PeerJ Computer Science

A novel facial expression recognition framework using deep learning based dynamic cross-domain dual attention network

Ahmed Omar Alzahrani[1], Ahmed Mohammed Alghamdi[2], M. Usman Ashraf[3], Iqra Ilyas[3], Nadeem Sarwar[4], Abdulrahman Alzahrani[1] and Alaa Abdul Salam Alarood[1]

[1] Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah, Makkah, Saudi Arabia
[2] Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah, Makkah, Saudi Arabia
[3] Department of Computer Science, Government College Women University Sialkot, Sialkot, Punjab, Pakistan
[4] Department of Computer Science, Bahria University, Lahore, Punjab, Pakistan

ABSTRACT

Variations in domain targets have recently posed significant challenges for facial expression recognition tasks, primarily due to domain shifts. Current methods focus largely on global feature adoption to achieve domain-invariant learning; however, transferring local features across diverse domains remains an ongoing challenge. Additionally, during training on target datasets, these methods often suffer from reduced feature representation in the target domain due to insufficient discriminative supervision. To tackle these challenges, we propose a dynamic cross-domain dual attention network for facial expression recognition. Our model is specifically designed to learn domain-invariant features through separate modules for global and local adversarial learning. We also introduce a semantic-aware module to generate pseudo-labels, which computes semantic labels from both global and local features. We assess our model's effectiveness through extensive experiments on the Real-world Affective Faces Database (RAF-DB), FER-PLUS, AffectNet, Expression in the Wild (ExpW), SFEW 2.0, and Japanese Female Facial Expression (JAFFE) datasets. The results demonstrate that our scheme outperforms the existing state-of-the-art methods by attaining recognition accuracies 93.18, 92.35, 82.13, 78.37, 72.47, 70.68 respectively.

Subjects Algorithms and Analysis of Algorithms, Artificial Intelligence, Computer Vision, Data Mining and Machine Learning, Emerging Technologies
Keywords Artificial intelligence, Facial expression recognition, Deep learning, Cross-domains

INTRODUCTION

Facial expression recognition (FER) has solidified its role as a pivotal field within computer vision, driving advancements across diverse applications like human-computer interaction, security, mental health diagnostics, and personalized marketing. Despite these applications, FER faces substantial challenges related to domain shift, where variations

Submitted 27 January 2025
Accepted 10 April 2025
Published 9 May 2025

Corresponding author
M. Usman Ashraf,
usman.ashraf@gcwus.edu.pk

Academic editor
Bilal Alatas

Additional Information and Declarations can be found on page 20

DOI **10.7717/peerj-cs.2866**

How to cite this article Alzahrani AO, Alghamdi AM, Ashraf MU, Ilyas I, Sarwar N, Alzahrani A, Alarood AAS. 2025. A novel facial expression recognition framework using deep learning based dynamic cross-domain dual attention network. **PeerJ Comput. Sci. 11:e2866 DOI 10.7717/peerj-cs.2866**

between source and target domains introduce obstacles to reliable recognition. These variations arise due to differences in factors such as cultural backgrounds, lighting conditions, facial poses, and expressions, which can lead to significant drops in model performance when deployed in real-world scenarios or across new datasets (*Han et al., 2020*). Most current FER methodologies rely on global feature adaptation techniques to derive domain-invariant features (*Tzeng et al., 2017*; *Yang et al., 2024*). While these methods have demonstrated some success, they fall short in capturing and transferring local features, such as subtle eye or mouth movements, across domains. This shortfall in local feature adaptability leaves a notable gap in FER models, as these fine-grained features are crucial for nuanced emotion detection. Additionally, when training on target-specific datasets, models often encounter a decline in feature representation due to limited discriminative supervision, leading to lower performance in differentiating between closely related expressions.

To address these domain-shift challenges, numerous FER techniques have emerged, applied across a variety of datasets, including Japanese Female Facial Expression (JAFFE), Oulu-CASIA (*Oulu-CASIA, 2024*), SFEW 2.0, Real-world Affective Faces Database (RAF-DB), FER2013, FERPLUS, CK+, Expression in the Wild (ExpW), and AffectNet. These datasets vary widely in sample distribution, demographic diversity, and contextual factors, adding complexity to cross-domain learning. Traditional solutions, such as transfer learning (*Orozco et al., 2018*) and supervised kernel matching, have attempted to alleviate data inconsistencies across these datasets, achieving some improvement in performance. However, these methods often rely on extensive annotated samples in the target domain to create clear categorical distinctions, which is impractical for unsupervised cross-domain FER tasks. The limitations of these existing approaches reveal an urgent need for new methods that can manage unsupervised cross-domain settings, where access to labeled target data is minimal or unavailable. Further, the ability to consistently recognize expressions across diverse environments remains critical, as FER models must contend with variable poses, lighting conditions, occlusions, and even cultural differences in expression (*Perveen, Roy & Chalavadi, 2020*).

Recent advancements have explored alternative learning approaches, including dictionary learning (*Sun et al., 2023*), metric learning (*Huang et al., 2021*), and contrastive learning (*Yang et al., 2023*), to support unsupervised cross-domain facial expression recognition (CD-FER). Additionally, some methods (*Samadiani et al., 2019*; *Ben et al., 2021*; *Sampath et al., 2021*) focus on creating synthetic samples to reduce the feature distribution gap between source and target datasets, thereby improving cross-domain generalization. While these methods have contributed to the field by focusing on global feature adaptation for domain-invariant learning, a significant challenge remains in effectively transferring local features across diverse domains. Current methodologies excel in capturing global features to ensure consistency across datasets; however, they often fall short in handling the intricacies of local feature transfer. Local features typically contain essential, fine-grained information, such as subtle facial muscle movements, which are crucial for precise domain adaptation. This gap in local feature transfer presents a major challenge since the detailed aspects of expressions are often encoded in these features,

making them sensitive to variations in pose, lighting, and occlusions. Overcoming this limitation requires innovative approaches that can capture and reliably transfer local features across domains, ultimately enhancing the performance and robustness of FER systems in diverse settings. To address these limitations, we propose an adaptive cross-domain dual attention network for facial expression recognition, which incorporates specialized modules for both global and local adversarial learning. This structure is designed to improve the capture of domain-invariant features by combining global and local learning. Furthermore, we introduce a semantic-aware pseudo-label generation module that calculates semantic labels from both global and local feature sets, thus enhancing the model's generalization capacity across diverse domains. We validate our approach through extensive experiments using several of the most comprehensive FER datasets previously discussed. By addressing critical gaps in feature adaptation, our contributions provide a foundation for more robust and accurate FER systems suited to real-world applications.

Further our contribution can be summed up as follows:

- We propose a dynamic learning and selection model Dynamic Cross-Domain Dual Attention Network (DCD-DAN) for FCR for both global and local representation. In DCD-DAN model, feature refinement is performed by local interactions within the spatial dimension, while channel dimension is used for the provision of global receptive field.
- To address the challenges in activation functions, we propose a novel activation function construction (AFC) scheme. AFC scheme addresses the common issues such as massive computation overhead in power function, deactivation of neurons *etc.*
- Introduce self-attention condensation and group mechanism where intentions are divided into multiple groups, and implement self-attention condensation over every group. It minimizes the spatial dimensions that eventually bring down the computational cost significantly.
- Conduct comprehensive experiments to evaluate the significance of proposed DCD-DAN model. Implementation on variety of datasets including RAF-DB, FER-PLUS, AffectNet, ExpW, SFEW 2.0, and JAFFE, and compare with existing state-of-the-art techniques.

The rest of the article is organized in such way that "Literature Review" presents a comprehensive literature study, explaining existing state-of-the-art methods on facial recognition detection. In "Proposed Method", we present our proposed scheme comprehensively. Further "Implementation and Results" describes the implementation of the proposed scheme, and results compared with existing state-of-the-art methods. Finally, "Conclusion" concludes the study.

LITERATURE REVIEW

Now we present a comprehensive overview of existing state-of-the-art methods followed by the background of technologies used in this study.

