PromptEmo: Learning Emotion with Bilateral Textual Prompts in Multi-Domain Open-set Scenarios

Xinyi Zeng[1*], Yuxiang Yang[1*], Pinxian Zeng[1], Wenxia Yin[1], Bo Liu[2†], Xi Wu[3], Yan Wang[1†]

[1]College of Computer Science, Sichuan University
[2]Department of Computing, The Hong Kong Polytechnic University
[3]School of Computer Science, Chengdu University of Information Technology
perperstudy@gmail.com, yangyuxiang3@stu.scu.edu.cn, zengpinxian@stu.scu.edu.cn, wangyanscu@hotmail.com

Abstract

Facial Expression Recognition (FER) is crucial to human-computer interaction. Existing cross-domain FER (CD-FER) methods mainly focus on single-source closed-set scenarios, transferring knowledge from a single source domain to a target domain with identical class sets. However, CD-FER faces two real-world challenges: 1) the need to leverage information from multiple sources, leading to multi-domain shift, and 2) the necessity to recognize unseen target classes, resulting in class shift. These issues give rise to a novel and challenging task, which we define as Multi-domain Open-set FER (MO-FER). In this paper, we propose PromptEmo, a novel CLIP-based framework that leverages bilateral textual prompts to address both shifts in the MO-FER task. Leveraging the generalizability of LLM, PromptEmo constructs trainable positive prompts with LLM-generated emotion descriptions for seen classes, as well as template-derived negative prompts to enhance the reasoning for unseen classes. Then, we introduce a modal-task optimization paradigm organized from two perspectives: textual semantics and visual domains, yielding Intra-modal Space-specific Optimization (ISO) and Cross-modal Emotion-aware Interaction (CEI) strategies. ISO refines the CLIP-based textual space to ensure semantic separation between bilateral prompts and improves the latent visual space by promoting inter-domain alignment. Founded on ISO, CEI facilitates effective vision-language interactions, resulting in four joint loss terms that improve emotion recognition by shaping a domain-invariant, discriminative feature space. PromptEmo surpasses the current SOTA method by 7.7% AUC on unseen classes across four FER datasets, serving as a strong baseline for the MO-FER task.

Code — https://github.com/PerPerZXY/PromptEmo

Introduction

Facial Expression Recognition (FER) plays a vital role in human-computer interaction by interpreting emotions through facial cues (Li and Deng 2020). Advances in deep learning paired with large well-annotated datasets have greatly improved the effectiveness of FER models (Wang et al. 2022). However, conventional methods assume that both training and testing datasets are drawn from the same domain with identical distributions, an assumption that often breaks down in real-world applications. This issue, known as *domain shifts*, arises when data distribution in the target domain (testing dataset) differs from that of the source (training dataset) domain, leading to degraded target performance. To address this, Cross-Domain FER (CD-FER) has emerged (Wang et al. 2022). By employing techniques like adversarial training (Ji et al. 2021; Chen et al. 2021b) and metric learning (Ji et al. 2019) for inter-domain alignment, CD-FER methods aim to bridge the domain gap and extract domain-invariant features from a labeled source domain to enable accurate recognition in the unlabeled target domain.

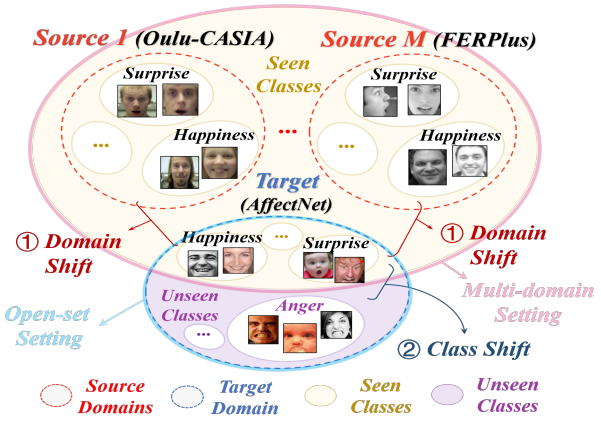


Figure 1: **MO-FER** features both multi-domain shifts (varied data distributions in **M**ulti-domain setting) and class shifts (unseen target classes in **O**pen-set setting) in **FER**.

Despite progress, CD-FER still faces two key challenges that limit its practical applicability. One major challenge is the ***multi-domain shifts***. Most CD-FER approaches typically focus on extracting knowledge from a single source domain. However, in real-world scenarios, labeled data often originates from diverse sources with varying acquisition conditions, lighting environments, and demographic characteristics (Barsoum et al. 2016; Zhao and Liu 2021). Leveraging these heterogeneous sources can greatly expand the scale and diversity of training samples, thereby facilitating the learning of more generalizable representations for complex

*These authors contributed equally.
†Corresponding author.

target domains (Mollahosseini, Hasani, and Mahoor 2017). In light of this, LA-CMFER (Yang et al. 2024b) has introduced the concept of Cross-Multidomain FER (CM-FER), aiming to extract valuable insights from multiple labeled sources and transfer knowledge to an unlabeled target. Notably, by employing inter-domain alignment at both sample and cluster levels, LA-CMFER demonstrates notable performance improvements over traditional FER baselines.

The aforementioned FER methods are typically grounded in a closed-set setting, assuming that training and testing sets share an identical set of class labels. However, different datasets (domains) may exhibit divergent class label sets due to variations in annotation protocols or collection scenarios, leading to ***class shifts***, another challenge in open-world scenarios. Such shifts can occur when unseen classes emerge during inference, a key issue addressed in Open-set Recognition (OSR) for natural images (Luo et al. 2020). For instance, Li, Yang, and Hu (2023) proposed combining self-supervised ViT (Dosovitskiy et al. 2020) with a nearest-neighbor strategy, using adaptive thresholds to detect samples of unseen classes. Nonetheless, due to the subtle differences in facial expression categories, unseen classes often exhibit high resemblance to seen classes, which constrains the effectiveness of common OSR methods in the FER field. Notably, Zhang et al. (2024b) were the first to explore the open-set FER (OS-FER) setting, which frames the identification of unseen samples as a noise-label detection problem using consistency of attention maps. Despite promising results, OS-FER does not account for cross-domain (let alone multi-domain) scenarios, limiting its broader applications.

