Journal of King Saud University Computer and Information Sciences https://doi.org/10.1007/s44443-026-00705-7

Article in Press

A context-aware dropout-based occlusion-adaptive network for robust facial landmark and emotion detection

Received: 01 Jan 2026

Accepted: 25 Mar 2026

Published online: 16 April 2026

Cite this article as: Sadiq, M., Zhang, Y., Zhou, Y. *et al.* A context-aware dropout-based occlusion-adaptive network for robust facial landmark and emotion detection. *J. King Saud Univ. Comput. Inf. Sci.* (2026). https://doi.org/10.1007/s44443-026-00705-7

Muhammad Sadiq, Yunsheng Zhang, Yu Zhou, Mohammad Mahmud, Muhammad Azhar, Muhammad Durad & Junwei Liang

A context-aware dropout-based occlusion-adaptive network for robust facial landmark and emotion detection

Abstract

We present CAD-Net, a Context-Aware Dropout-based Occlusion-Adaptive Network for robust facial landmark detection (FLD) and facial expression recognition (FER) under challenging real-world conditions. Clinical and in-the-wild settings frequently involve strong occlusions (e.g., masks, medical devices), illumination changes, and large pose variations, where conventional convolutional neural networks (CNNs) often experience significant performance degradation. Rather than treating occlusion primarily as an isolated visibility-estimation problem, CAD-Net is designed to preserve structural facial consistency under partial visibility by jointly modeling long-range geometry, feature reliability, and compact landmark regression. CAD-Net comprises three complementary components: (i) a deep geometry-aware block that leverages criss-cross attention to preserve global facial structure and propagate information between spatially distant but correlated facial regions, (ii) an attentive dropout block that combines channel-wise attention with learnable dropout masks to down-weight unreliable or occluded regions, and (iii) a low-rank learning block that regularizes the regression head to obtain compact and stable landmark predictions. These modules are trained jointly within a unified framework that can be instantiated for both image-based and video-based facial analysis and naturally extended to a multi-task setting that couples FLD and FER. By integrating geometry-aware reasoning, selective occlusion suppression, and low-rank regularization in a single end-to-end architecture, CAD-Net improves robustness in a lightweight and practically deployable setting, while maintaining moderate computational overhead. Extensive experiments on standard FLD benchmarks (300W, COFW, AFLW, 300VW, Menpo) and FER datasets demonstrate that CAD-Net achieves competitive or superior performance compared with recent occlusion-aware methods, particularly under severe occlusions and pose variations. We further strengthen the empirical evaluation by reporting unified-protocol comparisons where feasible, as well as additional analyses of efficiency, stability across multiple runs, and cross-dataset transfer. The proposed design improves robustness without incurring prohibitive computational overhead, making CAD-Net suitable for time-sensitive biomedical and health informatics applications such as telemedicine, mental health monitoring, and elderly care.

Keywords: facial landmark detection, facial expression recognition, occlusion handling, context-aware dropout, low-rank learning, telemedicine

1 Introduction

Facial landmarks are a fundamental representation for higher-level facial analysis, consisting of a set of anatomically meaningful keypoints distributed over the face. These landmarks typically correspond to the corners of the eyes, the tip and bridge of the nose, the contours of the mouth and lips, and the boundaries of the chin and jawline. Accurate localization of such landmarks underpins a wide range of downstream tasks, including face alignment, recognition, expression analysis, and 3D facial modeling. As these applications continue to expand in areas such as human–computer interaction, security, and healthcare, improving the precision and robustness of facial landmark detection (FLD) remains a central research challenge.

Emotion and expression analysis based on facial landmarks is particularly attractive for healthcare and biomedical applications. In mental health monitoring, landmark-based models can support the assessment of depression, anxiety, and other affective states, while in telemedicine they can help clinicians interpret subtle non-verbal cues during remote consultations Onishi (2021); Fu et al. (2023). Integrated into wearable devices or telehealth platforms Liedtke et al. (2018), such systems enable the continuous tracking of behavioural and emotional patterns and facilitate personalised, data-driven interventions Yildirim-Celik et al. (2022). However, real-world clinical environments frequently involve severe occlusions caused by masks, medical devices, hair, or hands, as well as strong illumination changes and pose variations, all of which make robust landmark localization and expression analysis considerably more difficult.