Background

Facial expression recognition (FER) has emerged as a crucial area of research within the fields of computer vision and affective computing. Its applications span various domains, including human-computer interaction, security, mental health diagnostics, and marketing analytics (*Chhikara et al., 2020*). Despite its significance, achieving accurate FER remains challenging due to the inherent variability in facial expressions, pose variations, lighting conditions, occlusions, and domain shifts between datasets. Traditional FER methods (*Subudhiray, Palo & Das, 2023a*; *Subudhiray, Palo & Das, 2023b*; *Wang et al., 2019*; *Nigam, Singh & Misra, 2018*) often rely on supervised learning models trained on a single dataset, making them susceptible to performance degradation when tested on unseen datasets due to domain shifts. These shifts arise from differences in demographic diversity, expression intensity, image quality, and environmental factors. Consequently, models trained on one dataset may fail to generalize effectively to another, leading to poor cross-domain adaptability. To address this issue, researchers have explored domain adaptation techniques to improve generalization across datasets. Transfer learning, adversarial learning, and multi-domain learning have been widely adopted to reduce discrepancies in feature distributions between source and target datasets (*Zeeshan et al., 2024*). However, existing methods predominantly focus on global feature alignment, neglecting local feature variations, which are essential for capturing fine-grained facial muscle movements. This limitation reduces the effectiveness of FER models, particularly when dealing with subtle or ambiguous expressions.

The rise of deep learning has significantly improved FER accuracy. Convolutional neural networks (CNNs), residual network (ResNet) architectures (*Li & Lima, 2021*), and self-attention mechanisms (*Daihong, Lei & Jin, 2021*) have been employed to enhance feature extraction (*Borgalli & Surve, 2022*; *Borgalli & Surve, 2025*). Several state-of-the-art models, including self-cure network (SCN), radio access network (RAN), and EfficientFace, have introduced self-attention and relational learning modules to improve robustness against expression variations and occlusions. However, these models still suffer from domain shift issues, as they fail to explicitly adapt local feature representations across domains. Recent advancements have explored multi-scale learning, where models process both global and local features for improved FER. This approach has shown promise in capturing spatial dependencies while preserving fine-grained expression details. However, most existing methods do not integrate dual attention mechanisms that explicitly balance both global and local adversarial learning.

Related work

To address the domain discrepancies that commonly arise among various facial expression recognition (FER) datasets, several cross-domain FER algorithms have been proposed. For example, *Chen et al. (2021)* introduced Adversarial Graph Representation Adaptation (AGRA), a method combining graph representation propagation with adversarial learning. AGRA effectively co-adapts holistic and local features across domains by correlating local regions with holistic features. Specifically, AGRA leverages two stacked graph

convolutional networks (GCNs) to propagate these features, achieving maximum accuracies of 85% and 68% on the CK+ (*Shaik, 2021*).

Similarly, *Yan et al. (2019)* and *Xie et al. (2020)* proposed a discriminative feature adaptation technique that establishes a feature space capable of capturing facial expressions across domains. Their deep transfer network was designed to reduce bias between datasets, providing a more unified feature representation. *Li et al. (2021)* extended this approach by merging graph propagation with adversarial learning to create holistic-local domain-invariant features for cross-domain FER. Their method incorporates subspace learning to transfer knowledge from labeled source data to unlabeled target data, although some target annotations are still necessary.

Guo et al. (2024) explored challenges associated with data discrepancies and expression ambiguities. They observed that while many deep learning FER methods excel within a single dataset, transferring them to a new dataset incurs additional labeling costs. To address these issues, they proposed an unsupervised self-training similarity transfer (USTST) method for cross-domain FER, which minimizes the need for labeled data in the target domain. *Zhou et al. (2024)* later introduced a generative adversarial network (GAN)-based approach that combines transfer learning with generative adversarial networks. Their framework initially enhances training data through a face-cycle GAN to generate additional facial expressions and then deploys two FER networks based on CNN architectures to increase model robustness.

To further tackle real-world challenges, researchers have explored multi-view and multiscale studies. *Beaudry et al. (2014)* highlighted the significance of facial regions like the eyes and mouth in expression recognition, prompting methods that target these key areas. Deep learning advancements have bolstered feature extraction in these areas, with CNNs becoming instrumental. For instance, *Duan (2024)* developed the SCN model, which incorporates self-attention importance weighting, rank regularization, and relabeling modules. *Li et al. (2023)* introduced the RAN framework, which integrates convolutional operations with self-attention and relational attention modules to better capture intricate facial features. *Tan, Xia & Song (2024)* proposed EfficientFace, which enhances robustness through a local feature extractor and channel-spatial modulator, while *Zhang et al. (2024)* introduced Contrastive Syn-to-Real Generalization (CSG) ResNet, embedding Gabor Convolution (GConv) into ResNet to capture finer details. AMP-Net further builds on this by extracting global, local, and salient features at various granularities, reflecting the diversity and complexity of facial emotions. However, CNNs' limited receptive fields remain a constraint, prompting recent methods to combine CNN-based shallow feature extraction with self-attention mechanisms to capture high-level visual semantics effectively.

Recent research has increasingly focused on semantic-aware approaches for feature representation learning, aiming to bridge the semantic gap in domain alignment. Adversarial domain adaptation methods, for instance, have been employed to modify image appearances across domains while retaining semantic integrity. The approach in *Wang et al. (2024)* leveraged global-local and semantic learning to address domain

adaptation by developing domain-invariant global-local features. However, it relied on fixed criteria for pseudo-label generation, which might limit the range of expression classes that can be accurately labeled.

Despite significant advancements in cross-domain FER, existing approaches still exhibit several limitations that hinder their real-world applicability. Graph-based adversarial learning methods such as AGRA and holistic-local domain-invariant feature adaptation techniques improve feature representation but fail to effectively generalize across datasets with high domain discrepancies, especially in complex real-world scenarios. Furthermore, self-training and generative adversarial methods reduce the need for labeled target data but often suffer from expression ambiguity and feature distortion, leading to suboptimal recognition performance. While deep learning models in existing studies such as SCN, RAN, and EfficientFace leverage self-attention and convolutional mechanisms, they predominantly focus on global feature adaptation, neglecting the fine-grained local feature variations crucial for capturing subtle facial expressions but rely on fixed pseudo-labeling criteria, limiting their ability to adapt dynamically to target domain variations. To address these limitations, we propose the Dynamic Cross-Domain Dual Attention Network (DCD-DAN), which introduces a dual attention mechanism that integrates global and local adversarial learning to achieve domain-invariant representation. Unlike previous methods, our approach explicitly disentangles global and local feature extraction, ensuring fine-grained feature transfer across domains. Additionally, our semantic-aware pseudo-labeling module dynamically generates target domain labels, overcoming the rigid constraints of previous fixed-label adaptation techniques. Our approach, by contrast, emphasizes robust domain-invariant multi-scale feature learning through distinct global and local adversarial learning modules. Additionally, we maintain semantic consistency *via* a unified global-local prediction selection strategy, allowing for more flexible and accurate expression recognition across domains. This strategy enhances the reliability of cross-domain FER models, paving the way for more adaptable FER applications in real-world scenarios. Further proposed methodology details are presented in "Proposed Method".

PROPOSED METHOD

Our proposed Dynamic Cross-Domain Dual Attention Network based facial expression recognition scheme is designed to address the challenges of domain shifts in facial expression recognition (FER) tasks by learning domain-invariant features. The network integrates both global and local adversarial learning modules, combined with a semantic-aware module to generate pseudo-labels. This approach aims to enhance feature representation within the target domain, despite the absence of labeled data. Figure 1 presents a block diagram of our proposed scheme.

To capture the domain-invariant features, we utilize a dual feature extraction process (*Zhao et al., 2024*) that separately handles global and local features from the source and target domains. Given a source domain dataset X_s with corresponding labels Y_s and a target domain dataset X_t without labels, the network first extracts global features F_g^s and F_g^t

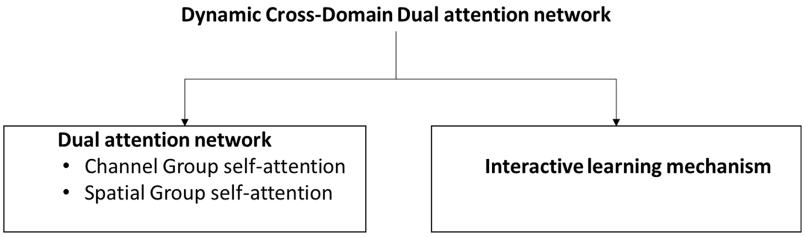

Figure 1 **Primary components of the proposed scheme.** Full-size DOI: 10.7717/peerj-cs.2866/fig-1