For FER analysis in the wild, it is common to encounter both diverse data distributions and divergent class sets, making isolated solutions in CM-FER or OS-FER insufficient. To bridge this gap, we present a novel and challenging task, **M**ulti-domain **O**pen-set **F**acial **E**xpression **R**ecognition (**MO-FER**), shown in Figure 1. Notably, MO-FER is particularly challenging due to three key factors: 1) *the need for a unified framework that provides a robust feature space* generalized across diverse domains and inconsistent class sets; 2) *the necessity of incorporating additional semantic guidance* for the discrimination between seen and unseen classes, thereby enhancing performance; and 3) *the requirement for an elaborately designed optimization strategies* that effectively addresses both multi-domain shifts and class shifts while avoiding objective conflicts. Recently, advances in vision-language models (VLMs) (Zhang et al. 2024a), particularly CLIP (Radford et al. 2021), have inspired proposals and solutions of various generalization-oriented tasks (Yu, Yoo, and Lin 2024; Yang et al. 2024a). By employing self-supervised contrastive pretraining on vast amounts of image-text pairs, CLIP shapes a robust plug-and-play vision-language feature space that enables zero-shot recognition with given textual prompts. Such generalizability makes CLIP an ideal backbone for tackling the challenging MO-FER task. By adapting CLIP with semantically-informed textual prompts and carefully-crafted optimization strategies, this paper seeks to create a cross-domain, well-aligned feature space that accurately identifies unseen emotions while preserving the discriminability of known ones.

Motivated by the above insights, we propose **PromptEmo**, a novel CLIP-based framework using bilateral textual prompts to address both multi-domain and class shifts in the MO-FER task. Specifically, for seen classes, we utilize large language models (LLMs) to generate fine-grained emotion descriptions, which are combined with trainable tokens to create adaptive positive prompts rich in high-level semantics. For unseen classes, we design template-derived negative prompts that explicitly exclude features from specific seen classes. Bilateral prompts are processed through a frozen CLIP text encoder to construct a preliminary CLIP-based textual space. Next, we delve deeper into the MO-FER task and CLIP structure, devising the modal-task optimization paradigm from two perspectives: textual semantics (positive and negative prompts) and visual domains (source and target features). This results in two strategies: Intra-modal Space-specific Optimization (ISO) and Cross-modal Emotion-aware Interaction (CEI). As a preliminary foundation, ISO refines the textual space to ensure semantic separation between bilateral prompts while shaping the latent visual space by promoting inter-domain alignment, yielding two specific loss terms. Building upon ISO, CEI facilitates effective vision-language interactions by incorporating four joint loss terms, unifying dual specific spaces into a cross-domain, modal-consistent feature space. ISO and CEI work synergistically to address multi-domain and class shifts, enabling the extraction of domain-invariant features for both seen and unseen emotions. Our contributions are fourfold:

- We introduce the challenging multi-domain open-set facial expression recognition (MO-FER) task and propose PromptEmo, a strong baseline that addresses both multi-domain and class shifts through bilateral textual prompts and carefully-crafted optimization strategies.
- We adapt the CLIP framework by designing bilateral textual prompts that employ detailed LLM-generated emotion descriptions as positive prompts and exclusionary templates as negative prompts, providing high-level semantic guidance for effective expression representation.
- Guided by the modal-task optimization paradigm, we devise the Intra-modal Space-specific Optimization (ISO) and Cross-modal Emotion-aware Interaction (CEI) strategies. ISO refines intra-modal space to ensure textual semantic separation and visual inter-domain alignment, while CEI unifies the dual spaces into a cross-domain, modal-consistent feature space, mitigating dual shifts.
- PromptEmo surpasses the current SOTA method by 7.7% AUC on unseen classes across four FER datasets, serving as a strong baseline for the challenging MO-FER task.

Related Works

From CD-FER to CM-FER

Cross-Domain Facial Expression Recognition (CD-FER) addresses performance degradation caused by domain distribution discrepancies (Zhang, Song, and Zheng 2022). By leveraging Unsupervised Domain Adaptation (UDA) techniques, such as metric learning (Ji et al. 2019; Li et al. 2021) and adversarial learning (Ji et al. 2021), CD-FER

