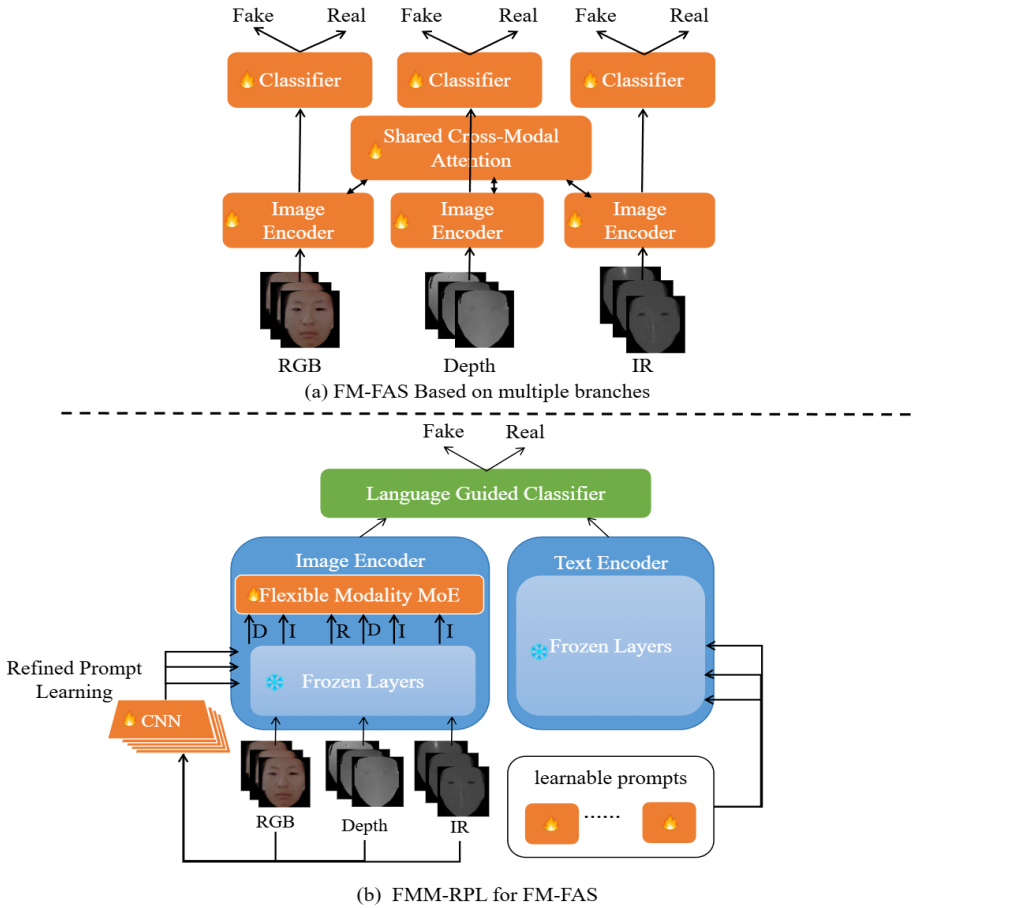
Flexible Modality Mixture of Experts with Refined Prompt Learning for Flexible Face Anti-Spoofing

Xu Zhang, Hui Ma, Yefan Li, Fuqing Duan

Frontiers of Computer Science, DOI: 10.1007/s11704-026-50892-9

Problems & Ideas

- Problems of conventional Flexible Face Anti-Spoofing approaches：
  - ➢ Designing different branches for each modality leads to an increase in the model parameters as it requires training three separate branches for the different modalities.
  - ➢ Their visual encoders exhibit insufficient adaptability to the specific characteristics of each modality, which may result in unstable feature extraction across modalities
- Ideas：
  - ➢ We design a network architecture that can adjust its parameters based on input modality without needing separate models, thereby learning more robust modality-agnostic representations.



(a) FM-FAS Based on multiple branches

(b) FMM-RPL for FM-FAS

(a) The multiple branches method requires dedicated image encoder branches for each modality to capture modality-specific representations, and a shared cross-modal attention layer to extract modality-invariant features through joint reasoning across modalities.(b) FMM-RPL eliminates redundant architectural components by employing a Flexible Modality MoE (FMM) module to dynamically allocate tasks to expert modules and using Refined Prompt Learning (RPL) to hierarchically accumulate both visual learnable prompts from original face images and randomly initialized textual learnable prompts

Main Contributions

- Main Contributions
  - To further assist visual feature extraction in capturing detailed information, we propose a dynamic expert fusion module within the visual encoder of CLIP that activates different experts based on input modality characteristics, improving adaptability and feature richness.
  - To ensure that the text and vision encoding phases fully learn the modality-agnostic features, we introduce a hierarchical prompt learning method that integrates visual prompts from original images and randomly initialized textual prompts, enabling more effective modality-agnostic representations across both visual and text branches.
  - We design a Refined Cross-level Loss function that leverages local image patches and global features, facilitating the extraction of fine-grained cues across modalities and enhancing the robustness of FAS.