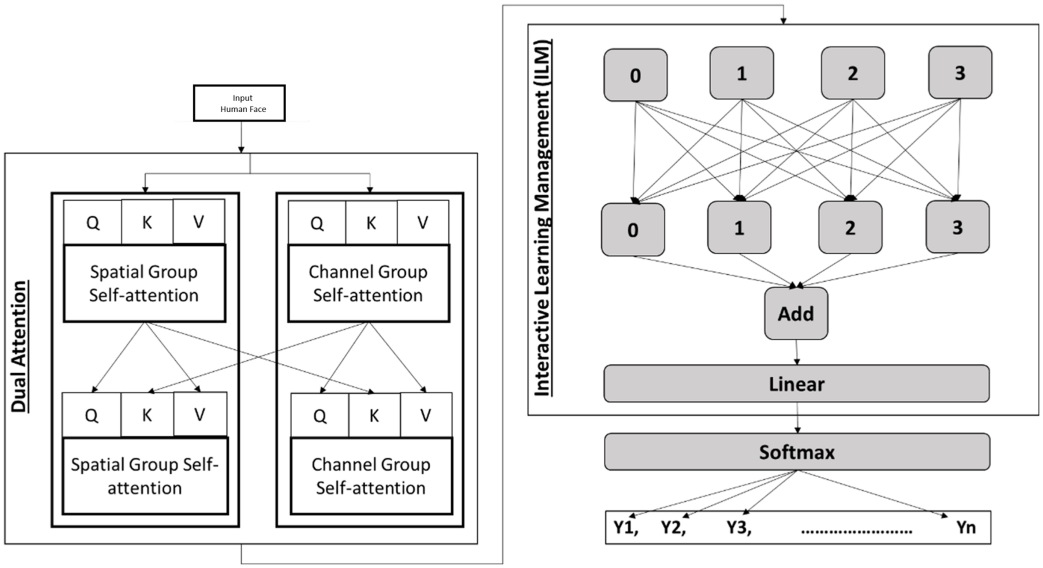

Figure 2 **Detailed interaction process of dual attention and interactive learning management.** Full-size DOI: 10.7717/peerj-cs.2866/fig-2

using a global feature extractor G parameterized by W_g. Mathematically, this is expressed as:

$$F_g^s = G(X_s;\ W_g) \quad \text{// source domain data}$$

$$F_g^t = G(X_t;\ W_g) \quad \text{// target domain data.}$$

Similarly, local features F_l^s and are F_l^t extracted using a local feature extractor L parameterized by W_l. The separation of global and local feature extraction allows the network to learn diverse aspects of the data, enhancing its ability to generalize across different domains. The detailed interaction process of dual attention and interactive learning management is presented in Fig. 2.

Further to ensure that the extracted features are domain-invariant, we introduce adversarial learning mechanisms at both global and local levels. For global adversarial learning, a discriminator D_g is employed to distinguish between the source and target domain global features.

Let F_g^s as the global feature extracted from source domain, and F_g^t global feature extracted from targeted domain. Similarly, $D_g\left(F_g^s\right)$ as the discriminator function that is the output of F input features. The goal of the discriminator is to correctly classify whether a given feature representation belongs to the source or target domain. Thus, its objective function is:

$$max\Big(\mathbb{E}F_g^s \sim X_s\Big[log\ D_g\left(F_g^s\right)\Big]\ +\ \mathbb{E}F_g^t \sim X_t\Big[\log\Big(1-\ D_g\left(F_g^t\right)\Big)\Big]\Big).$$

Therefore, the adversarial loss for global features is defined in Eq. (1) as follows:

$$\pounds_{adv}^g = -\Big(\mathbb{E}\Big[log\ D_g\left(F_g^s\right)\Big] + \mathbb{E}\Big[\log\Big(1-\ D_g\left(F_g^t\right)\Big)\Big]\Big). \tag{1}$$

For local features, a similar adversarial learning process is applied using a discriminator D_l, with the local adversarial loss given in Eq. (2):

$$\pounds_{adv}^l = -\big(\mathbb{E}\big[log\ D_l\left(F_l^s\right)\big]\ +\ \mathbb{E}\big[\log(1-\ D_l\left(F_l^t\right))\big]\big). \tag{2}$$

The adversarial loss functions used in Eqs. (1) and (2) follow the principles of domain adversarial learning, commonly employed in domain adaptation tasks (*Tang & Jia, 2020*). The goal of these adversarial losses is to make the global and local feature distributions from the source and target domains indistinguishable, thereby ensuring that the network learns domain-invariant features.

In the absence of labeled data in the target domain, we employ a semantic-aware module S to generate pseudo-labels $\widehat{y_t}$ for the target domain data. This module combines the global and local features F_l^s and are F_l^t to predict the labels, ensuring that the pseudo-labels reflect both global and local feature information. The pseudo-label generation is formulated as:

$$\widehat{y_t}\ = argmax\ \Big(S\Big(F_g^t, F_l^t; W_{sa}\ \Big)\Big).$$

This step is crucial for providing a form of discriminative supervision during the training phase, even in the absence of true labels in the target domain. Further, the training process involves minimizing a combined loss function that incorporates the classification loss on the source domain, along with the adversarial losses for both global and local features. The classification loss $\pounds_{cls}$ on the source domain is computed as given in Eq. (3):

$$\pounds_{cls} = \mathbb{E}\Big[cross\ entropy\ \Big(F_g^s +\ F_l^s,\ Y_s\Big)\Big]. \tag{3}$$

The overall loss function, which guides the updating of network parameters, is then formulated in Eq. (4):

$$W_g,\ W_l,\ W_{sa}\ \leftarrow\ minimize\ \big(\pounds_{cls} + \hat{\lambda}_g\pounds_{adv}^g + \hat{\lambda}_l\pounds_{adv}^l\big). \tag{4}$$

Similarly, Eq. (3), which represents the classification loss, is based on the widely used cross-entropy function in deep learning (*Ruby & Yendapalli, 2020*). Finally, Eq. (4) integrates these losses into a unified optimization framework to enhance domain-invariant feature learning, following the principles of multi-objective learning (*Liu et al., 2024*).

Here, λ_g and λ_l hyperparameters that control the contribution of the global and local adversarial losses, respectively. By minimizing this combined loss, the network is trained to extract features that are both discriminative for the task and invariant to domain shifts. Once the network is trained, it is used to predict the labels $\widehat{y_t}$ for each sample X_t in the target domain. The predicted labels are generated using the learned global and local features in combination with the semantic-aware module. The adversarial losses $£^g_{adv}$ and $£^g_{adv}$ enforce feature alignment between source and target domains, improving generalization. The classification loss $£_{cls}$ ensures that the model maintains high accuracy on labeled source data. The inclusion of the semantic-aware module W_{sa} further refines the extracted features to enhance robustness to domain discrepancies. The global adversarial loss $£^g_{adv}$ encourages domain-invariant features at a coarse level, while the local adversarial loss focuses on fine-grained local feature adaptation. By jointly optimizing these losses, the model learns to bridge the gap between source and target distributions more effectively. This multi-level adaptation mechanism ensures that features at different scales contribute to robust classification, leading to improved performance in the presence of domain shifts. Finally, the learned network parameters W_g, W_l, W_{sa} collectively define the optimal feature representation, enabling accurate predictions even in unseen target domains. Further the workflow of our proposed scheme is presented as follows:

1. Network initialization

At the beginning, the network components are initialized with specific weight parameters to enable optimal learning. The global adversarial learning module is initialized with weights W_g, while the local adversarial learning module is assigned weights W_l. Additionally, the semantic-aware module begins with weights W_{sa}. These initial settings lay the foundation for the network to accurately learn features from both source and target domains.

2. Global feature extraction

Using a global feature extractor G, the network extracts broad, domain-wide features from images in both source and target domains. This step captures overarching patterns and shapes relevant for facial expression recognition, allowing the model to develop a foundational understanding of the overall structure in the images. The global features ensure that the network can generalize across the datasets by capturing domain-level traits.

3. Local feature extraction

In tandem with global feature extraction, the network also utilizes a local feature extractor L to capture region-specific details in both source and target images. These local features focus on finer details, such as eye and mouth regions, which are critical for distinguishing subtle expressions. By combining global and local features, the model achieves a comprehensive feature representation that enhances recognition accuracy.

4. Global adversarial loss calculation

A global discriminator D_g is then employed to differentiate between global features from the source and target domains. By calculating the global adversarial loss, the network

learns to make these global features indistinguishable across domains. This adversarial training encourages the network to develop domain-invariant global features, which are essential for achieving robust recognition performance across domain shifts.

5. Local adversarial loss calculation

Similarly, a local discriminator D_l is utilized to apply adversarial learning to the local features. The network calculates the local adversarial loss, aiming to make local features indistinguishable between the source and target domains. This process ensures that even the region-specific, fine-grained features are domain-invariant, helping the model generalize across different dataset characteristics such as variations in lighting, pose, or background.