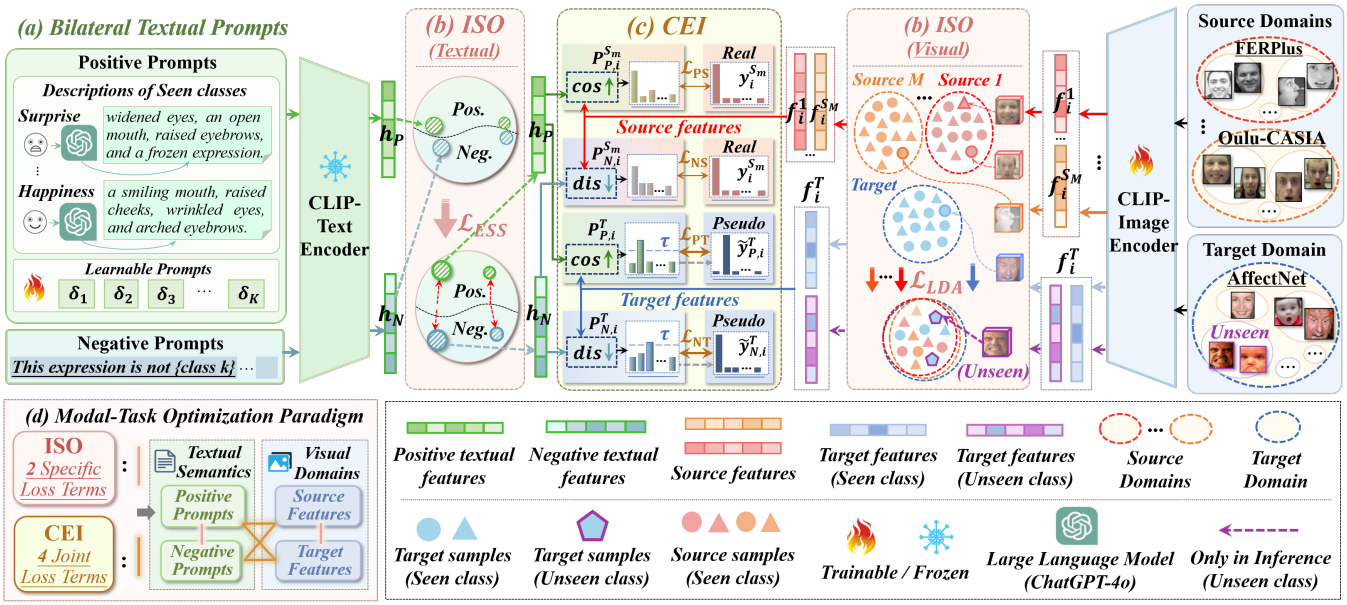


Figure 2: Illustration of our proposed PromptEmo. Built on CLIP with a fixed text encoder, PromptEmo fine-tunes its visual encoder for the MO-FER task by integrating (a) Bilateral Textual Prompts, plus two strategies (b) Intra-modal Space-specific Optimization (ISO) and (c) Cross-modal Emotion-aware Interaction (CEI) under the (d) Modal-Task Optimization Paradigm.

transfers knowledge from labeled source data to unlabeled target data, enhancing model generalization. However, CD-FER focuses on single-source domains, neglecting the diversity of multiple sources. Multi-Source Domain Adaptation (MDA), which combines labeled data from multiple sources, has made progress in natural image tasks (Yang et al. 2025; Zhao et al. 2018), but still faces challenges in FER due to boundary ambiguities. LA-CMFER (Yang et al. 2024b) first introduces the Cross-Multidomain FER task and tackles it by aligning source and target domains at both sample and cluster levels, achieving notable improvements.

From OSR to OS-FER

In Open-Set Recognition (OSR) tasks, the test set may contain unseen classes, requiring models to classify seen and recognize unseen ones (Luo et al. 2020). OSR methods can be divided into discriminative models (Li, Yang, and Hu 2023), which identify unseen classes through exclusion rules, and generative models (Rakshit et al. 2020), which generate synthetic samples to aid recognition. However, OSR methods face limitations in generalization to the FER field due to subtle inter-class differences in facial expressions. Zhang et al. (Zhang et al. 2024b) first proposed the task and the solution for Open-set FER (OS-FER), framing the identification of unseen samples as noise label detection using attention map consistency.

Although recent pioneering works such as LA-CMFER and OS-FER have advanced FER for broader applications, in-the-wild FER analysis often entails both diverse data distributions and divergent class sets. This paper presents Multi-domain Open-set FER (MO-FER), a challenging task that addresses both challenges rather than in isolation.

Adapting CLIP to MO-FER

Vision-Language Pretraining has demonstrated strong generalization capabilities (Gan et al. 2022). As one of the milestone models, CLIP (Radford et al. 2021) performs contrastive pretraining over 400 million image-text pairs to encourage positive proximity and negative separation, creating a robust multi-modal feature space that generalizes well across new tasks without requiring extra annotations. However, directly applying CLIP's standard textual prompts and general contrastive objective to the MO-task is less effective due to the lack of fine-grained cues in prompts and the significant challenge of reducing dual shifts. To better fit MO-FER, PromptEmo adapts CLIP with semantically-informed textual prompts and carefully-crafted optimization strategies. By establishing a well-aligned, cross-domain feature space, PromptEmo identifies unseen emotions while preserving the discriminability of known ones.

Methodology

Task Definition and Model Overview

Task Definition. For the CM-FER task, there are M labeled source datasets, where the m-th source S_m is denoted as $\mathcal{D}^{S_m} = \{(x_i^{S_m}, y_i^{S_m})\}_{i=1}^{|S_m|}$. $x_i^{S_m}$ is the i-th image and $y_i^{S_m} \in \{0,1\}^{|C^S|}$ is its corresponding one-hot label, with C^S being the shared seen-class set for all source domains. $K = |C^S|$ represents the number of shared categories. The target domain T contains an unlabeled dataset $\mathcal{D}^T = \{x_i^T\}_{i=1}^{|T|}$, where x_i^T denotes the i-th unlabeled target image. In CM-FER, which only involves seen classes, the category set in the target domain is assumed to match sources, i.e., $C^T = C^S$.

In the proposed MO-FER task, the category set for the target domain may include expression categories that are unseen in the source domains, leading to $C^T \neq C^S$, and $\overline{C^T} = C^T \setminus C^S$ represents the unseen classes. During inference, the goal of MO-FER is to classify target samples into one of the known categories from C^S, or determine whether they belong to one of the unseen categories from $\overline{C^T}$.

Model Overview. As shown in Figure 2, PromptEmo adapts two encoder branches of CLIP: (1) the ViT-based visual encoder $E_I(\cdot)$, which processes source and target images $x_i^{S_m}/x_i^T$ to generate the corresponding domain-specific visual features $f_i^{S_m}/f_i^T$; and (2) the Transformer-based textual encoder $E_T(\cdot)$, which maps K adaptive LLM-generated positive prompts T_P, as well as K fixed template-derived negative prompts T_N into textual features h_P/h_N, forming the preliminary CLIP-based textual space. Based on the proposed modal-task joint optimization paradigm, we define two strategies for h_P/h_N with different semantics and $f_i^{S_m}/f_i^T$ from different domains, including Intra-modal Space-specific Optimization (ISO) and Cross-modal Emotion-aware Interaction (CEI). As a preliminary foundation, ISO refines the textual space by increasing the distance between h_P/h_N using $\mathcal{L}_{\text{ESS}}$, while aligning the latent visual space by minimizing the distance between $f_i^{S_m}/f_i^T$ through $\mathcal{L}_{\text{LDA}}$. Building upon ISO, CEI enables comprehensive vision-language interaction and generates both positive- and negative-oriented predictions ($P_{P,i}^{S_m}/P_{P,i}^T$ and $P_{N,i}^{S_m}/P_{N,i}^T$), contributing to four joint losses $\mathcal{L}_{\text{PS}}$, $\mathcal{L}_{\text{PT}}$, $\mathcal{L}_{\text{NS}}$, and $\mathcal{L}_{\text{NT}}$.

