [1]	22.34	84.97
w/ IGMA	21.12	85.50
w/ IGMA+CPB (With P_t)	20.55	88.41
w/ IGMA+CPB (With P_t&P_c)	10.44	93.87
BIG-MoE (Ours)	**10.96**	**94.35**

Fig. 4: t-SNE visualization when respectively tested on CeFA, PADISI, SURF, and WMCA domains.

demonstrate the synergistic effects across modalities and features, but also highlight the substantial refinement in cue detection and decision-making capabilities afforded by an optimal prompt combination.

3.5. Visualization and Analysis

t-SNE was used for dimensionality reduction and visualization of complex data, effectively showing its utility with ViT and BIG-MoE methods. Fig. 4 illustrates BIG-MoE's advanced classification, aided by CPB technology in capturing fine feature differences. However, to enhance model generalizability across domains, optimizing multi-domain training samples is needed due to variations in feature representation from different training datasets.

4. CONCLUSION

This paper introduces BIG-MoE, integrating the Isolated Gating Mechanism Adapter and Convolutional Prompt Bypass for generalized multimodal face anti-spoofing (FAS). The former detects subtle spoofing cues with fine-grained experts and efficient key retrieval, while the latter extracts local features and boosts model perception via attention mechanisms. Our method demonstrates superior performance in generalized multimodal FAS through extensive experiments. Future work will focus on improving MoE's generalization with limited samples and in multimodal settings.

5. REFERENCES

[1] Xu Owen He, "Mixture of a million experts," *arXiv preprint arXiv:2407.04153*, 2024.

[2] Zitong Yu, Yunxiao Qin, Xiaobai Li, Zezheng Wang, Chenxu Zhao, Zhen Lei, and Guoying Zhao, "Multi-modal face anti-spoofing based on central difference networks," in *CVPR*, 2020, pp. 650–651.

[3] Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao, "Deep learning for face anti-spoofing: A survey," *TPAMI*, vol. 45, no. 5, pp. 5609–5631, 2022.

[4] Anjith George and Sébastien Marcel, "Cross modal focal loss for rgbd face anti-spoofing," in *CVPR*, 2021, pp. 7882–7891.

[5] Anjith George and Sébastien Marcel, "Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks," *TIFS*, vol. 16, pp. 361–375, 2020.

[6] Ajian Liu and Yanyan Liang, "Ma-vit: Modality-agnostic vision transformers for face anti-spoofing," *arXiv preprint arXiv:2304.07549*, 2023.

[7] Ajian Liu, Zichang Tan, Zitong Yu, Chenxu Zhao, Jun Wan, Yanyan Liang, Zhen Lei, Du Zhang, Stan Z Li, and Guodong Guo, "Fm-vit: Flexible modal vision transformers for face anti-spoofing," *TIFS*, vol. 18, pp. 4775–4786, 2023.

[8] Zitong Yu, Rizhao Cai, Yawen Cui, Ajian Liu, and Changsheng Chen, "Visual prompt flexible-modal face anti-spoofing," *arXiv preprint arXiv:2307.13958*, 2023.

[9] Zitong Yu, Rizhao Cai, Yawen Cui, Xin Liu, Yongjian Hu, and Alex C Kot, "Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing," *IJCV*, pp. 1–22, 2024.

[10] Zitong Yu, Ajian Liu, Chenxu Zhao, Kevin HM Cheng, Xu Cheng, and Guoying Zhao, "Flexible-modal face anti-spoofing: A benchmark," in *CVPR*, 2023, pp. 6346–6351.

[11] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng, "Provable dynamic fusion for low-quality multimodal data," in *ICML*. PMLR, 2023, pp. 41753–41769.

[12] Xun Lin, Shuai Wang, Rizhao Cai, Yizhong Liu, Ying Fu, Wenzhong Tang, Zitong Yu, and Alex Kot, "Suppress and rebalance: Towards generalized multi-modal face anti-spoofing," in *CVPR*, 2024, pp. 211–221.

[13] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby, "Scaling vision with sparse mixture of experts," *NeurIPS*, vol. 34, pp. 8583–8595, 2021.

[14] William Fedus, Barret Zoph, and Noam Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *JMLR*, vol. 23, no. 120, pp. 1–39, 2022.

[15] Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus, "St-moe: Designing stable and transferable sparse expert models," *arXiv preprint arXiv:2202.08906*, 2022.

[16] Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You, "Open-moe: An early effort on open mixture-of-experts language models," *arXiv preprint arXiv:2402.01739*, 2024.

[17] Yuqi Yang, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, and Bo Li, "Multi-task dense prediction via mixture of low-rank experts," in *CVPR*, 2024, pp. 27927–27937.

[18] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Shouhong Ding, and Lizhuang Ma, "Adaptive mixture of experts learning for generalizable face anti-spoofing," in *ACM MM*, 2022, pp. 6009–6018.

[19] Chenqi Kong, Anwei Luo, Song Xia, Yi Yu, Haoliang Li, and Alex C Kot, "Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection," *arXiv preprint arXiv:2404.08452*, 2024.

[20] Ajian Liu, "Ca-moeit: Generalizable face anti-spoofing via dual cross-attention and semi-fixed mixture-of-expert," *IJCV*, pp. 1–14, 2024.

[21] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu, "Visual prompt multi-modal tracking," in *CVPR*, 2023, pp. 9516–9526.

[22] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li, "Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *ICCV*, 2021, pp. 1179–1187.

[23] Mohammad Rostami, Leonidas Spinoulas, Mohamed Hussein, Joe Mathai, and Wael Abd-Almageed, "Detection and continual learning of novel face presentation attacks," in *ICCV*, 2021, pp. 14851–14860.

[24] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li, "Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing," *IEEE T-BIOM*, vol. 2, no. 2, pp. 182–193, 2020.

[25] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *TIFS*, vol. 15, pp. 42–55, 2019.

[26] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen, "Single-side domain generalization for face anti-spoofing," in *CVPR*, 2020, pp. 8484–8493.

[27] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang, "Domain generalization via shuffled style assembly for face anti-spoofing," in *CVPR*, 2022, pp. 4123–4133.

[28] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Ran Yi, Shouhong Ding, and Lizhuang Ma, "Instance-aware domain generalization for face anti-spoofing," in *CVPR*, 2023, pp. 20453–20463.

[29] 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 *ECCV*. Springer, 2022, pp. 37–54.

[30] 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," *arXiv preprint arXiv:2010.11929*, 2020.