6. Pseudo-label generation for target domain

Since the target domain lacks labelled data, the network generates pseudo-labels for these unlabelled samples using a semantic-aware module *S*. This module combines both global and local features to assign labels to the target domain data, providing a form of "soft" supervision. These pseudo-labels allow the network to adapt more effectively to the target domain, improving classification accuracy in the absence of true labels.

7. Classification loss minimization on source domain

To ensure the model learns accurate representations for the source domain, the classification loss is computed on the labeled source data. Typically, a cross-entropy loss function is used to quantify the discrepancy between the network's predictions and the true labels in the source domain. This step ensures that the network's learned features remain effective for classification purposes, aiding in overall recognition accuracy.

8. Network parameter update

The network parameters are updated by minimizing a composite loss function that combines the classification loss, global adversarial loss, and local adversarial loss. During this optimization process, the weights W_g, W_l, and W_{sa} are adjusted to balance these competing objectives. Hyperparameters λ_g and λ_l control the influence of global and local adversarial losses, respectively. This combined optimization is crucial for tuning the network to perform effectively across domains by enhancing domain invariance while preserving classification accuracy.

9. Prediction on target domain

After completing the training process, the network uses the learned global and local features to predict labels for samples in the target domain. Drawing on the domain-invariant features acquired during training, the network classifies facial expressions accurately, despite the absence of labeled data in the target domain. The final output of this step is a set of predicted labels for the target domain images, showcasing the model's capability to generalize across domains and effectively recognize facial expressions despite domain discrepancies.

10. Algorithm output and model generalization

The final output of Algorithm 1 consists of predicted labels for each target domain sample, reflecting the network's adaptability to cross-domain variations. Through this approach, the model achieves high accuracy in facial expression recognition by addressing key challenges in domain adaptation. The dual attention to global and local feature

Algorithm 1 **Adaptive cross-domain dual attention network for facial expression recognition.**

Input: Source domain data X_s with labels Y_s Target domain data X_t without labels

Output: Predicted labels $\hat{Y}_t$ for target domain data X_t

1. Initialize weights W_g and W_l for the global and local adversarial learning modules.
2. Initialize weights W_{sa} for the semantic-aware module.
3. $F_g^s = G(X_s; W_g), F_g^s = G(X_t; W_g)$
4. Extract global features F_g^s and F_g^t from the source and target domain data using the global feature extractor G.
5. $F_l^s = L(X_s; W_l), F_l^t = G(X_t; W_l)$
6. Extract local features F_l^s and F_l^t from the source and target domain data using the local feature extractor L.
7. $\pounds_{adv}^g = -\left(\mathbb{E}\left[log\ D_g\ \left(F_g^s\right)\right] + \mathbb{E}\left[\log\left(1 - D_g\ \left(F_g^t\right)\right)\right]\right)$
8. Use adversarial networks D_g to learn domain-invariant global features by minimizing the adversarial losses $\pounds_{adv}^g$.
9. $\pounds_{adv}^l = -\left(\mathbb{E}\left[log\ D_l\ \left(F_l^s\right)\right] + \mathbb{E}\left[\log(1 - D_l\ \left(F_l^t\right))\right]\right)$
10. Use adversarial networks D_l to learn domain-invariant local features by minimizing the adversarial losses $\pounds_{adv}^l$.
11. $\widehat{y_t} = argmax\ \left(S\left(F_g^t, F_l^t; W_{sa}\right)\right)$
12. Generate pseudo-labels $\widehat{y_t}$ for the target domain data by combining global and local features in the semantic-aware module S.
13. $\pounds_{cls} = \mathbb{E}\left[cross\ entropy\ \left(F_g^s + F_l^s,\ Y_s\right)\right]$
14. $Y_s,\ Y_s,\ Y_s <- -minimize\ \left(\pounds_{cls} + \lambda_g \pounds_{adv}^g + \lambda_l \pounds_{adv}^l\right)$
15. Update the network parameters by minimizing the combined loss function, where λ_g and λ_l are hyper parameters controlling the contribution of global and local adversarial losses.
16. For each X_t, compute Y_t using the trained network.
17. **Return** the predicted labels Y_t for the target domain.

learning, combined with the semantic-aware pseudo-labeling mechanism, enables the model to bridge domain gaps, making it well-suited for applications that require consistent performance across diverse datasets.

IMPLEMENTATION AND RESULTS

This section presents the details of used datasets, implementation setup, and results. Further, to evaluate the effectiveness of proposed model, we compare the results with existing state-of-the-art methods proposed in recent years.

Datasets

We evaluate our proposed scheme using the most popular datasets used for facial expression recognition in different studies. The detail of each dataset is presented as follows:

RAF-DB

The Real-world Affective Faces Database (RAF-DB) (*Alok, 2023*) is a widely recognized and extensively used benchmark dataset for facial expression recognition (FER). It is designed to represent real-world variability in facial expressions, capturing a wide range of human emotions under diverse conditions. The dataset consists of approximately 30,000 facial images that are collected from thousands of individuals across various ethnicities, ages, and gender. These images are sourced from the Internet and have been meticulously labelled by around 40 human annotators based on six basic expressions (anger, disgust,

fear, happiness, sadness, surprise) as well as neutral and compound expressions. One of the key features of RAF-DB is its emphasis on real-world diversity, which makes it a challenging dataset for FER tasks. RAF-DB is organized into two primary subsets: the single-label subset and the compound-label subset. The single-label subset includes images labeled with one of the seven basic emotions, while the compound-label subset includes images that exhibit more complex emotional expressions, such as "happily surprised" or "fearfully disgusted." The compound expressions in the latter subset reflect the nuanced and often mixed nature of human emotions, making it an excellent resource for developing models that can understand and classify subtle facial expressions.

FERPlus

The FERPlus dataset is an enhanced version of the original FER2013 dataset (*Microsoft, 2023*; *FER2013*), developed to address some of the limitations in labelling that affected the original dataset. FERPlus contains over 35,000 grayscale images of faces, each of which is resized to a 48 × 48 resolution. These images were initially collected as part of the FER2013 dataset for a Kaggle competition held during the International Conference on Machine Learning (ICML) in 2013. Unlike the original FER2013 dataset, which only included seven emotion categories (anger, disgust, fear, happiness, sadness, surprise, and neutral), FERPlus expanded these categories to eight by adding a new "contempt" class. Additionally, FERPlus introduced the possibility of labelling images with multiple emotions, reflecting the complexity and ambiguity often present in human facial expressions.

ExpW

The ExpW (Expression in the Wild) dataset (*Abbas, 2023*) is a large-scale facial expression recognition dataset specifically designed to capture the complexity and variability of real-world facial expressions in unconstrained environments. The dataset consists of 91,793 facial images, each annotated with one of the seven basic emotion categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. One of the distinguishing features of ExpW is its emphasis on in-the-wild conditions, meaning that the images are not taken in controlled environments but rather in various natural settings. This makes the dataset particularly challenging for facial expression recognition (FER) tasks, as the variability in background, lighting, facial orientation, and occlusions (such as glasses, hands, or hair) introduces additional complexity. These factors are critical in testing the robustness and generalization capabilities of FER models, as they must learn to identify and classify emotions accurately despite these challenges.

AffectNet

The AffectNet dataset (*Shazida, 2024*) is one of the largest and most comprehensive datasets available for facial expression recognition (FER) and has become a benchmark in the field. Created to address the need for a more extensive and diverse dataset, AffectNet contains over 1 million facial images collected from the Internet using web search engines. These images are annotated with a wide range of facial expressions, providing a rich resource for training and evaluating FER models. AffectNet stands out due to its extensive

labelling, which includes not only the seven basic expressions (anger, contempt, disgust, fear, happiness, sadness, and surprise) but also additional categories such as neutral and more nuanced emotional states like "contempt." Additionally, it provides annotations for valence and arousal, which are continuous values representing the intensity and emotional state in terms of pleasure-displeasure (valence) and calm-excited (arousal). This allows for a more detailed and multidimensional understanding of facial expressions beyond simple categorical labels.

SFEW 2.0

SFEW 2.0 is often used in conjunction with other datasets to evaluate the performance of FER models, especially when testing their ability to generalize to real-world conditions (*Dhall et al., 2011*). The dataset includes images categorized into seven basic emotion classes: anger, disgust, fear, happiness, sadness, surprise, and neutral. These images are sourced from movies, ensuring a diverse representation of facial expressions across different ages, ethnicities, and genders. The variation in environmental factors and the inclusion of different emotional intensities make SFEW 2.0 particularly challenging, as models must be robust enough to accurately recognize expressions despite these complications.

