Prompt-Based Cross-Modal Feature Alignment for Weakly Supervised IFER

Hanqin Shi, Xiaofeng Kang, Jiaxiang Wang, Aihua Zheng, and Wenjuan Cheng

***Abstract*—Infrared Facial Expression Recognition (IFER) encounters challenges in data acquisition and annotation under low-light conditions, making fully supervised training difficult. Although pre-trained Vision-Language Models (VLMs) can enhance generalization for downstream tasks, their insufficient attention modeling in cross-domain scenarios leads to ineffective local semantic correlation. To address this, we propose a Prompt-based Cross-modal feature Alignment (PCA) method that improves weakly supervised IFER performance by leveraging RGB facial expression data. The PCA framework comprises two key components: (1) a Cross-modal Prompt Transfer (CPT) strategy that integrates category-specific information to distinguish expressions, and (2) an Image-Guided Alignment (IGA) module that achieves feature alignment using dual-domain feature banks. Experimental results on two benchmark datasets demonstrate that our method significantly outperforms current state-of-the-art approaches, confirming its effectiveness and superiority.**

***Index Terms*—Cross-modal prompt transfer (IFER), image-guided alignment, infrared facial expression recognition.**

I. Introduction

Facial Expression Recognition (FER) infers psychological states by analyzing changes in facial features and has broad applications in human-computer interaction [1], security surveillance [2], healthcare [3], education [4], and customer service [5]. Current research primarily focuses on emotion recognition in visible light environments [3], [6], [7], while studies on IFER under low-light or adverse weather conditions (e.g., darkness or fog) remain limited. Due to challenges in data collection and annotation, high-quality labeled datasets are scarce, which restricts the ability of models to interpret emotional states. To address these challenges, we adopt partial labels for a study on weakly supervised IFER.

Traditional FER primarily focuses on facial expression feature extraction, where deep learning methods are crucial. For instance, Li et al. [8] proposed an attention-based convolutional neural network to identify and emphasize unoccluded facial regions. Xie et al. [9] integrated a region attention mechanism into the convolutional neural network architecture to enhance the focus on local features. Similarly, Wang developed a region attention network to prioritize key facial regions, addressing challenges such as occlusion and gesture variations. Zeng et al. [10] introduced the Meta-Face2Exp framework, which combines prior knowledge of facial expressions with pseudo-label optimization to improve single-domain FER performance.

Large-scale visual language models (VLMs) have recently demonstrated remarkable generalization performance across various downstream tasks. Models such as CLIP [11] excel in cross-modal semantic alignment, enabling them to associate semantic features effectively and enhance the generalization capabilities of both unsupervised and weakly supervised models. Fine-tuning these VLMs allows for improved adaptation to various downstream tasks. Prompt tuning has gained significant attention as a prominent fine-tuning technique, exemplified by methods such as CoOp [12] and MaPLe [13]. CoOp employs soft prompts to learn optimal textual representations, while MaPLe advances this approach by incorporating both visual and verbal prompts, ensuring synergy between modalities. Compared to CLIP, MaPLe demonstrates superior domain alignment, as evidenced by its lower KL divergence and MMD values, which indicate that prompt tuning can help mitigate domain discrepancy. Although prompt adaptation enhances the recognition capabilities of large models, existing methods struggle to sufficiently attend to sample features under weak supervision [14], [15]. Optimal transport theory [16] demonstrates that cross-domain feature alignment can effectively reinforce feature mining, thereby improving representation learning for weakly supervised samples [17]. This necessitates an in-depth investigation of cross-domain prompt arrangement strategies to advance performance in weakly supervised IFER recognition.