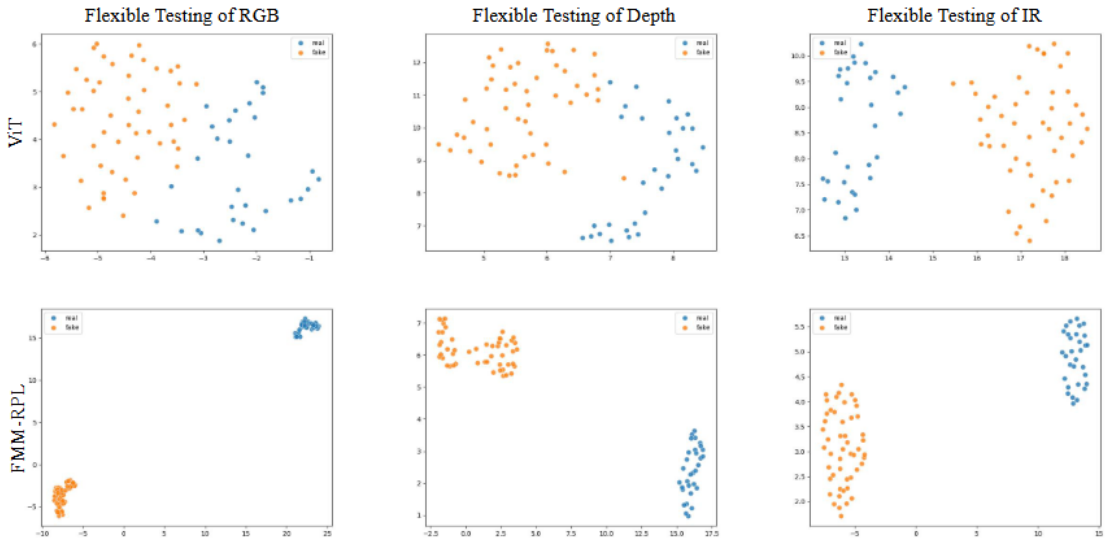


Fig. 5 Comparing the linear divisibility of visual features.

Table 1 Comparing flexible modal results (%) based on multi-modal datasets. 'SOTA' denotes the state-of-the-art method with publicly reported results on the respective dataset. R-D-I refers to approaches that utilize synchronized RGB (R), depth (D), and infrared (IR) data samples.

Method	Train	Test	SURF			CeFA (Protocol 4)			WMCA (Protocol "seen")		
			APCER	BPCER	ACER	APCER	BPCER	ACER	APCER	BPCER	ACER
FM-ViT [25]	R&D&I	R	8.77	16.00	12.38	36.61±9.51	5.50±1.32	21.06±4.90	2.26	3.48	2.87
		D	5.14	1.83	3.49	2.79±0.44	1.71±1.13	2.25±0.36	2.04	2.61	2.32
		I	1.34	3.83	2.59	3.43±2.73	2.33±1.91	2.88±2.23	3.39	0.87	2.13
CMS-Enhancer [27]	R&D&I	R	18.09	2.50	10.30	21.56±2.99	12.53±3.51	14.57±1.41	2.71	1.74	2.23
		D	3.35	3.00	3.18	3.50±1.95	2.00±0.75	2.42±0.79	4.30	0.00	2.15
		I	4.13	0.33	2.23	2.94±3.42	2.25±1.09	2.59±2.24	2.04	1.74	1.89
FM-CLIP [27]	R&D&I	R	18.59	1.83	10.21	14.4±3.37	9.3±2.98	11.87±3.14	2.49	0.87	1.68
		D	3.21	2.83	3.02	4.16±2.08	1.08±1.01	2.29±0.68	3.85	0.00	1.92
		I	2.01	2.00	**2.01**	2.15±1.37	2±1.39	**2.07±1.36**	3.39	0.00	1.7
FMM-RPL	R&D&I	R	10.22	0.83	**5.53**	11.09±0.65	4.75±2.04	**7.92±0.72**	0.00	0.00	**0.00**
		D	1.40	1.33	**1.36**	2.77±0.53	1.58±0.59	**2.18±0.56**	1.20	0.00	**0.60**
		I	4.36	2.33	3.34	2.63±1.28	1.67±0.62	2.15±0.90	0.30	0.86	**0.58**