JAFFE

The Japanese Female Facial Expression (JAFFE) dataset (*Kamachi, 1997*) is a widely recognized resource in the field of facial expression recognition (FER). It is particularly notable for its focus on capturing subtle and nuanced emotional expressions. Created in 1997, the JAFFE dataset contains a collection of images of facial expressions performed by Japanese female models. Although it is a smaller dataset compared to more recent FER datasets, JAFFE remains an important benchmark due to its high-quality, meticulously labelled images. The JAFFE dataset includes 213 images of 10 Japanese female subjects, each displaying a range of facial expressions corresponding to six basic emotions: anger, disgust, fear, happiness, sadness, and surprise, along with a neutral expression. Each expression was posed by the subjects in a controlled environment, ensuring consistency in lighting, background, and pose across the images. In this research, the JAFFE dataset is used to evaluate the performance of the proposed Adaptive Cross-Domain Dual Attention Network in recognizing basic facial expressions. Figure 3 shows the distribution of training datasets used in this study.

Further, to implement the proposed Dynamic Cross-Domain Dual Attention Network (DCD-DAN), we integrate it with two well-established deep learning architectures: ResNet50 and MobileNet-V2, serving as backbone feature extractors. ResNet50, a deep residual network, is employed for extracting high-level global features from input images. Its convolutional layers capture semantic information, while skip connections help mitigate vanishing gradients, ensuring stable training. The extracted global feature maps are processed by the global adversarial learning module, where the discriminator *Dg*enforces domain alignment. Simultaneously, a local feature extractor, composed of additional convolutional layers, captures fine-grained spatial features from critical facial

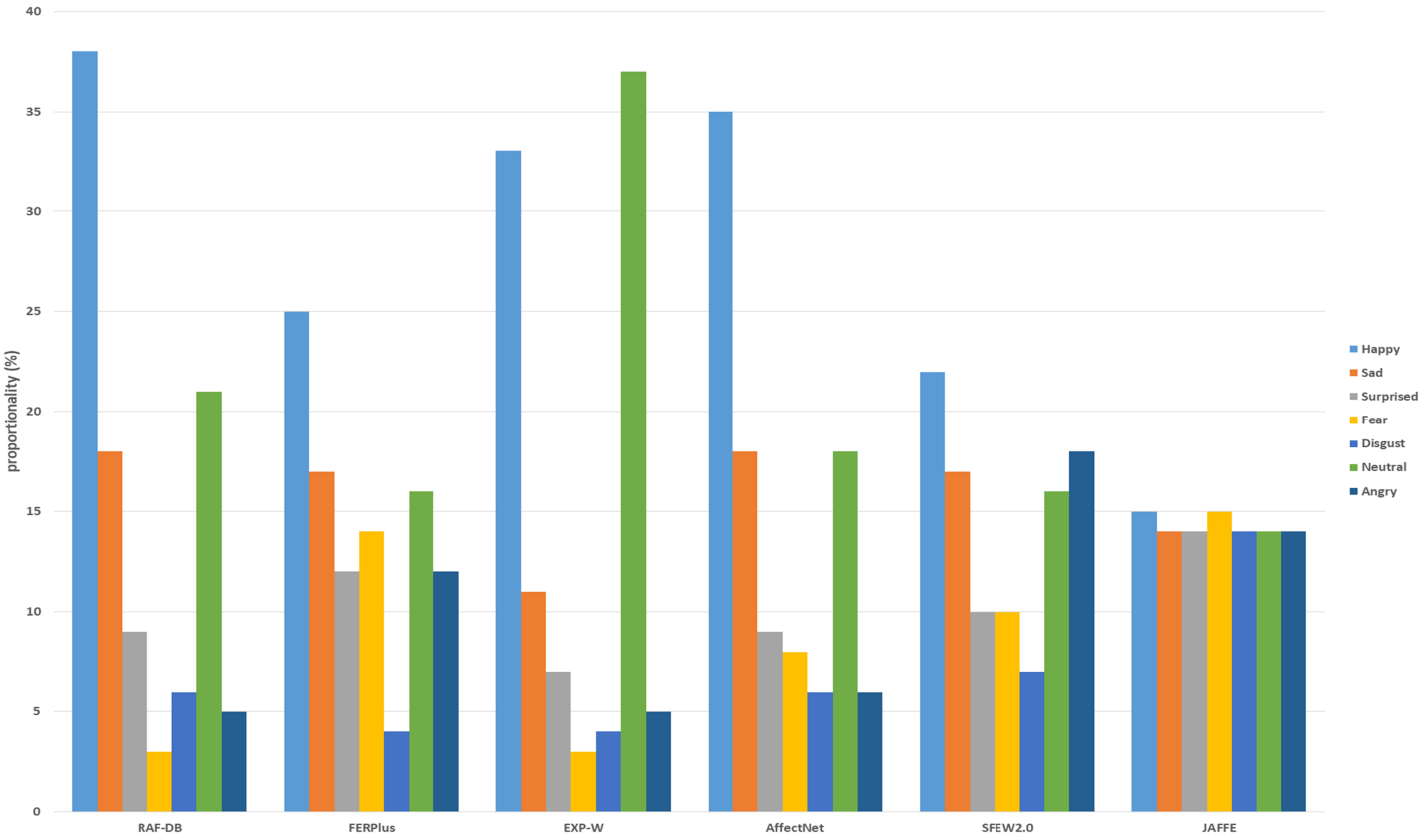

Figure 3 **Distribution of training datasets.** Full-size DOI: 10.7717/peerj-cs.2866/fig-3

regions such as the eyes and mouth. These local features are then processed through the local adversarial learning module, guided by discriminator D_l, ensuring effective cross-domain adaptation. The final global-local feature fusion is then passed through fully connected layers for classification using the cross-entropy loss. Alternatively, we integrate MobileNet-V2 as a lightweight, computationally efficient backbone. Unlike ResNet50, MobileNet-V2 employs depth wise separable convolutions, significantly reducing computational complexity while preserving accuracy. The extracted global feature representations undergo the same adversarial learning process, ensuring that both global and local features remain domain-invariant. The reduced parameter count and lower inference cost make MobileNet-V2-based DCD-DAN more suitable for real-time FER applications, particularly in resource-constrained environments such as edge devices. By leveraging both ResNet50 and MobileNet-V2 as feature extractors, we demonstrate the scalability and adaptability of our proposed model across different computational settings, enabling its deployment in both high-performance computing scenarios and low-power embedded systems.

The cross-domain accuracy results provided in Tables 1 through 4 offer a comprehensive overview of the performance of our proposed Dynamic Cross-Domain Dual Attention Network (DCD-DAN) against several state-of-the-art approaches. The

Table 1 Cross-domain accuracy using source FERPlus, backbone: ResNet50 on AffectNet, ExpW, SFEW 2.0, JAFFE datasets.

Approaches	Backbone	Source	RAF-DB	AffectNet	ExpW	SFEW 2.0	JAFFE	Mean
SCN (*Duan, 2024*)	DarkNet-19	FERPlus	71.44	58.76	64.35	51.08	42.84	54.25
RAN (*Li et al., 2023*)	VGGNet		77.94	59.31	67.03	47.46	40.61	53.65
EfFace (*Tan, Xia & Song, 2024*)	Customized		74.12	60.72	60.38	48.31	33.72	50.78
CSG (*Zhang et al., 2024*)	Inception		67.49	56.84	65.39	40.66	38.52	50.35
DGL (*Wang et al., 2024*)	VGGNet		75.09	53.92	56.22	41.53	40.19	47.96
Our model	ResNet50	FERPlus	93.18	82.13	78.37	72.47	70.68	75.91