This paper proposes a novel Prompt-based Cross-modal Feature Alignment (PCA) method, which enhances weakly supervised infrared facial expression recognition performance by aligning cross-modal facial emotion features. The framework consists of two core components: a Cross-modal Prompt Transfer (CPT) strategy and an Image-Guided Alignment (IGA) module. First, a vision transformer is employed to extract low-level feature embeddings. Then, the CPT strategy integrates category-specific information into prompts to enhance domain-transferred emotion feature representations. Finally, the IGA module leverages CLIP's vision-language capabilities for cross-modal

Received 23 June 2025; revised 5 August 2025; accepted 15 August 2025. Date of publication 21 August 2025; date of current version 29 August 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62102344, in part by the Basic Science Major Foundation (Natural Science) of the Jiangsu Higher Education Institutions of China under Grant 22KJA520012, in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2022-036, and in part by Xuzhou Science and Technology Plan Project under Grant KC22305. The associate editor coordinating the review of this article and approving it for publication was Prof. Parvaneh Saeedi. *(Corresponding author: Jiaxiang Wang.)*

Hanqin Shi and Xiaofeng Kang are with the School of Information Engineering (School of Big Data), Xuzhou University of Technology, Xuzhou 221018, China (e-mail: shihanqin0726@sina.com; kxfeng07@163.com).

Jiaxiang Wang is with the School of Artificial Intelligence, Anhui University of Science and Technology, Hefei 231131, China (e-mail: Netizenwjx@foxmail.com).

Aihua Zheng is with the School of Artificial Intelligence, Anhui University, Hefei 230601, China (e-mail: ahzheng214@foxmail.com).

Wenjuan Cheng is with the School of Computer and Information, Hefei University of Technology, Hefei 230009, China (e-mail: cheng@ah.edu.cn).

Digital Object Identifier 10.1109/LSP.2025.3601513

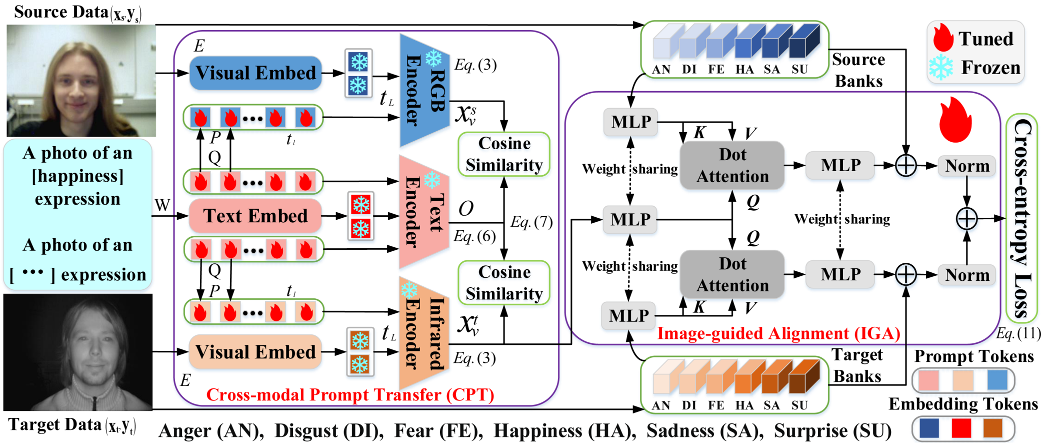


Fig. 1. The PCA method overview involves two core components. The CPT strategy facilitates the transfer of emotional features across domains by incorporating category-specific information as prompts to precisely define various expression features. In addition, the IGA module leverages the visual language features of the CLIP model to bridge the gap between the visible and infrared image feature banks, thereby achieving cross-modal feature alignment.

association learning, connecting visible and infrared emotion feature spaces, thereby significantly improving weakly supervised IFER performance. The main contributions of this study are as follows:

- We propose a CPT strategy, which incorporates category-specific information to explicitly distinguish expression categories, thereby enhancing cross-domain transfer of emotional features.
- We design an IGA module that leverages the vision-language capabilities of CLIP to establish cross-modal connections between visible and infrared feature spaces.
- Experiments on NVIE [18] (thermal infrared data, denoted as N_E) and Oulu-CASIA [19] (near-infrared and dark data, denoted as O_{NI} and O_{Dark}) demonstrate that the proposed method achieves state-of-the-art (SOTA) performance in weakly supervised IFER tasks.