Bilateral Textual Prompts

PromptEmo adapts CLIP with FER-tailored textual semantics using bilateral prompts. Positive prompts use LLM-generated emotion descriptions, combined with trainable tokens to enrich high-level semantics for seen classes. Negative prompts, derived from templates, capture unseen-class features by excluding specific seen-class patterns.

Positive Textual Prompts. Standard CLIP prompts, such as "a photo of {*class*}," are limited in expressiveness in the MO-FER task, as they fail to describe fine-grained emotions and capture the subtle differences between FER classes. To address it, PromptEmo utilizes a pre-trained large language model (LLM), i.e., ChatGPT-4o (Hurst et al. 2024), to mine richer semantic details. Specifically, we pose structured queries to LLM for each seen class, such as: *"Please list the typical visual features of* {*class*} *from the perspective of facial behavior,"* and refine the responses through multi-turn dialogues, yielding detailed descriptions of facial movements (shown in Figure 2(a)). PromptEmo also introduces a trainable prompt vector with learnable tokens, which is concatenated with the LLM-generated descriptions to adaptively refine positive prompts for each class. The k-th ($k \in [1, K]$) positive prompt $T_P(k) \in R^{N_T \times L}$ is defined as:

$$T_P(k) = [\delta_k, tok(LLM(k))], \quad \delta_k \in R^{N_0 \times L}, \quad (1)$$

where N_T and N_0 are the token dimensions of complete textual prompts and the trainable vector δ_k, and L is the embedding dimension. $LLM(k)$ is the LLM-generated description for the k-th seen class, and $tok(\cdot)$ denotes tokenization. Then, $T_P(k)$ is passed through the frozen CLIP textual encoder $E_T(\cdot)$ to yield the k-th positive embedding $h_P(k)$.

Negative Textual Prompts. In OSR, negative representation learning aids in uncovering discriminative cues for unseen classes by creating a semantically opposite space for seen classes (Chen et al. 2021a). In the MO-FER task, detecting negative patterns from specific known FER categories also enhances the model's ability to distinguish subtle differences between seen and unseen expressions. Specifically, PromptEmo leverages templates with "*not*"-phrases to form explicit negative prompts, encouraging the model to explore patterns that do not belong to known expression categories. The k-th negative prompt is expressed as:

$$T_N(k) = tok(\text{"}\textit{This expression is not } \{\textit{class k}\}\text{"}), \quad (2)$$

where $T_N(k) \in R^{N_T \times L}$ is also fed into $E_T(\cdot)$ to yield the k-th negative embedding, aiming to exclude features from the k-th seen class. The CLIP-based textual space is shaped by K positive embeddings $h_P = \{h_P(1), h_P(2), \ldots, h_P(K)\}$ and K negative ones $h_N = \{h_N(1), h_N(2), \ldots, h_N(K)\}$.

Modal-Task Optimization Paradigm

When adapting the CLIP architecture with bilateral textual prompts to the MO-FER task with dual shifts (multi-domain and class), the general contrastive objective of CLIP is no longer applicable. In this case, objectives should be considered from two modal-specific aspects: (1) the semantic differences in textual prompts (positive and negative) designed for the open-set setting, and (2) the distributional differences in visual domains (source and target features) inherent in the multi-domain setting. This insight leads to our ***Modal-Task Optimization Paradigm***, shown in Figure 2(d), which includes two strategies: Intra-modal Space-specific Optimization (ISO) with two specific losses, and Cross-modal Emotion-aware Interaction (CEI) with four joint losses.

Intra-modal Space-specific Optimization. As shown in Figure 2(b), ISO learning strategies are tailored to the inherent structural properties of modal-specific feature spaces.

For the *textual space*, semantic distinctions are enforced based on explicitly-defined exclusive relations of bilateral prompts (T_P and T_N), encouraging the formation of a robust and well-clustered textual space that can effectively guide cross-domain visual representations. To achieve this, an *Explicit Semantic Separation (ESS)* loss is proposed to adjust the learnable tokens δ_k to adaptively push the textual embeddings h_P and h_N apart in textual feature space:

$$\mathcal{L}_{\text{ESS}} = 1 - \|h_P - h_N\|_2, \quad (3)$$

where $\|\cdot\|_2$ denotes the L2 normalization loss.

For the *visual space*, domain alignment is performed to mitigate multi-domain shifts between sources and the target by aligning feature distributions into a shared latent space. The *Latent Domain Alignment (LDA)* loss is formulated as:

$$\mathcal{L}_{\text{LDA}} = \sum_{m=1}^{M} \|\bar{\phi}^{S_m} - \bar{\phi}^T\|_\kappa^2, \quad (4)$$

$$\bar{\phi}^a = \frac{1}{|B^a|}\sum_{i=1}^{|B^a|}\phi_i^a, \quad \phi_i^a = \hat{D}_\kappa(f_i^a), \quad a = S_m \,/\, T,$$

where $\bar{\phi}^a$ is the average mapping of samples in the m-th source or the target batch B^a, and $\hat{D}_\kappa(\cdot)$ is the mapping to the Reproducing Kernel Hilbert Space. By optimizing $\mathcal{L}_{\text{LDA}}$, our PromptEmo is guided to learn domain-invariant visual features, thereby alleviating the multi-domain shift.

Cross-modal Emotion-aware Interaction. As shown in Figure 2(c), with ISO as the modal-specific representation foundation, CEI enables cross-modal interaction by deriving two types of predictions and four joint losses, enhancing the model's awareness of domain-invariant discriminative emotion features from both seen and unseen classes.