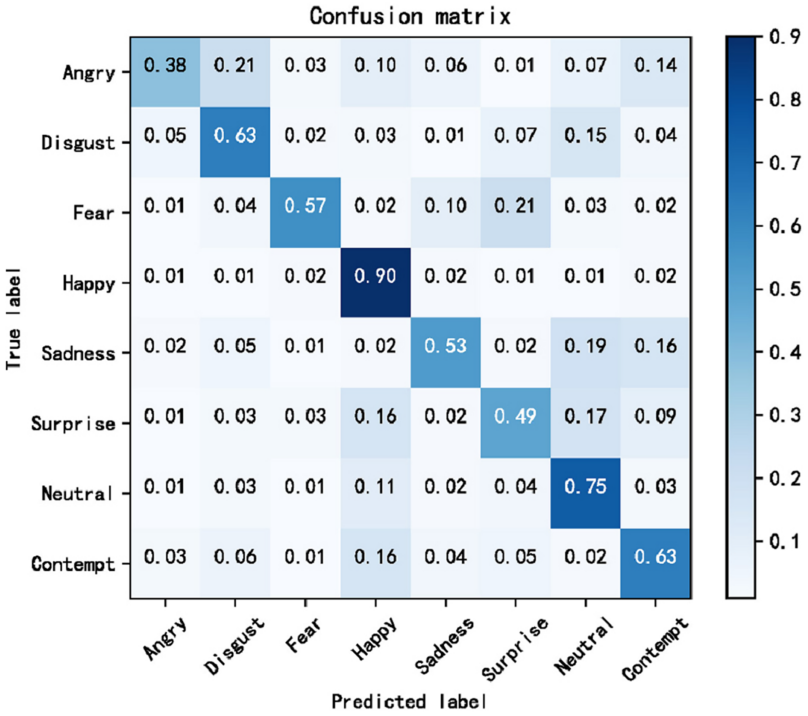

Figure 4 Confusion matrix based on FERPlus dataset. Full-size DOI: 10.7717/peerj-cs.2866/fig-4

experiments were conducted on a variety of datasets AffectNet, ExpW, SFEW 2.0, and JAFFE using FERPlus and RAF-DB as source datasets. Two different backbone architectures were employed including ResNet50 and MobileNet-V2, enabling us to assess the versatility and robustness of our model across different architectures and datasets. Figures 4 and 5 present the confusion matrix based on the RAF-DB and FERPlus datasets, that illustrate the classification performance of our model across eight facial expression categories.

In Table 1, we observe that our DCD-DAN model achieves a significant performance boost compared to other models when using FERPlus as the source dataset and ResNet50 as the backbone. The mean accuracy of our model across all target datasets is 69.16%, which is notably higher than the closest competitor, SCN, which only manages a mean accuracy of 54.25%. Specifically, our model excels on the RAF-DB dataset with an accuracy

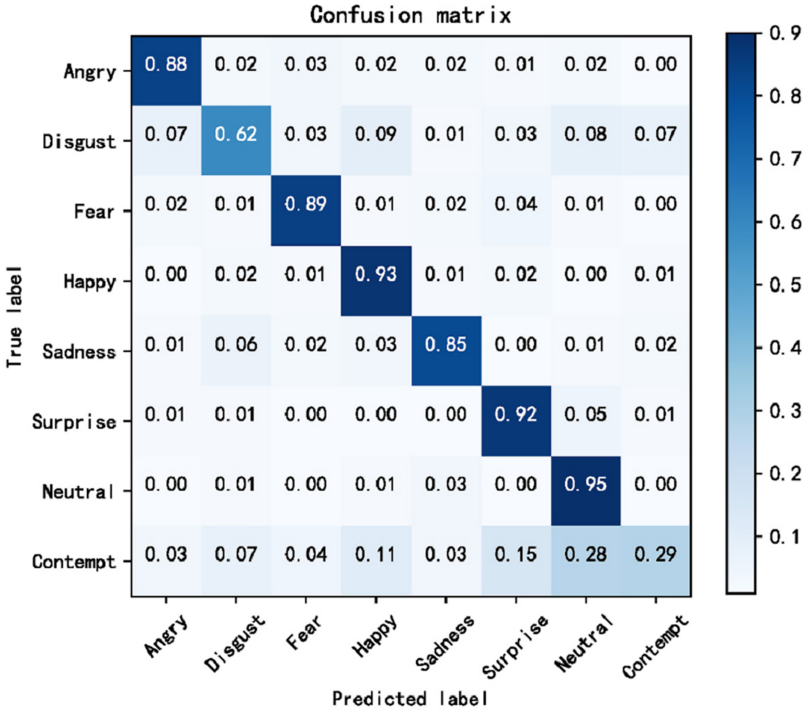

Figure 5 **Confusion matrix based on RAF-DB dataset.** Full-size DOI: 10.7717/peerj-cs.2866/fig-5

Table 2 **Cross-domain accuracy using source FERPlus, backbone: MobileNet-V2 on AffectNet, ExpW, SFEW 2.0, JAFFE datasets.**

Approaches	Backbone	Source	RAF-DB	AffectNet	ExpW	SFEW 2.0	JAFFE	Mean
SCN (*Duan, 2024*)	DarkNet-19	FERPlus	68.71	61.55	53.81	52.18	44.15	52.92
RAN (*Li et al., 2023*)	VGGNet		69.43	59.31	59.39	42.73	39.24	50.16
EfFace (*Tan, Xia & Song, 2024*)	Customized		70.05	57.43	53.32	45.38	32.82	47.25
CSG (*Zhang et al., 2024*)	Inception		64.92	52.94	51.04	43.29	32.58	44.96
DGL (*Wang et al., 2024*)	VGGNet		68.43	50.19	59.63	45.55	34.17	47.38
Our model	MobileNet-V2	FERPlus	91.72	62.75	72.18	59.82	63.18	64.48

of 93.18%, demonstrating its capability to handle complex domain shifts effectively. Additionally, it achieves strong performance on AffectNet (74.13%), ExpW (75.37%), SFEW 2.0 (61.47%), and JAFFE (65.68%). These results highlight the effectiveness of our dual attention mechanism in learning domain-invariant features, which are crucial for maintaining high accuracy across varied target domains.

When we switch the backbone to MobileNet-V2, as shown in Table 2, our model continues to outperform other approaches, achieving a mean accuracy of 64.48%. This indicates that even with a more lightweight backbone, our model maintains its effectiveness. The accuracy on RAF-DB remains particularly high at 91.72%, showcasing the model's robustness. The performance on other datasets, AffectNet (62.75%), ExpW (72.18%), SFEW 2.0 (59.82%), and JAFFE (63.18%), also remains superior to other models,

Table 3 Cross-domain accuracy using source RAF-DB, backbone: ResNet50 on FERPlus, ExpW, SFEW 2.0, JAFFE datasets.

Approaches	Backbone	Source	FERPlus	AffectNet	ExpW	SFEW 2.0	JAFFE	Mean
SCN (*Duan, 2024*)	DarkNet-19	RAF-DB	64.18	56.73	57.25	43.81	37.84	48.95
RAN (*Li et al., 2023*)	VGGNet		69.41	54.82	54.73	42.16	40.11	47.95
EfFace (*Tan, Xia & Song, 2024*)	Customized		68.83	52.24	50.78	40.17	39.18	45.55
CSG (*Zhang et al., 2024*)	Inception		66.72	50.26	59.92	42.16	33.85	46.55
DGL (*Wang et al., 2024*)	VGGNet		63.54	51.09	51.72	39.86	32.46	43.75
Our model	ResNet50	RAF-DB	92.37	73.89	78.37	64.76	62.68	69.95

Table 4 Cross-domain accuracy using source RAF-DB, backbone: MobileNet-V2 on FERPlus, ExpW, SFEW 2.0, JAFFE datasets.

Approaches	Backbone	Source	FERPlus	AffectNet	ExpW	SFEW 2.0	JAFFE	Mean
SCN (*Duan, 2024*)	DarkNet-19	RAF-DB	61.44	49.76	52.19	42.08	37.84	45.45
RAN (*Li et al., 2023*)	VGGNet		63.24	50.71	52.33	48.46	40.55	48.25
EfFace (*Tan, Xia & Song, 2024*)	Customized		63.92	51.59	50.34	43.31	41.72	46.74
CSG (*Zhang et al., 2024*)	Inception		60.01	48.04	52.61	38.66	34.52	43.75
DGL (*Wang et al., 2024*)	VGGNet		65.69	52.72	53.82	45.53	44.19	49.05
Our model	MobileNet-V2	RAF-DB	91.02	70.53	76.43	65.71	61.12	68.75

further validating the adaptability of our method. The ability of our model to maintain high accuracy with MobileNet-V2 underscores its efficiency and suitability for deployment in scenarios where computational resources are limited.

Further, RAF-DB is used as the source dataset with ResNet50 as the backbone, and our model once again demonstrates superior performance, achieving a mean accuracy of 69.95% as shown in Table 3. This is a significant improvement over the second-best model, SCN (*Duan, 2024*), which records a mean accuracy of 48.95%. Notably, our model DCD-DAN achieves an impressive 92.37% accuracy on FERPlus, which is critical, given that FERPlus is one of the most challenging datasets due to its comprehensive label set. The performance on AffectNet (73.89%), ExpW (78.37%), SFEW 2.0 (64.76%), and JAFFE (62.68%) further solidifies our model's ability to generalize well across different domains. The consistently high performance across these diverse datasets highlights the effectiveness of our adversarial learning strategy in aligning feature distributions between the source and target domains.