Despite substantial progress in deep learning-based FLD, performance degradation under heavy occlusion and extreme pose remains a persistent limitation. For example, on challenging subsets such as 300W-Challenging, existing occlusion-aware methods report NRMSE values above 5.8×10^{-2}, whereas performance under relatively controlled conditions is significantly lower. This gap reflects structural instability rather than merely increased point-wise error. Under severe occlusion, predictions often violate global facial symmetry and relative geometric consistency, which in turn propagates errors to downstream tasks such as facial expression recognition (FER).

Classical FLD approaches can be broadly grouped into regression based, template-based, and deep learning-based methods. Regression-based techniques learn a direct mapping from image features to landmark coordinates without explicit shape models, often achieving a fast and simultaneous prediction of all landmarks while implicitly maintaining geometric consistency Wu and Ji (2019). Template-based approaches rely on statistical models, such as principal component analysis (PCA), to capture facial shape variations from annotated datasets Zhu and Ramanan (2012). Although effective under relatively controlled conditions, these methods typically exhibit limited robustness to occlusions and large pose changes.

The advent of deep learning, in particular convolutional neural networks (CNNs), has significantly advanced FLD by alleviating many limitations of handcrafted features and shallow models. More recent work has introduced occlusion-adaptive architectures that explicitly estimate landmark visibility alongside landmark locations and exploit occlusion-aware feature representations to improve robustness in the presence of partial occlusions. While these methods have shown promising results, FLD in unconstrained and clinical environments remains highly sensitive to occlusions, extreme viewpoints, uneven illumination, and missing or corrupted facial regions. Under such conditions, existing deep models often struggle to distinguish informative features from occluded ones, tend to overemphasize highly salient yet incomplete regions, and may fail to preserve the global geometric structure of the face.

A central limitation of many existing approaches is insufficient modeling of long-range facial geometry when parts of the face are invisible. Convolutional layers mainly capture local patterns, and when these local cues are corrupted by occlusion, the model lacks an explicit mechanism to infer missing structural relationships from distant but correlated facial regions. As a result, structural consistency across the face is not explicitly enforced, leading to unstable predictions and reduced cross-dataset generalization.

Another practical limitation is that many recent methods improve performance by relying on increasingly heavy backbones or large-capacity architectures, which makes it more difficult to determine whether robustness comes from the proposed occlusion-handling strategy itself or simply from increased model scale. This also raises concerns about deployment in time-sensitive and resource-constrained environments.

In this work, we explicitly formulate *occlusion-induced geometric inconsistency* as the core failure mode to be addressed. Rather than focusing solely on local landmark accuracy or visibility estimation, we aim to preserve global structural coherence under partial visibility.

In this work, we propose CAD-Net, a Context-Aware Dropout-based Occlusion-Adaptive Network designed to address these limitations in a unified framework for FLD and facial expression recognition (FER). CAD-Net builds on the strengths of occlusion-adaptive deep networks while introducing three tightly coupled components: (i) a deep geometry-aware block that leverages criss-cross attention to capture long-range structural dependencies across the face, (ii) an attentive dropout block that combines channel-wise attention with learnable dropout masks to down-weight unreliable or occluded regions, and (iii) a low-rank learning block that regularizes the regression head to produce compact and stable landmark predictions.

To ensure that the contribution of the proposed modules can be evaluated in a controlled and deployment-conscious setting, we intentionally adopt a lightweight backbone. This choice allows the empirical gains to be attributed more directly to the geometry-aware, attentive dropout, and low-rank components rather than to backbone scaling alone.

Criss-cross attention aggregates contextual information along horizontal and vertical directions passing through each spatial location, enabling each pixel to access structurally related facial regions such as symmetric eye or mouth components. Unlike full pairwise attention, this directional context modeling provides an efficient approximation of global dependency learning while maintaining moderate computational cost. By allowing visible facial regions to contribute geometric cues for occluded parts, the model explicitly supports structural inference under partial visibility.

Importantly, CAD-Net is not a simple aggregation of independent modules. The geometry-aware block, attentive dropout mechanism, and low-rank regression head are jointly optimized so that geometry propagation is guided by reliability-aware features, and regression capacity is constrained to prevent overfitting to occlusion-specific artifacts. This inter-dependent design differentiates CAD-Net from prior approaches that incorporate attention, occlusion modeling, or regularization as isolated enhancements.