II. Method

A. Overview of Infrared Facial Expression Recognition

This paper addresses the challenge of weakly supervised IFER by leveraging fully labeled visible images $((x_i^s, y_i^s)_{i=1}^{n_s} \sim D_s)$ in the source domain and a limited number of labeled infrared images $((x_i^t, y_i^t)_{i=1}^{n_t} \sim D_t)$ in the target domain. Both domains share the same set of expression labels (y_i^s, y_i^t), which denotes the hot ground truth of $\{1, 2, \ldots, C\}$, where C denotes the total number of categories. The traditional IFER framework comprises a feature extractor F and a linear classification head. However, the limited availability of labeled infrared data necessitates the development of more efficient solutions. Leveraging the strong generalization capabilities of the pre-trained CLIP [11], we propose a prompt-based cross-modal feature alignment method (as shown in Fig. 1), which fine-tunes CLIP by incorporating prompts into visual and textual inputs, thereby enhancing the generalization performance of IFER. The training pseudo-code is shown in Algorithm 1.

B. Cross-Modal Prompt Transfer Strategy

The significant heterogeneity between visible and infrared images poses a challenge for the ViT-based CLIP model to directly learn cross-modal prompts [20], as it cannot dynamically adjust and align the representation spaces required for IFER. To address this, we introduce textual descriptions of emotion categories as intermediate features, connecting the visual and language branches through a cross-modal prompt transfer network, thereby enhancing the consistency of cross-modal visual feature representations.

The proposed network architecture consists of two image encoders and one text encoder. In the image encoder, we integrate learnable visual prompts into the visual branch of the CLIP model. The input image $x \in R^{3\times H\times W}$ is divided into M patches x_m, which are assigned to patch embeddings $E_0 \in R^{M\times d_v}$. Simultaneously, we define a set of k learnable prompt vectors $P = \{p^i \in R^{d_v}\}_{i=1}^{k}$. The patch embeddings E_l, prompt vectors, and a learnable class token t_l are jointly fed into the $(l+1)$-th transformer block of the visual branch and processed through L consecutive layers, as follows:

$$[t_l, E_l, \_] = F_l([t_{l-1}, E_{l-1}, P_{l-1}]) \quad l = 1, 2, \ldots, J, \tag{1}$$

$$[t_l, E_l, P_l] = F_l([t_{l-1}, E_{l-1}, P_{l-1}]) \quad l = J+1, \ldots, L, \tag{2}$$

where $[t_l, E_l, P_l]$ is embedded in the space $R^{(1+M+k)\times d_v}$. These feature vectors, along with the learnable prompt set P, are processed through the layers F of the image encoder up to the J-th layer. Each layer F_l consists of a multi-head self-attention mechanism and a feed-forward neural network, both enhanced by layer normalization [21] and residual connections [22]. The $IPro$ extracts the image feature x_v from the top-layer class token embedding t_L, calculated as follows:

$$x_v = IPro([t_L]). \tag{3}$$

To integrate prompt information into the language context, the text branch of the CLIP model introduces k learnable prompt vectors $Q = \{q^i \in R^{d_w}\}_{i=1}^{k}$, which are randomly initialized via a normal distribution. The text encoder combines Q with the initial word embedding vectors $W_0 = [w_0^1, w_0^2, \ldots, w_0^N] \in R^{N\times d_w}$ to form a new sequence $[Q, W_0]$, which is then fed into the $(l+1)$-th transformer module G_{l+1} of the text branch for detailed analysis, as follows:

$$[\_, W_l] = G_l([Q_{l-1}, W_{l-1}]) \quad l = 1, 2, \ldots, J, \tag{4}$$

$$[Q_l, W_l] = G_l([Q_{l-1}, W_{l-1}]) \quad l = J+1, \ldots, L. \tag{5}$$

To obtain the final textual representation o, we use the $TPro$ model to map the L-layer output W_L of G_L, thus projecting the text embedding into the aligning embedding space.