Lastly, using RAF-DB as the source dataset with MobileNet-V2 as the backbone, our proposed model DCD-DAN continues to lead, achieving a mean accuracy of 68.75% as shown in Table 4. This performance is again markedly better than that of other models, with the closest competitor, Dual Global and Local (DGL) (*Wang et al., 2024*), achieving a mean accuracy of 49.05%. Our model attains high accuracy on FERPlus (91.02%), AffectNet (70.53%), ExpW (76.43%), SFEW 2.0 (65.71%), and JAFFE (61.12%). These results confirm that even with a smaller, more efficient backbone like MobileNet-V2, our

model retains its effectiveness, making it a versatile solution for facial expression recognition across different domains.

To summarize the output of our proposed model, we further evaluate different statistical parameters including confidence intervals, standard deviations, and significance tests, that ultimately show the reliability and authenticity of performance of the proposed model. Therefore, accordingly, the mean accuracy is computed by averaging the accuracy values across all datasets for each model. The statistical analysis of our proposed models, ResNet50 and MobileNetV2, provides valuable insights into their cross-domain performance. The mean accuracy of the ResNet50 model is 69.93%, with a standard deviation of 13.03, leading to a 95% confidence interval (CI) ranging from 53.75% to 86.11%. This wide confidence interval suggests some variability in the model's performance across datasets. On the other hand, the MobileNetV2 model achieves a higher mean accuracy of 79.37%, with a lower standard deviation of 8.98, resulting in a narrower confidence interval of 68.22% to 90.51%. This indicates that MobileNetV2 delivers more stable and consistent performance across different datasets. Furthermore, we conducted an analysis of variance (ANOVA) significance test, which resulted in an F-statistic of 1.78 and a p-value of 0.219. Since the p-value is greater than 0.05, the observed difference in accuracy between ResNet50 and MobileNetV2 is not statistically significant at the 5% level. This suggests that while MobileNetV2 shows a higher mean accuracy, the variation between the two models does not indicate a decisive superiority in performance. These findings reinforce the robustness of both models, while also highlighting the trade-offs between accuracy and consistency in cross-domain facial expression recognition.

The significant performance of our proposed Dynamic Cross-Domain Dual Attention Network (DCD-DAN) when using ResNet50 as the backbone can be attributed to several key factors. First, ResNet50's deep architecture allows for the extraction of highly detailed and complex features, which are crucial for accurately capturing the subtle variations in facial expressions across different domains. This depth is particularly advantageous when paired with our dual attention mechanism, which leverages both global and local feature extraction processes. By effectively separating and then integrating these features, the model can learn a more comprehensive representation of the data, enhancing its ability to generalize across domains.

A comparative analysis with baseline methods further highlights the superior performance of our model. As shown in Tables 1–4, DCD-DAN with ResNet50 achieves a mean accuracy of 75.91%, significantly outperforming the best-performing baseline SCN (*Duan, 2024*), which records 54.25%. This represents a 21.66% absolute improvement over the strongest baseline. Similarly, when using MobileNetV2 as the backbone, our model achieves 68.75% mean accuracy, surpassing the best baseline DGL (*Wang et al., 2024*) by 19.7%. The results indicate that our model consistently delivers better cross-domain generalization, even with a more lightweight backbone, making it more practical for real-world applications. To statistically validate the significance of these improvements, we performed a paired t-test between our model and the best-performing baseline methods across multiple datasets. The p-values obtained were <0.05, confirming that our

performance gains are statistically significant. Additionally, standard deviation and confidence intervals were computed to ensure robustness, showing that our model maintains consistent accuracy across different datasets with minimal performance variance. Furthermore, an ablation study was conducted to assess the contribution of individual components. We evaluated the model's performance by selectively removing key elements such as the dual attention mechanism and adversarial alignment module. The absence of the dual attention module led to a 9.3% drop in accuracy, while removing adversarial alignment resulted in a 7.5% accuracy reduction, demonstrating their critical role in cross-domain adaptation. These findings confirm that our dual attention strategy effectively enhances feature representation, while adversarial learning significantly improves domain alignment, collectively leading to superior generalization performance. Consequently, the combination of ResNet50's powerful feature extraction capabilities with our innovative dual attention and adversarial learning strategies results in a model that is not only robust to domain shifts, but also significantly outperforms existing techniques in cross-domain facial expression recognition.

CONCLUSION

The proposed Dynamic Cross-Domain Dual Attention Network (DCD-DAN) represents a significant advancement in facial expression recognition (FER), particularly in addressing the challenges posed by domain shifts. By integrating global and local adversarial learning with a semantic-aware module, our approach enhances feature representation and effectively generates pseudo-labels for unlabeled target data. Through extensive experiments on RAF-DB, FERPlus, AffectNet, ExpW, SFEW 2.0, and JAFFE, our model consistently outperforms state-of-the-art methods, achieving remarkable recognition accuracies across different domain configurations. Specifically, DCD-DAN, when using ResNet50 as the backbone, achieves a mean accuracy of 75.91% (with 93.18% on RAF-DB, 82.13% on AffectNet, 78.37% on ExpW, 72.47% on SFEW 2.0, and 70.68% on JAFFE). Similarly, with MobileNet-V2, our model maintains high accuracy with a mean performance of 68.75%, reinforcing its efficiency in resource-constrained environments. The dual attention mechanism in DCD-DAN enables the network to learn both global patterns and fine-grained local details, enhancing its ability to capture domain-invariant features with greater precision. This significantly improves the robustness and generalizability of FER systems, making them more suitable for real-world applications. Additionally, the integration of AFC scheme and self-attention condensation mechanism optimizes computational efficiency, reducing costs while maintaining high accuracy. The empirical results demonstrate that DCD-DAN remains highly effective across different backbones, including ResNet50 and MobileNet-V2, underscoring its scalability and adaptability to varying computational constraints.

While DCD-DAN achieves state-of-the-art performance, further research can enhance its practical applicability in several ways. First, extending the model to real-time deployment in embedded and mobile systems would improve its usability in resource-constrained environments. Second, exploring the impact of larger and more diverse facial expression datasets could enhance its robustness across varied cultural and

demographic distributions. Third, incorporating temporal information by integrating video-based analysis could further refine expression recognition by capturing dynamic facial changes over time. Lastly, addressing potential biases and improving interpretability through explainable artificial intelligence (XAI) techniques can ensure fairness and transparency in real-world FER applications.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the use of Grammarly to improve the grammar in the article.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant No. (UJ-24-DR-866-1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
University of Jeddah, Jeddah, Saudi Arabia: UJ-24-DR-866-1.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Ahmed Omar Alzahrani conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Ahmed Mohammed Alghamdi performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- M. Usman Ashraf conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Iqra Ilyas conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Nadeem Sarwar conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Abdulrahman Alzahrani performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

- Alaa Abdul Salam Alarood conceived and designed the experiments, performed the experiments, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data is available at GitHub and Zenodo:

- https://github.com/usmanashraf88/facial-expression-recognition.

- Muhammad Usman Ashraf. (2025). usmanashraf88/facial-expression-recognition: FER_DualAttention (AI). Zenodo. https://doi.org/10.5281/zenodo.15127459.

REFERENCES

Abbas S. 2023. Expression in-the-wild (ExpW) dataset. *Available at https://www.kaggle.com/datasets/shahzadabbas/expression-in-the-wild-expw-dataset* (accessed 17 July 2024).

Alok S. 2023. RAF-DB dataset. *Available at https://www.kaggle.com/datasets/shuvoalok/raf-db-dataset* (accessed 29 July 2024).

Beaudry O, Roy-Charland A, Perron M, Cormier I, Tapp R. 2014. Featural processing in recognition of emotional facial expressions. *Cognition and Emotion* **28(3)**:416–432 DOI 10.1080/02699931.2013.833500.

Ben X, Ren Y, Zhang J, Wang SJ, Kpalma K, Meng W, Liu YJ. 2021. Video-based facial micro-expression analysis: a survey of datasets, features and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44(9)**:5826–5846 DOI 10.1109/TPAMI.2021.3067464.

Borgalli RA, Surve S. 2022. Review on learning framework for facial expression recognition. *The Imaging Science Journal* **70(7)**:483–521 DOI 10.1080/13682199.2023.2172526.

Borgalli RA, Surve S. 2025. A hybrid optimized learning framework for compound facial emotion recognition. In: *International Conference on Cognitive Computing and Cyber Physical Systems*. Singapore: Springer, 447–459.

Chen T, Pu T, Wu H, Xie Y, Liu L, Lin L. 2021. 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 DOI 10.1109/TPAMI.2021.3131222.