In addition to improving landmark localization, this coupled design also provides a stronger basis for multi-task facial expression recognition. Stable landmark geometry offers a structured representation of facial

Figure 1 Examples from the COFW dataset Burgos-Artizzu et al. (2013) illustrating typical occlusions caused by hair, hands, accessories, and food. Such occlusions pose significant challenges for accurate facial landmark detection in unconstrained environments.

deformation, which is particularly useful when expression-related appearance cues are partially missing due to occlusion or pose variation.

Empirically, this design leads to consistent improvements across multiple challenging benchmarks. For example, on the 300W Full subset, CAD-Net reduces NRMSE to 2.90×10^{-2}, outperforming recent occlusion-aware baselines, while on the 300W-Challenging subset it achieves 5.21×10^{-2}, demonstrating improved robustness under severe occlusion. Similar gains are observed on COFW and AFLW under pose variation.

The revised manuscript further strengthens this empirical evidence by including unified-protocol baselines, repeated-run stability analysis, efficiency evaluation, and cross-dataset transfer experiments, so that the practical and methodological benefits of the proposed design can be assessed more transparently.

To improve clarity and focus, we summarize the contributions in a problem–method–benefit structure:

1. **Problem:** Occlusion and pose variation cause geometric inconsistency and unstable landmark topology. **Method:** We introduce a geometry-aware modeling block based on recurrent criss-cross attention to capture long-range structural dependencies. **Benefit:** Improved landmark stability and structural coherence under severe occlusion.
2. **Problem:** Occlusion-corrupted activations mislead feature learning. **Method:** We propose an attentive dropout mechanism that estimates feature reliability and suppresses unreliable regions while preserving salient structural cues. **Benefit:** Enhanced robustness and cross-dataset generalization.
3. **Problem:** High-capacity regression heads overfit to spurious occlusion patterns. **Method:** We constrain landmark regression to a compact low-rank subspace. **Benefit:** Stable predictions with moderate computational overhead suitable for practical deployment.

Beyond standalone FLD, CAD-Net supports a unified multi-task setting combining FLD and FER. By preserving stable geometric representations, the framework improves expression recognition under occlusion and pose variation without requiring explicit action unit supervision.

The remainder of the paper also reflects this revised focus on controlled evaluation and practical feasibility. In particular, the experimental section now distinguishes reported and unified-protocol comparisons, includes efficiency and stability analyses, and discusses limitations and deployment considerations more explicitly.

The remainder of this paper is organized as follows. Section 2 reviews related work on facial landmark detection, occlusion handling, and attention-based deep architectures. Section 3 presents the overall architecture of the proposed *CAD-Net* and details its core components. Section 4 describes the optimization strategy and the learning objectives used to train the network. Experimental setups and results for FLD are reported in Section 5, followed by an analysis of FER performance in Section 6. The ablation study is discussed in Section 7, and potential application domains and limitations are outlined in Section 8. Finally, Section 9 concludes the paper and highlights directions for future work.

Figure 2 Examples of occluded faces from the AffectNet dataset Mollahosseini et al. (2017), where masks, hairs, hands, and objects obscure key regions of the face. These cases demonstrate the difficulty of recognizing emotions under occluded and extreme pose conditions.

2 Related Work

Facial landmark detection (FLD) aims to accurately localize a predefined set of keypoints on facial images. In unconstrained environments, however, it faces significant challenges due to occlusions that are often unpredictable and complex, as illustrated in Figures 1 and 2. A number of strategies have been proposed to mitigate these effects, including supervised regression for iteratively updating landmark visibility probabilities Wu and Ji (2019), occlusion dictionaries that explicitly model typical occlusion patterns Xing et al. (2017), and adaptive regression combined with shape modeling to infer landmark visibility Liu et al. (2016). More recent advances Sadiq et al. (2022); Deng et al. (2020) have further improved the accuracy and robustness of facial point localization by leveraging stronger backbones and improved face detection. Occlusion-adaptive deep networks such as ODN introduce joint modeling of landmark positions and visibility, demonstrating that explicitly estimating occlusion can substantially enhance robustness under challenging conditions, but they still rely largely on generic CNN backbones and limited modeling of global facial geometry.

Although visibility-aware regression improves robustness, many existing approaches primarily refine local features and treat occlusion as a per-landmark confidence estimation problem. Consequently, occlusion handling is often implemented as a confidence re-weighting strategy rather than as an explicit structural reasoning process. The preservation of long-range geometric dependencies across distant facial regions remains insufficiently explored, especially under severe occlusions where entire facial components may be invisible.