$$o = TPro([W_L]). \tag{6}$$

Initial textual prompts are randomly generated from a normal distribution, while visual prompt P is generated through a vision-language linear transformation mapping function $T(.)$, denoted as $P_k = T_k(Q_k)$. In the IFER task, manually crafted text prompts construct an expression mapping with category labels $y \in \{1, 2, \ldots, C\}$. The predicted label $\overline{y}$ is determined by identifying the label with the highest cosine similarity to the image x. Finally, the vision-prompt-based network is trained on the image dataset $(x_i, y_i)_{i=1}^{n_s/n_t}$ by minimizing the cross-entropy loss, as described below:

$$p(\overline{y}|x_v) = \frac{exp(sim(x_v, o_{\overline{y}}))}{\Sigma_{i=1}^{C} exp(sim(x_v, o_i))}, \tag{7}$$

$$L_{cls} = E_{(x_i,y_i)\sim\{D_s/D_t\}} \left(\sum_{\overline{y}} (-y_i logp(\overline{y}|x_v)) \right). \tag{8}$$

C. Image-Guided Alignment Module

The IGA module narrows the visible-infrared modal gap by aligning the dual domain feature bank. First, CLIP identifies the highest-probability emotion representations, with top-$N(5)$ visual features selected per category for bank construction. The source domain feature bank has C categories, each with N samples; the target domains have few samples per category. Central features are then computed for each category to form the source (x_{sc}) and target (x_{tc}) domain feature banks. Finally, the IGA module uses these feature banks for cross-domain feature alignment. A three-layer shared MLP f_{pr} is employed to transform the image features $x_{s/t}$, x_{sc}, and x_{tc} into query, key, and value vectors. The transformation is as follows:

$$[Q_i, K_i, V_i] = f_{pr}(x_i) \quad i \in \{s/t, sc, tc\}, \tag{9}$$

we then obtain the enhanced feature xa_i with the help of another weight-sharing projector f_{po}, as described below:

$$xa_i = f_{po}(softmax\left(\frac{Q_{s/t}K_i^T}{\epsilon}\right) V_i) \quad i \in \{sc, tc\}, \tag{10}$$

where ϵ denotes the scaling factor, and T represents the transpose operation. The adjusted visual features x_i are combined with the original features and normalized to obtain new features $xa'_i, i \in \{sc, tc\}$. By summing $xa'_{sc} + xa'_{tc}$, the alignment features x' for dual-domain images are obtained. Finally, the cross-domain alignment image dataset $(x'_i, y_i)_{i=1}^{n_s/n_t}$ is used to train the image alignment network by minimizing contrastive loss, as defined below:

$$p(\overline{y}|x') = \frac{exp(sim(x', o_{\overline{y}}))}{\Sigma_{i=1}^{C} exp(sim(x', o_i))}, \tag{11}$$

$$L_{iga} = E_{(x_i,y_i)\sim\{D_s/D_t\}} \left[\sum_{\overline{y}} (-y_i logp(\overline{y}|x')) \right], \tag{12}$$

we combine the cross-entropy loss from (8) with the image-guided contrastive loss defined in (12), λ_1 and λ_2 to composite to obtain total loss:

$$L_{final} = \lambda_1 L_{cls} + \lambda_2 L_{iga} \tag{13}$$

Algorithm 1: Learning Process of PCA Method.

Input: Multi-modal training data x^s, x^t, W.
Parameter: Learnable tokens t_l, two image encoder $IPro$, a text encoder $TPro$ and Prompt vectors (P, Q).
Output: The final loss $\mathcal{L}_{final}$.
Initialize $IPro, TPro$ from the pre-trained CLIP.
for *n* in [1, epochs] **do**
  // Learning patch embeddings (E, W), prompt vectors (P, Q), and class token (t).
  $[t_l, E_l, P_l] = F_l([t_{l-1}, E_{l-1}, P_{l-1}])$, (2)
  $[Q_l, W_l] = G_l([Q_{l-1}, W_{l-1}])$, (5)
  for *i* in [1, iterations] **do**
    // Sample a batch samples from x_i^s, x_i^t.
    $x_v = IPro([t_L])$, (3)
    $o = TPro([W_L])$, (6)
    // Cross-domain transfer.
    $p(\overline{y}|x_v) = \frac{exp(sim(x_v, o_{\overline{y}}))}{\Sigma_{i=1}^{C} exp(sim(x_v, o_i))}$, (7)
    // Cross-domain alignment.
    $p(\overline{y}|x') = \frac{exp(sim(x', o_{\overline{y}}))}{\Sigma_{i=1}^{C} exp(sim(x', o_i))}$ (11)
    Calculate features for cross-domain alignment via (9) and (10).
  end for
end for