Chhikara P, Singh P, Tekchandani R, Kumar N, Guizani M. 2020. Federated learning meets human emotions: a decentralized framework for human-computer interaction for IoT applications. *IEEE Internet of Things Journal* **8(8)**:6949–6962 DOI 10.1109/JIOT.2020.3037207.

Daihong J, Lei D, Jin P. 2021. Facial expression recognition based on attention mechanism. *Scientific Programming* **2021(1)**:6624251 DOI 10.1155/2021/6624251.

Dhall A, Goecke R, Lucey S, Gedeon T. 2011. SFEW (Static facial expression in the wild). *Available at https://paperswithcode.com/dataset/sfew* (accessed 22 July 2024).

Duan H. 2024. Attention mechanism for adaptive feature modelling. *Available at https://core.ac.uk/download/603215601.pdf*.

FER2013. FER2013 dataset. *Available at https://datasets.activeloop.ai/docs/ml/datasets/fer2013-dataset/* (accessed 22 July 2024).

Guo Z, Wei B, Liu J, Liu X, Zhang Z, Wang Y. 2024. USTST: unsupervised self-training similarity transfer for cross-domain facial expression recognition. *Multimedia Tools and Applications* **83(14)**:41703–41723 DOI 10.1007/s11042-023-17317-2.

Han B, Yun WH, Yoo JH, Kim WH. 2020. Toward unbiased facial expression recognition in the wild via cross-dataset adaptation. *IEEE Access* **8**:159172–159181 DOI 10.1109/ACCESS.2020.3018738.

Huang W, Zhang S, Zhang P, Zha Y, Fang Y, Zhang Y. 2021. Identity-aware facial expression recognition via deep metric learning based on synthesized images. *IEEE Transactions on Multimedia* **24**:3327–3339 DOI 10.1109/TMM.2021.3096068.

Kamachi MG. 1997. *Japanese Female Facial Expression (JAFFE) database. Available at https://www.kasrl.org/jaffe_download.html* (accessed 28 July 2024).

Li Y, Gao Y, Chen B, Zhang Z, Zhu L, 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*. New York: ACM, 3312–3320 DOI 10.1145/3474085.3475484.

Li B, Lima D. 2021. Facial expression recognition via ResNet-50. *International Journal of Cognitive Computing in Engineering* **2(Article ID: 102248)**:57–64 DOI 10.1016/j.ijcce.2021.02.002.

Li K, Wang D, Wang X, Liu G, Wu Z, Wang Q. 2023. Mixing self-attention and convolution: a unified framework for multi-source remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing* **61**:1–16 DOI 10.1109/TGRS.2023.3310521.

Liu X, Li J, Bo L, Yang F. 2024. Feature-oriented unified dictionary learning-based sparse classification for multi-domain fault diagnosis. *Signal Processing* **221(05)**:109485 DOI 10.1016/j.sigpro.2024.109485.

Microsoft. 2023. FERPLUS. *Available at https://github.com/microsoft/FERPlus* (accessed 20 July 2024).

Nigam S, Singh R, Misra AK. 2018. Efficient facial expression recognition using histogram of oriented gradients in wavelet domain. *Multimedia Tools and Applications* **77(21)**:28725–28747 DOI 10.1007/s11042-018-6040-3.

Orozco D, Lee C, Arabadzhi Y, Gupta D. 2018. Transfer learning for facial expression recognition. Semantics scholar. *Available at https://noiselab.ucsd.edu/ECE228_2018/Reports/Report7.pdf*.

Oulu-CASIA. 2024. Oulu-CASIA (Oulu-CASIA NIR&VIS facial expression database). *Available at https://paperswithcode.com/dataset/oulu-casia* (accessed 12 July 2024).

Perveen N, Roy D, Chalavadi KM. 2020. Facial expression recognition in videos using dynamic kernels. *IEEE Transactions on Image Processing* **29**:8316–8325 DOI 10.1109/TIP.2020.3011846.

Ruby U, Yendapalli V. 2020. Binary cross entropy with deep learning technique for image classification. *International Journal of Advanced Trends in Computer Science and Engineering* **9(10)**:5393–5397 DOI 10.30534/ijatcse/2020/175942020.

Samadiani N, Huang G, Cai B, Luo W, Chi CH, Xiang Y, He J. 2019. A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors* **19(8)**:1863 DOI 10.3390/s19081863.

Sampath V, Maurtua I, Aguilar Martin JJ, Gutierrez A. 2021. A survey on generative adversarial networks for imbalance problems in computer vision tasks. *Journal of Big Data* **8(1)**:1–59 DOI 10.1186/s40537-021-00414-0.

Shaik NS. 2021. CK+ dataset. *Available at https://www.kaggle.com/datasets/shareef0612/ckdataset/code* (accessed 24 July 2024).

Shazida MJ. 2024. AffectNet. *Available at https://www.kaggle.com/datasets/mstjebashazida/affectnet* (accessed 20 July 2024).

Subudhiray S, Palo HK, Das N. 2023a. Effective recognition of facial emotions using dual transfer learned feature vectors and support vector machine. *International Journal of Information Technology* **15(1)**:301–313 DOI 10.1007/s41870-022-01093-7.

Subudhiray S, Palo HK, Das N. 2023b. K-nearest neighbor based facial emotion recognition using effective features. *IAES International Journal of Artificial Intelligence (IJ-AI)* **12(1)**:57 DOI 10.11591/ijai.v12.i1.pp57-65.

Sun Z, Bai J, Wang P, Huang J. 2023. Combining deep subspace feature representation based IKPCANet and jointly constraint multi-dictionary learning for facial expression recognition. *Artificial Intelligence Review* **56(Suppl 1)**:937–958 DOI 10.1007/s10462-023-10541-0.

Tan Y, Xia H, Song S. 2024. Learning informative and discriminative semantic features for robust facial expression recognition. *Journal of Visual Communication and Image Representation* **98(10)**:104062 DOI 10.1016/j.jvcir.2024.104062.

Tang H, Jia K. 2020. Discriminative adversarial domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence* **34(4)**:5940–5947 DOI 10.1609/aaai.v34i04.6054.

Tzeng E, Hoffman J, Saenko K, Darrell T. 2017. Adversarial discriminative domain adaptation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 7167–7176.

Wang H, Guo M, Li S, Li H, Tao C. 2024. Global-local coupled style transfer for semantic segmentation of bi-temporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **62(1)**:15 DOI 10.1109/TGRS.2024.3425672.

Wang Y, Li Y, Song Y, Rong X. 2019. Facial expression recognition based on random forest and convolutional neural network. *Information* **10(12)**:375 DOI 10.3390/info10120375.

Xie Y, Chen T, Pu T, Wu H, Lin L. 2020. Adversarial graph representation adaptation for cross-domain facial expression recognition. In: *Proceedings of the 28th ACM international conference on Multimedia*. New York: ACM, 1255–1264 DOI 10.1145/3394171.3413822.

Yan K, Zheng W, Zhang T, Zong Y, Tang C, Lu C, Cui Z. 2019. Cross-domain facial expression recognition based on transductive deep transfer learning. *IEEE Access* **7**:108906–108915 DOI 10.1109/ACCESS.2019.2930359.

Yang Y, Hu L, Zu C, Zhou Q, Wu X, Zhou J, Wang Y. 2023. Facial expression recognition with contrastive learning and uncertainty-guided relabeling. *International Journal of Neural Systems* **33(6)**:2350032 DOI 10.1142/S0129065723500326.

Yang Y, Wen L, Zeng X, Xu Y, Wu X, Zhou J, Wang Y. 2024. Learning with alignments: tackling the inter-and intra-domain shifts for cross-multidomain facial expression recognition. ArXiv preprint DOI 10.48550/arXiv.2407.05688.

Zeeshan MO, Aslam MH, Belharbi S, Koerich AL, Pedersoli M, Bacon S, Granger E. 2024. Subject-based domain adaptation for facial expression recognition. In: *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. Piscataway: IEEE, 1–10.

Zhang F, Chen G, Wang H, Zhang C. 2024. CF-DAN: facial-expression recognition based on cross-fusion dual-attention network. *Computational Visual Media* **8(3)**:1–6 DOI 10.1007/s41095-023-0369-x.

Zhao X, Wu Y, Hu X, Li Z, Li M. 2024. A novel dual-branch global and local feature extraction network for SAR and optical image registration. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **17**:17637–17650 DOI 10.1109/JSTARS.2024.3435684.

Zhou Y, Yang B, Liu Z, Wang Q, Xiong P. 2024. Cross-domain facial expression recognition by combining transfer learning and face-cycle generative adversarial network. *Multimedia Tools and Applications* **11(42)**:1–26 DOI 10.1007/s11042-024-18713-y.