Beyond regression-based approaches, recent FLD models increasingly adopt heatmap-based localization and transformer-style global modeling. Heatmap regression strategies improve spatial precision by learning dense probability maps, while transformer architectures attempt to model global dependencies through self-attention mechanisms. Although these methods enhance contextual modeling, they are often optimized primarily for point-wise localization accuracy and may not explicitly enforce structural consistency under partial visibility. Moreover, the computational overhead of full self-attention can limit practicality in time-sensitive or resource-constrained settings.

Recent progress has also shown that stronger backbones can substantially improve in-dataset performance. However, when evaluating an occlusion-handling strategy itself, heavier architectures may obscure whether the gain originates from improved geometric reasoning or simply from increased model capacity. This distinction is important for both fair methodological comparison and practical deployment.

The *Facial Action Coding System* (FACS) Ekman and Friesen (1978); Hager et al. (2002) provides a detailed framework for systematically describing human facial movements and expressions. As illustrated in Table 1, it decomposes facial activity into a set of *Action Units* (AUs), which are anatomically grounded and comprise roughly 30 distinct units, 12 associated with the upper face and 18 with the lower face. FACS captures all observable movements of facial muscles, including those not directly linked to emotions or specific psychological states. While FACS has been widely adopted by behavioural scientists for the precise delineation of facial expressions, manual AU coding is labor-intensive and time-consuming, motivating research in automatic AU detection and FER.

From a computational perspective, FACS implicitly emphasizes that facial expressions arise from coordinated geometric interactions among multiple facial regions. However, many deep FER approaches primarily focus on appearance-based feature extraction and classification, with limited explicit incorporation of landmark topology or structural priors derived from facial geometry. As a result, the relationship between landmark configuration and emotion prediction is often weakly constrained, particularly under occlusion or pose variation.

This observation is particularly relevant for cross-dataset FER, where appearance statistics can vary substantially across collections, while expression-related geometric deformations remain more stable. A geometry-aware representation can therefore serve as a useful structural prior when transferring across datasets, identities, and capture conditions.

In recent years, increasing attention has been devoted to *automated* AU detection. Several methods analyse localized facial patches and model AU co-occurrence in a multi-label learning framework Zhao et al. (2015), while others employ localized classifiers targeting distinct facial regions Jaiswal et al. (2015). More recent studies integrate attention mechanisms into weakly supervised settings to highlight regions most relevant to each AU Shao et al. (2018b,a). AU-based modeling plays a crucial role in interpretable facial expression analysis Tian et al. (2001); however, approaches that are strictly tied to a predefined AU taxonomy may overlook the broader range of subtle muscle activations and geometric deformations through which human emotions are expressed.

Deep learning-based methods currently dominate automatic facial expression recognition (FER). Compared to traditional approaches relying on handcrafted features or rule-based systems, deep models—particularly convolutional neural networks (CNNs)—have demonstrated superior accuracy and robustness in emotion classification Kennedy and Balint (2016); Kowalski et al. (2017); Lopes et al. (2015). By learning hierarchical representations, CNNs capture both fine-grained local details (e.g., around the eyes and mouth) and global facial patterns, and when trained on large-scale datasets they exhibit strong generalization across variations in illumination, pose, and ethnicity.

Architectures inspired by the *Inception* model Szegedy et al. (2015) have further advanced FER by enabling multi-scale feature representation. These multi-branch designs allow simultaneous extraction of local and global features and have shown improved robustness to facial variations and partial occlusions Hasani and Mahoor (2017); Xia et al. (2017).

More recently, FER research has also moved toward stronger CNN hybrids and transformer-inspired models that improve recognition accuracy through richer context modeling. While these methods are effective for classification, they do not necessarily clarify how much of the improvement comes from explicit facial geometry modeling, nor do they always maintain moderate computational cost for deployment-oriented settings.

More broadly, recent advances in image processing and representation learning emphasize robust context modeling, structured feature aggregation, and efficient attention mechanisms Wang et al. (2025b); Liu et al. (2025); Wang et al. (2025a). These developments highlight the importance of balancing global dependency modeling with computational efficiency. CAD-Net aligns with this direction by adopting structured context propagation (criss-cross attention) rather than full pairwise attention, thereby achieving global geometric reasoning with moderate computational overhead.