For known classes, we follow CLIP's procedure and compute the cosine similarity $cos(\cdot)$ between visual features f_i^a and all positive embeddings h_P, where *higher $cos(\cdot)$ values indicate greater similarity*. This results in the positive-oriented prediction $P_{P,i}^a$ for both source and target domains:

$$P_{P,i}^a = \textit{softmax}\left(cos(f_i^a, h_P)\right), \quad a = S_m/T. \tag{5}$$

To promote negative representation learning, we calculate the Euclidean distance $dis(\cdot)$, *where smaller $dis(\cdot)$ values indicate greater similarity*, between f_i^a and all negative embeddings h_N, which helps filter out patterns with *"related-but-opposite"* correlations. Specifically, for a sample with class k, a dual negation is applied to negate the existence of patterns specific to class k in other classes *(with greater similarity to $h_N(j \neq k)$ and smaller $dis(\cdot)$ values)*, suggesting that the sample is more likely to exhibit patterns from the k-th class *(with a larger $dis(\cdot)$ value to $h_N(k)$)*. The negative-oriented prediction is computed as:

$$P_{N,i}^a = \textit{softmax}\left(dis(f_i^a, h_N)\right), \quad a = S_m/T. \tag{6}$$

With $P_{P,i}^a$ and $P_{N,i}^a$, CEI employs four joint cross-entropy classification terms, ensuring comprehensive vision-language interaction. For positive-oriented predictions $P_{P,i}^a$, $P_{P,i}^{S_m}$ is supervised by the real label $y_i^{S_m}$ to learn discriminative features from seen classes, while $P_{P,i}^T$ is supervised by the pseudo-label $\tilde{y}_{P,i}^T$, assigned when maximum in $P_{P,i}^T$ exceeds the threshold τ. Two positive-oriented loss terms are:

$$\mathcal{L}_{\text{PS}} = \frac{1}{|B^{S_m}|}\sum_{i=1}^{|B^{S_m}|}\text{CE}\left(P_{P,i}^{S_m}, y_i^{S_m}\right), \tag{7}$$

$$\mathcal{L}_{\text{PT}} = \frac{1}{|B^T|}\mathbf{1}\left[\text{Max}(P_{P,i}^T) > \tau\right]\sum_{i=1}^{|B^T|}\text{CE}\left(P_{P,i}^T, \tilde{y}_{P,i}^T\right). \tag{8}$$

Negative-oriented predictions $P_{N,i}^a$ offer an extra view to enhance recognition of unseen classes. Similarly, two negative terms are defined, with $\tilde{y}_{N,i}^T$ derived from $P_{N,i}^T$:

$$\mathcal{L}_{\text{NS}} = \frac{1}{|B^{S_m}|}\sum_{i=1}^{|B^{S_m}|}\text{CE}\left(P_{N,i}^{S_m}, y_i^{S_m}\right), \tag{9}$$

$$\mathcal{L}_{\text{NT}} = \frac{1}{|B^T|}\mathbf{1}\left[\text{Max}(P_{N,i}^T) > \tau\right]\sum_{i=1}^{|B^T|}\text{CE}\left(P_{N,i}^T, \tilde{y}_{N,i}^T\right). \tag{10}$$

Training and Inference

In PromptEmo, the total training loss $\mathcal{L}_{\text{total}}$ is a weighted sum of loss terms in ISO and CEI, with the four CEI terms combined into a source term $\mathcal{L}_S$ and a target term $\mathcal{L}_T$:

$$\mathcal{L}_S = \sum_{m=1}^{M}(\mathcal{L}_{\text{PS}} + \mathcal{L}_{\text{NS}}), \quad \mathcal{L}_T = \mathcal{L}_{\text{PT}} + \mathcal{L}_{\text{NT}}, \tag{11}$$

$$\mathcal{L}_{\text{total}} = \left[\mathcal{L}_S + \alpha\mathcal{L}_T\right] + \left[\beta\mathcal{L}_{\text{ESS}} + \gamma\mathcal{L}_{\text{LDA}}\right], \tag{12}$$

where α, β, and γ are used to balance different terms.

During inference, the model generates final predictions by averaging the positive and negative-oriented target predictions, $P_{P,i}^T$ and $P_{N,i}^T$, and employs a simple dynamic threshold strategy from DAML (Noguchi and Shirakawa 2024) to distinguish between seen and unseen class samples.

Experiments

Experimental Details

Implementation Details. We implemented PromptEmo in PyTorch and conducted experiments on an RTX 3090 GPU. Preprocessing was performed using RetinaFace (Deng et al. 2020), resizing images to 224×224. A pretrained ViT-B/32 CLIP with a fixed textual encoder was utilized. The learning rates for visual encoder $E_I(\cdot)$ and learnable prompt δ_k were set to 0.00001 and 0.001. Training ran for 50 epochs with a batch size of 256, adjusting the learning rate every 10 epochs using MultiStepLR with a gamma of 0.1. Results were averaged across three runs with different seeds. Hyper-parameters were set to $\alpha = 0.1$, $\beta = 0.3$, $\gamma = 0.2$, $\tau = 0.8$, and the learnable vector length, N_0, was set to 8.

Dataset and Settings. For datasets, we selected four widely used public FER benchmarks, following the division scheme from OS-FER (Zhang et al. 2024b): (1) **AffectNet** (Mollahosseini, Hasani, and Mahoor 2017) is the most challenging FER dataset, from which we extracted 280,000 training images and 4,000 test images. (2) **Oulu-CASIA** (Zhao et al. 2011) comprises 2,880 facial expression video sequences from 80 subjects under laboratory lighting conditions, yielding a total of 2,988 images. (3) **RAF-DB** (Li, Deng, and Du 2017) consists of 29,672 facial expression images from real-world scenarios, with 12,271 for training and 3,068 for testing. (4) **FERPlus** (Barsoum et al. 2016) expands on FER2013 with 35,887 grayscale images, offering 28,709 training, 3,589 validation, and 3,589 test images.