III. Experiments

A. Implementation Details

This paper is implemented using the PyTorch framework and accelerated with a Tesla V100 32 GB GPU. Based on the pre-trained ViT-B/16 model, the CLIP architecture is the baseline. The proposed PCA method consists of two key components: (1) a cross-modal prompt transfer strategy, the core mechanism of which lies in prompt sequence design that directly governs parameter efficiency, with empirical validation confirming that the dual prompt configuration delivers optimal performance; (2) an Image-Guided Alignment module requiring a few MLPs for training. Consequently, our approach achieves significant parameter reduction compared to conventional methods that require full-parameter training. The experimental settings include a batch size of 32, an initial learning rate of 0.003, and the SGD optimiser (momentum $= 0.9$, weight decay $= 0.0005$) for 30 training epochs. The tradeoff parameters λ_1 and λ_2 are set to 0.5, indicating the equal importance of the two modules.

B. Comparison to the State-of-The-Arts

To validate the effectiveness of our PCA approach, we conduct comparative experiments with several established FER techniques using two infrared and dark FER datasets, as shown in Table I. Results demonstrate that traditional FER techniques exhibit limited effectiveness for cross-domain transfer learning due to the large differences between domains. Methods such as LA [32] and AGLRLS [33] require extensive source data and precise landmarks, respectively, which existing datasets cannot satisfy, leading to suboptimal performance. Our proposed

TABLE I
COMPARISON OF DIFFERENT SOTA ALGORITHMS ON THE IFER TASK. # DENOTES THE RESULTS OBTAINED THROUGH ALGORITHM RETRAINING

Methods	Backbone	N_E	O_{NI}	O_{Dark}
LPL# [23]	ResNet [22]	19.94	44.05	42.20
DETN# [24]		18.82	18.76	16.70
ECAN# [25]		20.40	17.64	16.50
BNM# [26]		19.98	40.12	38.20
FixBi# [27]		21.72	38.91	36.90
DMSRL# [28]		19.912	25.32	22.20
PDF [29]		<u>24.50</u>	**67.22**	**58.30**
PCA (Ours)		**30.80**	<u>63.10</u>	<u>54.00</u>
CDTrans# [30]	ViT [20]	17.80	19.30	16.70
pmTrans# [31]		18.96	60.27	55.50
PDF [29]		<u>21.80</u>	<u>66.21</u>	<u>60.20</u>
LA# [32]		19.70	64.25	43.08
AGLRLS# [33]		17.21	20.9	20.09
PCA (Ours)		**36.10**	**72.10**	**63.30**

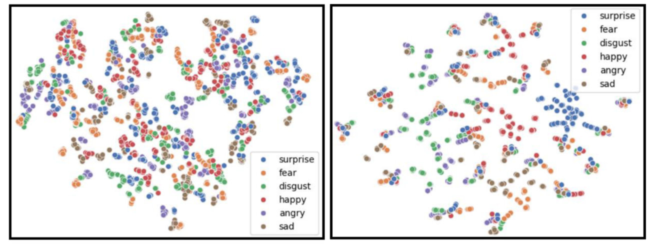
TABLE II
COMPONENT ABLATION EXPERIMENTS WITH THE PCA METHOD

Components		ResNet [22]			ViT [20]		
CPT	IGA	N_E	O_{NI}	O_{Dark}	N_E	O_{NI}	O_{Dark}
×	×	16.5	45.2	40.1	25.8	65.6	41.2
✓	×	**31.2**	60.4	51.9	34.5	69.3	62.4
×	✓	15.5	47.4	44.0	16.7	49.0	47.0
✓	✓	30.8	**63.1**	**54.0**	**36.1**	**72.1**	**63.3**