Recent FER and FLD models Zhu et al. (2019); Gao et al. (2020); Browatzki and Wallraven (2020); Sadiq and Shi (2022); Wan et al. (2023); Xiang et al. (2025); Sadiq et al. (2024) leverage deeper architectures and increasingly sophisticated attention mechanisms to improve classification and localization accuracy. Many of these frameworks integrate advanced face detectors, heatmap regression strategies, or occlusion-aware components to suppress background clutter and non-facial artifacts. Our comparative analysis also considers contemporary FER methods proposed in Xie et al. (2020); Peng et al. (2022); Liu et al. (2023), which typically adopt ResNet-based backbones and demonstrate competitive performance on in-the-wild benchmarks.

Despite these advances, a substantial portion of existing work remains inspired by generic object classification architectures and optimizes landmark localization and emotion recognition primarily through feature refinement and classification loss design. The intrinsic geometric topology of the human face, including symmetry, proportional constraints, and spatial interdependencies among distant landmarks, is often underutilized.

Furthermore, many improvements are achieved by stacking additional modules such as attention layers, auxiliary branches, or regression refinements without explicitly enforcing structural coherence under partial visibility. While such modular enhancements can improve performance incrementally, they do not necessarily address the core challenge of reconstructing consistent facial geometry when critical regions are occluded.

A second limitation is that many reported improvements are demonstrated mainly under in-dataset evaluation. Under distribution shift, such as unseen occlusion patterns, demographic diversity, or capture changes, the absence of explicit geometry preservation can lead to unstable landmark topology and degraded downstream recognition.

Another limitation concerns evaluation protocols. A considerable number of studies report results under in-dataset training and testing settings. When exposed to distribution shifts involving unseen occlusion types,

Table 1 Facial Action Coding System (FACS) configurations for basic emotions and their corresponding facial landmarks, based on the EMFACS methodology Hager et al. (2002); Ekman and Friesen (1978).

Emotion	AUs	Description	Facial Landmarks
Happiness	6, 12	Cheek raiser; lip corner puller	1–2, 14–15; 48–49, 53–55, 59–60, 64
Anger	4, 5+7, 23	Brow lowerer; upper lid raiser and lid tightener; lip tightener	17–22; 37–39, 42–44; 48–49, 53–67
Fear	1, 2, 4, 5+7, 20, 26	Inner and outer brow raiser; brow lowerer; lid actions; lip stretcher; jaw drop	17–26; 37–39, 42–44; 48–49, 53–55, 59–60, 64; 55–66
Sadness	1, 4, 15	Inner brow raiser; brow lowerer; lip corner depressor	17–21; 22–26; 48–49, 53–55, 59–60, 64
Disgust	9, 15, 16	Nose wrinkler; lip corner depressor; lower lip depressor	27–35; 48–49, 53–58, 59–60, 64
Surprise	1, 2, 5, 26	Inner and outer brow raiser; upper lid raiser; jaw drop	17–26; 37–39, 42–44; 55–66

demographic diversity, or pose variations, performance often degrades, indicating limited structural generalization across domains. This observation motivates approaches that explicitly model geometry and reliability to enhance cross-dataset robustness.

To address these limitations, we propose a geometry-aware and occlusion-adaptive approach that moves beyond purely appearance-based or AU-only modeling. Our method emphasizes the spatial geometry and dynamic interrelations of facial landmarks, using structural topology as an explicit modeling prior to enhance both landmark detection and emotion recognition.

Unlike prior works that incorporate attention, dropout, or low-rank constraints independently, the proposed framework integrates global geometry modeling, selective occlusion suppression, and compact regression learning within a unified optimization strategy. The objective is not merely to aggregate existing modules, but to couple them so that geometry-aware context reasoning guides occlusion suppression and low-rank regression stabilizes the resulting structural representation. This coordinated design explicitly targets structural consistency under partial visibility while maintaining computational practicality.

This positioning is important because the goal of CAD-Net is not to compete through backbone scale alone, but to show that a lightweight and controlled architecture can still achieve strong robustness when the occlusion-handling mechanism is explicitly tied to structural geometry and feature reliability.