For our proposed MO-FER task, it focuses on two real-world settings in FER. For the multi-domain setting, we treat each dataset as a target domain for evaluation, with the others serving as source domains. For the open-set setting, we assign one of the seven basic expressions as the unseen class in the target domain, with the others as seen classes. Notably, the training sets for source domains only contain seen classes, while the testing set for the target domain includes both seen and unseen expressions. Following Zhang et al. (Zhang et al. 2024b), we use Area Under Curve to evaluate performance on the *Unseen classes* (denoted AUC-*U* [%]), assessing the open-set recognition ability of the model. Classification accuracy is used for distinguishing *Seen classes* (denoted as ACC-*S* [%]).

Datasets	AffectNet								Oulu-CASIA							
Methods	Sur.	Fea.	Dis.	Hap.	Sad.	Ang.	Neu.	Avg.	Sur.	Fea.	Dis.	Hap.	Sad.	Ang.	Neu.	Avg.
Baseline	39.42	36.07	49.00	42.83	43.51	48.75	48.17	43.96	44.39	45.46	48.23	50.57	50.86	49.40	54.10	49.00
CoOp	58.55	41.09	63.69	54.09	61.59	64.69	65.00	58.39	46.68	55.29	63.73	48.10	49.13	66.22	60.58	55.68
NENO	55.91	55.68	55.68	60.94	61.48	56.61	59.30	57.94	57.04	52.92	52.84	59.56	56.82	59.67	56.50	56.48
OS-FER	55.16	43.04	58.49	55.25	61.35	60.79	62.60	56.67	52.15	49.71	33.35	54.62	59.03	49.82	64.23	51.84
DUML	60.30	55.80	50.47	70.79	62.94	67.73	67.21	62.18	58.11	54.30	57.22	60.21	57.14	60.00	58.23	57.89
LA-CMFER	58.82	59.84	52.29	74.53	64.17	**74.21**	68.98	64.69	59.71	55.62	54.21	**61.17**	58.04	59.69	58.07	58.07
PromptEmo	**62.15**	**61.74**	**65.43**	**76.07**	**67.67**	69.48	**69.60**	**67.45**	**70.83**	**57.25**	**67.60**	52.39	**76.41**	**71.95**	**85.50**	**68.85**
Datasets	FERPlus								RAF-DB							
Methods	Sur.	Fea.	Dis.	Hap.	Sad.	Ang.	Neu.	Avg.	Sur.	Fear	Dis.	Hap.	Sad.	Ang.	Neu.	Avg.
Baseline	40.17	44.19	43.93	43.55	49.19	44.81	50.31	45.16	43.57	58.62	54.57	52.96	55.04	56.77	57.04	54.08
CoOp	57.84	60.26	56.58	61.70	50.28	62.42	64.71	59.11	60.77	52.09	69.73	62.39	73.88	56.86	73.26	64.14
NENO	50.68	57.17	52.69	52.85	62.81	57.27	60.69	56.31	67.22	65.79	64.35	57.06	63.71	62.21	65.74	63.73
OS-FER	49.38	56.17	54.35	63.78	56.02	54.30	53.64	55.38	55.77	56.51	66.31	51.91	65.10	51.68	66.36	59.09
DUML	51.60	59.79	55.42	54.08	66.14	60.27	65.73	59.00	69.85	68.01	69.22	59.05	67.66	66.97	74.54	67.90
LA-CMFER	54.62	61.69	57.90	54.32	68.73	60.63	66.24	60.59	66.77	66.85	69.87	**69.44**	72.83	55.42	80.02	68.74
PromptEmo	**76.80**	**64.04**	**70.35**	**70.41**	**69.61**	**71.65**	**71.74**	**70.66**	**79.85**	**69.18**	**72.00**	68.43	**78.03**	**80.80**	**83.08**	**75.91**

Table 1: Comparison of AUC-*U* [%] on four public FER datasets for *Unseen Classes*.

Methods	Sur.	Fea.	Dis.	Hap.	Sad.	Ang.	Neu.	Avg.
Baseline	29.68	37.17	30.66	21.81	34.87	31.67	33.13	31.28
CoOp	42.68	41.03	42.98	27.24	39.54	39.85	52.56	40.84
LA-CMFER	44.82	61.91	**59.63**	40.61	**61.90**	39.64	50.23	51.25
PromptEmo	**57.62**	**64.58**	55.83	**55.66**	56.67	**63.25**	**66.22**	**59.98**

Table 2: ACC-*S* [%] on RAF-DB for *Seen Classes*.

Comparison with the State-of-the-art Methods

Compared Methods. We compare PromptEmo with state-of-the-art methods, including: 1) Baseline, which combines MMD loss (Zhu, Zhuang, and Wang 2019) for cross-domain alignment and SoftMax (Hendrycks and Gimpel 2016) scores for open-set recognition; 2) CoOp (Zhou et al. 2022), which uses CLIP model with adaptive prompts; 3) NENO (Li, Yang, and Hu 2023), a MDA method for open-set scenarios; 4) OS-FER (Zhang et al. 2024b), the first method for OS-FER task based on noise label learning; and 5) two latest CMFER methods: DUML (Liu et al. 2023) and LA-CMFER (Yang et al. 2024b). Results were either reported from original papers or reimplemented from publicly available codes. For fairness, methods not originally designed for open-set settings also use dynamic thresholding from DAML to identify unseen classes.

Comparison Results for Unseen Classes. We compared PromptEmo across all datasets to evaluate its OSR performance in the MO-FER task. As shown in Table 1, PromptEmo outperforms all methods in MO-FER tasks, with a notable average improvement of 7.7% AUC over the second-best method, LA-CMFER. Specifically, for CMFER Methods, DUML and LA-CMFER perform well on AffectNet but exhibit instability on certain classes (e.g., Dis.), and they also fall far behind PromptEmo on other datasets. The Baseline, which combines simple setting-specific techniques, struggles with domain and class shifts, leading to the poorest performance. For CoOp, while effective in many other tasks by adapting from CLIP, it lacks sufficient generalizability in the MO-FER task. NENO, designed for multi-domain OSR in natural images, performs less effectively on FER datasets due to the significant domain gap. In contrast, PromptEmo consistently shows superior performance, validating its effectiveness in recognizing unseen target samples.