PCA method addresses these limitations, achieving superior performance on both datasets. Notably, our model significantly outperforms the PDF [29] method due to better cross-domain alignment, even though the PDF method is specifically optimized for IFER. Furthermore, the ViT [20] demonstrates higher accuracy than ResNet [22] in constructing expression feature banks for target domains, enabling ViT-based CLIP models [11] to achieve exceptional IFER performance. Experimental results reveal that N_E data underperforms both O_{NI} and O_{Dark} datasets, attributable to weaker emotional signals in N_E and greater source-target domain discrepancy. The superior accuracy of NIR data O_{NI} over O_{Dark} suggests more distinct emotional expressions in infrared images, making them particularly suitable for accurate recognition tasks.

C. Ablation Experiments

Evaluation Components: To evaluate the independent contribution of each component, ablation experiments were conducted, and the results are presented in Table II. Due to cross-domain modality differences, the CPT employs textual prompts to guide the integration of category-specific information into the image representation, enhancing the model's ability to recognize different expression categories. The IGA establishes connections between visible and infrared expression features, enabling cross-domain feature alignment. However, the model performs poorly when using only the CPT due to the modality differences and the ambiguous class boundaries of emotional values. In contrast, the joint operation of CPT and IGA creates a synergistic effect, improving both emotion recognition and feature alignment. This combination achieves the best overall model performance, highlighting the complementary nature of the two strategies and their essential role in enhancing the IFER process.



(a) t-SNE of Baseline **(b) t-SNE of PCA**

Fig. 2. Visual representation of different emotions by color.

TABLE III
ANALYSIS OF THE NUMBER OF CPT MODULE PROMPTS

Num	k=1	k=2	k=3	k=4	k=5
N_E	33.2	36.1	35.1	36.2	36.4
O_{NI}	69.8	72.1	72.6	71.4	72.0
O_{Dark}	61.1	63.3	62.1	63.1	62.2

Evaluation of Feature Distribution: We employed t-SNE [34] to visualize the emotional features of the O_{NI} dataset. As shown in Fig. 2(a) and (b), compared to the baseline methods, our proposed approach results in a more compact distribution of infrared emotion features. This indicates that our proposed PCA method effectively learns infrared emotion features, thereby demonstrating the efficacy of our approach.

Evaluation of Prompt Numbers: Table III shows the comparison of accuracy under different numbers of prompts. The results indicate that multiple prompts outperform a single prompt. However, increasing the number of prompts does not significantly enhance the model's capability. Therefore, we adopt two learnable prompts in all experiments. This shows that our model can achieve superior performance with fewer training parameters.

IV. CONCLUSION

The letter presents an innovative PCA method to enhance the performance of weakly supervised IFER through CLIP techniques. The method incorporates a CPT strategy with an IGA module. The CPT ensures explicit discrimination between expression categories by integrating category-specific information into the prompts, while the IGA utilizes cross-domain feature banks to achieve self-enhancement and cross-domain feature fusion, facilitating feature alignment to minimize inter-domain differences. Extensive experiments on multiple datasets validate the effectiveness of the PCA method, which consistently shows significant advantages. The PCA method extends CLIP to weakly supervised infrared face emotion recognition, effectively addressing the current limitation of scarce labeled data. Given the transferability of prompt learning, we can further explore its applications in unsupervised domain adaptation and downstream tasks. Additionally, to address the uncertainty in prompt learning and the inherent noise and blurring effects, we further explore deterministic prompt alignment strategies and a noise-tolerant model [35] to optimize weakly supervised domain adaptation tasks.

REFERENCES

[1] J. Park, C. Hong, S. Baik, and K. M. Lee, "CoLaNet: Adaptive context and latent information blending for face image inpainting," *IEEE Signal Process. Lett.*, vol. 31, pp. 91–95, 2024.