Comparison Results for Seen Classes. In the MO-FER task, maintaining the discriminability of seen-class samples is also critical. As shown in Table 2, while the SOTA CMFER method, LA-CMFER, performs well for certain seen expressions, our PromptEmo, designed with a broader scope to handle both seen and unseen classes, outperforms it on all other seen classes, achieving an average 8.73% higher ACC. The poor performance of the Baseline and CoOp can be attributed to their lack of designs to effectively address domain shifts and inter-class similarities in the FER field.

In summary, by adapting CLIP framework with bilateral prompts and an elaborate modal-task optimization paradigm, PromptEmo injects high-level textual guidance to a cross-domain multi-modal feature space, enabling accurate identification of unseen emotions while preserving superior discriminability of seen ones.

Ablation Studies

Analysis of Key Components. Using AUC-*U* [%] for unseen classes and ACC-*S* [%] for seen classes, we conducted an ablation study to comprehensively evaluate the contribution of each module. Starting with a Baseline that uses a sim-

Ablation Models	AUC-*U* [%]	ACC-*S* [%]
Baseline	58.57	38.33
+ BTP & CEI (*Pos*)	66.64 (+8.07)	51.89 (+13.56)
+ BTP & CEI (*Neg*)	69.08 (+2.44)	55.51 (+3.62)
+ ISO ($\mathcal{L}_{\text{ESS}}$)	69.53 (+0.45)	57.28 (+1.77)
+ ISO ($\mathcal{L}_{\text{LDA}}$)	70.66 (+1.13)	58.04 (+0.76)

Table 3: Ablation studies of key components on FERPlus.

Positive Textual Prompts	AUC-*U* [%]	ACC-*S* [%]
(A) "a photo of $\{class\}$"	68.16	54.12
(B) "an expression of $\{class\}$"	68.23	55.53
(C) LLM-descriptions w/o δ_k	70.24	56.63
(D) LLM-descriptions w/ δ_k	70.66	58.04

Table 4: Analysis of positive prompt designs on FERPlus.

ple supervision loss for source domains with standard CLIP prompts, we progressively added the following components: (1) the proposed positive prompts T_P in bilateral textual prompts (BTP) and positive-oriented loss terms $\mathcal{L}_{\text{PS}}$/$\mathcal{L}_{\text{PT}}$ in the CEI strategy (denoted as BTP & CEI (*Pos*)), (2) their negative counterparts (*Neg*), i.e., T_N in BTP and $\mathcal{L}_{\text{NS}}$/$\mathcal{L}_{\text{NT}}$ in CEI, (3) Explicit Semantic Separation loss $\mathcal{L}_{\text{ESS}}$ for textual space in the ISO strategy, and (4) Latent Domain Alignment loss $\mathcal{L}_{\text{LDA}}$ for visual space in ISO, forming our PromptEmo.

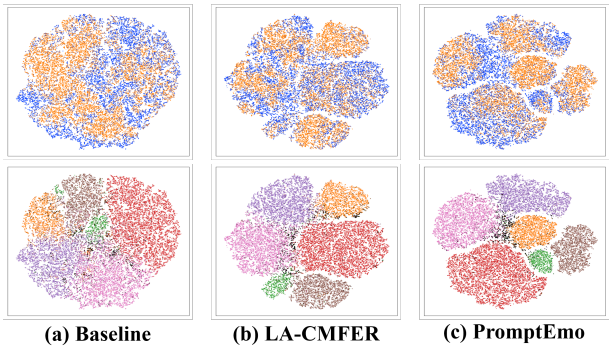
As shown in Table 3, integrating adaptive positive prompts with positive-oriented loss terms in CEI significantly improves the ability to recognize both seen and unseen expressions. The negative designs in prompt and CEI help exclude patterns from seen classes, thereby enhancing learning for unseen classes. ISO contributes by separating textual semantics and aligning visual domains, enabling more robust representations for CEI.

Analysis of Positive Prompt Design. We compared our adaptive LLM-generated positive prompt with three other designs: (A) the standard "a photo of $\{class\}$" prompt in CLIP, (B) the explicit "an expression of $\{class\}$" prompt for FER, and (C) an LLM-generated prompt without adaptive vectors δ_k. As shown in Table 4, our prompt method employs the widely-used GPT-4o with structured queries, outperforming two manually-designed prompts (A) and (B) by yielding high-quality descriptions that better capture fine-grained facial behavior knowledge of seen class expressions. Additionally, comparisons between (C) and (D) demonstrate that the inclusion of trainable prompt vectors δ_k ensures refined positive prompts in ISO optimization, leading to a more robust textual feature space and improved recognition performance of seen classes.

Analysis of CLIP Backbone. To determine whether the performance gains of PromptEmo arise solely from the generalized CLIP backbone, we enhanced LA-CMFER by adding an extra CLIP textual encoder with its standard prompts. As shown in Table 5, simply incorporating CLIP does not yield evident performance boosts for LA-CMFER, and CoOp also performs suboptimal in classifying seen classes. This suggests that direct adaptation of CLIP does not always yield favorable results in open-set tasks. In contrast, prompt designs and optimization strategies in PromptEmo are specifically tailored to both the MO-FER task and the multimodal architecture of CLIP, resulting in superior performance.

Methods	AUC-*U* [%]	ACC-*S* [%]
CoOp	59.11	38.02
LA-CMFER	60.59	39.85
LA-CMFER + CLIP	60.67	40.23
PromptEmo	70.66	58.04

Table 5: Incorporating CLIP backbone on FERPlus.



(a) Baseline (b) LA-CMFER (c) PromptEmo

Figure 3: Visualizations of feature embeddings on the target RAF-DB. The first row shows domain distributions (orange for source, blue for target), and the second row shows class distributions (black for unseen, others for seen classes).