[2] Y. Qian and S.-K. Tang, "Pose attention-guided paired-images generation for visible-infrared person re-identification," *IEEE Signal Process. Lett.*, vol. 31, pp. 346–350, 2024.

[3] S. Wang, D. Yang, P. Zhai, and L. Zhang, "CPR-CLIP: Multimodal pre-training for composite error recognition in CPR training," *IEEE Signal Process. Lett.*, vol. 31, pp. 211–215, 2024.

[4] Y. Guo et al., "Facial expressions recognition with multi-region divided attention networks for smart education cloud applications," *Neurocomputing*, vol. 493, pp. 119–128, 2022.

[5] F. Wu, Y. Ma, H. Jin, X.-Y. Jing, and G.-P. Jiang, "MFECLIP: Clip with mapping-fusion embedding for text-guided image editing," *IEEE Signal Process. Lett.*, vol. 31, pp. 116–120, 2024.

[6] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, Jan./Mar. 2015.

[7] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[8] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[9] S. Xie, H. Hu, and Y. Wu, "Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition," *Pattern Recognit.*, vol. 92, pp. 177–191, 2019.

[10] D. Zeng, Z. Lin, X. Yan, Y. Liu, F. Wang, and B. Tang, "Face2Exp: Combating data biases for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20291–20300.

[11] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[12] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.

[13] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "MAPLE: Multi-modal prompt learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19113–19122.

[14] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, 2018.

[15] F. Huo, W. Xu, J. Guo, H. Wang, and S. Guo, "C2KD: Bridging the modality gap for cross-modal knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16006–16015.

[16] D. Alvarez-Melis and N. Fusi, "Geometric dataset distances via optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21428–21439.

[17] W. Ma, S. Li, L. Cai, and J. Kang, "Learning modality knowledge alignment for cross-modality transfer," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 33777–33793.

[18] S. Wang et al., "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 682–691, Nov. 2010.

[19] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, 2011.

[20] D. Alexey et al., "An image is worth 16 x 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[21] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[23] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2852–2861.

[24] S. Li and W. Deng, "Deep emotion transfer network for cross-database facial expression recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2018, pp. 3092–3099.

[25] S. Li and W. Deng, "A deeper look at facial expression dataset bias," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 881–893, Apr./Jun. 2022.

[26] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, and Q. Tian, "Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3941–3950.

[27] J. Na, H. Jung, H. J. Chang, and W. Hwang, "FixBi: Bridging domain spaces for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1094–1103.

[28] Y. Li, Z. Zhang, B. Chen, G. Lu, and D. Zhang, "Deep margin-sensitive representation learning for cross-domain facial expression recognition," *IEEE Trans. Multimedia*, vol. 25, pp. 1359–1373, 2023.

[29] T. Ma, Y. Qi, Z. Wang, Y. Li, and J. Wang, "Prototype transferred diverse features for day-night cross-domain facial expression recognition," in *Proc. Int. Conf. Graph. Image Process.*, 2025, pp. 294–303.

[30] T. Xu, W. Chen, P. Wang, F. Wang, H. Li, and R. Jin, "Cdtrans: Cross-domain transformer for unsupervised domain adaptation," in *Proc. Int. Conf. Learn. Representations*, 2022.

[31] J. Zhu, H. Bai, and L. Wang, "Patch-mix transformer for unsupervised domain adaptation: A game perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 3561–3571.

[32] Y. Yang et al., "Learning with alignments: Tackling the inter-and intra-domain shifts for cross-multidomain facial expression recognition," in *Proc. ACM Int. Conf. Multimedia*, 2024, pp. 4236–4245.

[33] Y. Gao, Y. Xie, Z. Z. Hu, T. Chen, and L. Lin, "Adaptive global-local representation learning and selection for cross-domain facial expression recognition," *IEEE Trans. Multimedia*, vol. 26, pp. 6676–6688, 2024.

[34] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[35] C. Liang, L. Zhu, Z. Yang, W. Chen, and Y. Yang, "Noise-tolerant hybrid prototypical learning with noisy web data," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 20, no. 10, pp. 1–19, 2024.