Analysis of t-SNE Visualization. We also employed t-SNE visualizations to illustrate cross-domain transferability. As shown in Figure 3, the Baseline shows weak unseen-class recognition with ambiguous boundaries. While LA-CMFER forms clearer seen-class clusters, unseen-class samples still overlap greatly with seen ones. In contrast, PromptEmo achieves compact seen-class clusters and maps unseen-class samples to one independent region, showcasing superior cross-domain transfer and open-set recognition performance that aligns with its design goals for the MO-FER task.

More Details. Detailed descriptions of positive prompts, explorations of $\alpha, \beta, \gamma, \tau$ in loss terms, and adaptive vector length N_0 are provided in the supplementary materials.

Conclusion

We present PromptEmo, a novel CLIP-based framework designed to address the novel and challenging Multi-domain Open-set Facial Expression Recognition (MO-FER) task, which involves both multi-domain and class shifts. Specifically tailored for MO-FER and CLIP, PromptEmo devises bilateral textual prompts and integrates Intra-modal Space-specific Optimization (ISO) and Cross-modal Emotion-aware Interaction (CEI) strategies within a modal-task optimization paradigm. Experiments on four FER benchmarks show that PromptEmo effectively recognizes unseen emotions while maintaining the discrimination of seen ones.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62371325, in part by Sichuan Science and Technology Program under Grant 2025NSFJQ0050, in part by Sichuan Science and Technology Program under Grant 2024ZDZX0018, and in part by the Key Laboratory of Internet Natural Language Processing of Sichuan Provincial Education Department under Grant INLP202402.

References

Barsoum, E.; Zhang, C.; Ferrer, C. C.; and Zhang, Z. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, 279–283.

Chen, G.; Peng, P.; Wang, X.; and Tian, Y. 2021a. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 8065–8081.

Chen, T.; Pu, T.; Wu, H.; Xie, Y.; Liu, L.; and Lin, L. 2021b. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(12): 9887–9903.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5203–5212.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gan, Z.; Li, L.; Li, C.; Wang, L.; Liu, Z.; Gao, J.; et al. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4): 163–352.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ji, Y.; Hu, Y.; Yang, Y.; Shen, F.; and Shen, H. T. 2019. Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network. *Neurocomputing*, 333: 231–239.

Ji, Y.; Hu, Y.; Yang, Y.; and Shen, H. T. 2021. Region attention enhanced unsupervised cross-domain facial emotion recognition. *IEEE Transactions on Knowledge and Data Engineering*, 35(4): 4190–4201.

Li, J.; Yang, L.; and Hu, Q. 2023. Enhancing multi-source open-set domain adaptation through nearest neighbor classification with self-supervised vision transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4): 2648–2662.

Li, S.; and Deng, W. 2020. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3): 1195–1215.

Li, S.; Deng, W.; and Du, J. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2852–2861.

Li, Y.; Gao, Y.; Chen, B.; Zhang, Z.; Zhu, L.; and Lu, G. 2021. JDMAN: Joint discriminative and mutual adaptation networks for cross-domain facial expression recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3312–3320.

Liu, H.; Cai, H.; Lin, Q.; Li, X.; and Xiao, H. 2023. Learning from more: Combating uncertainty cross-multidomain for facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5889–5898.

Luo, Y.; Wang, Z.; Huang, Z.; and Baktashmotlagh, M. 2020. Progressive graph learning for open-set domain adaptation. In *International Conference on Machine Learning*, 6468–6478. PMLR.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1): 18–31.

Noguchi, M.; and Shirakawa, S. 2024. Simple domain generalization methods are strong baselines for open domain generalization. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rakshit, S.; Tamboli, D.; Meshram, P. S.; Banerjee, B.; Roig, G.; and Chaudhuri, S. 2020. Multi-source open-set deep adversarial domain adaptation. In *European conference on computer vision*, 735–750. Springer.

Wang, L.; Jia, G.; Jiang, N.; Wu, H.; and Yang, J. 2022. Ease: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks. In *Proceedings of the 30th ACM international conference on multimedia*, 218–227.

Yang, Y.; Hou, Y.; Wen, L.; Zeng, P.; and Wang, Y. 2024a. Semantic-aware adaptive prompt learning for universal multi-source domain adaptation. *IEEE Signal Processing Letters*, 31: 1444–1448.

Yang, Y.; Wen, L.; Zeng, X.; Xu, Y.; Wu, X.; Zhou, J.; and Wang, Y. 2024b. Learning with alignments: tackling the inter-and intra-domain shifts for cross-multidomain facial expression recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4236–4245.

Yang, Y.; Zeng, X.; Zeng, P.; Zu, C.; Yan, B.; Zhou, J.; and Wang, Y. 2025. Adaptive Hardness-Driven Augmentation

and Alignment Strategies for Multisource Domain Adaptations. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15.

Yu, X.; Yoo, S.; and Lin, Y. 2024. Clipceil: Domain generalization through clip via channel refinement and image-text alignment. *Advances in Neural Information Processing Systems*, 37: 4267–4294.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024a. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8): 5625–5644.

Zhang, W.; Song, P.; and Zheng, W. 2022. Joint local-global discriminative subspace transfer learning for facial expression recognition. *IEEE Transactions on Affective Computing*, 14(3): 2484–2495.

Zhang, Y.; Yao, Y.; Liu, X.; Qin, L.; Wang, W.; and Deng, W. 2024b. Open-set facial expression recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 646–654.

Zhao, G.; Huang, X.; Taini, M.; Li, S. Z.; and PietikäInen, M. 2011. Facial expression recognition from near-infrared videos. *Image and vision computing*, 29(9): 607–619.

Zhao, H.; Zhang, S.; Wu, G.; Moura, J. M.; Costeira, J. P.; and Gordon, G. J. 2018. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31.

Zhao, Z.; and Liu, Q. 2021. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM international conference on multimedia*, 1553–1561.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, Y.; Zhuang, F.; and Wang, D. 2019. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 5989–5996.