By capturing complex geometric deformations and expression patterns, the proposed framework provides a more nuanced and explainable representation of emotional cues. Specifically, we employ an advanced facial landmark detection architecture, *CAD-Net*, which integrates a deep geometry-aware block, an attentive dropout block, and a low-rank surrogate regularizer within the loss function to jointly improve landmark detection and emotion classification. This joint design enables the network to exploit landmark position dependencies for robust FER while maintaining strong occlusion robustness and cross-dataset generalization.

3 Detailed Structure of CAD-Net

To achieve discriminative feature extraction, robust occlusion handling, and efficient inference with moderate computational overhead, we propose a novel architecture termed *CAD-Net*. As illustrated in Figure 4, CAD-Net consists of three tightly integrated functional blocks:

1. Deep Geometry-Aware Block (DGB),
2. Attentive Dropout Block (ADB), and
3. Low-Rank Learning Block (LLB).

Although the individual ingredients are related to structured attention, attention-guided dropout, and low-rank factorization, the main contribution of CAD-Net lies in how these mechanisms are coupled to address a specific failure mode, namely occlusion-induced geometric inconsistency. The three blocks are designed to operate inter-dependently rather than independently: DGB propagates structured long-range geometry, ADB estimates feature reliability and suppresses occlusion-dominated activations before they distort global reasoning, and LLB constrains the final mapping so that residual noise does not translate into unstable landmark updates.

While the building blocks are related to existing ideas (structured attention, dropout, and low-rank factorization), CAD-Net differs in *how these mechanisms are coupled* to address occlusion-induced failure modes. Specifically, (i) DGB propagates *structured long-range geometric context* to infer missing relations, (ii) ADB estimates *feature reliability* and suppresses occlusion-dominated responses to avoid misleading context propagation, and (iii) LLB constrains the final mapping to a *compact geometric subspace* so that residual occlusion

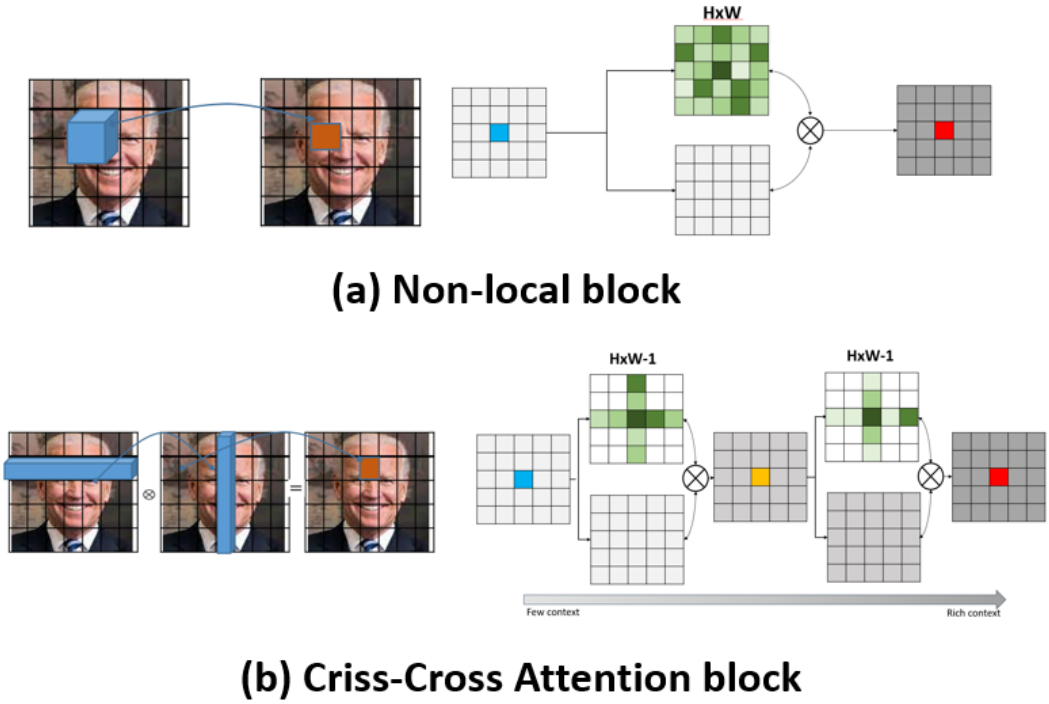

Figure 3 Structural comparison between a generic non-local block and the criss-cross attention (CCA) block adopted in our deep geometry-aware module. For readability.

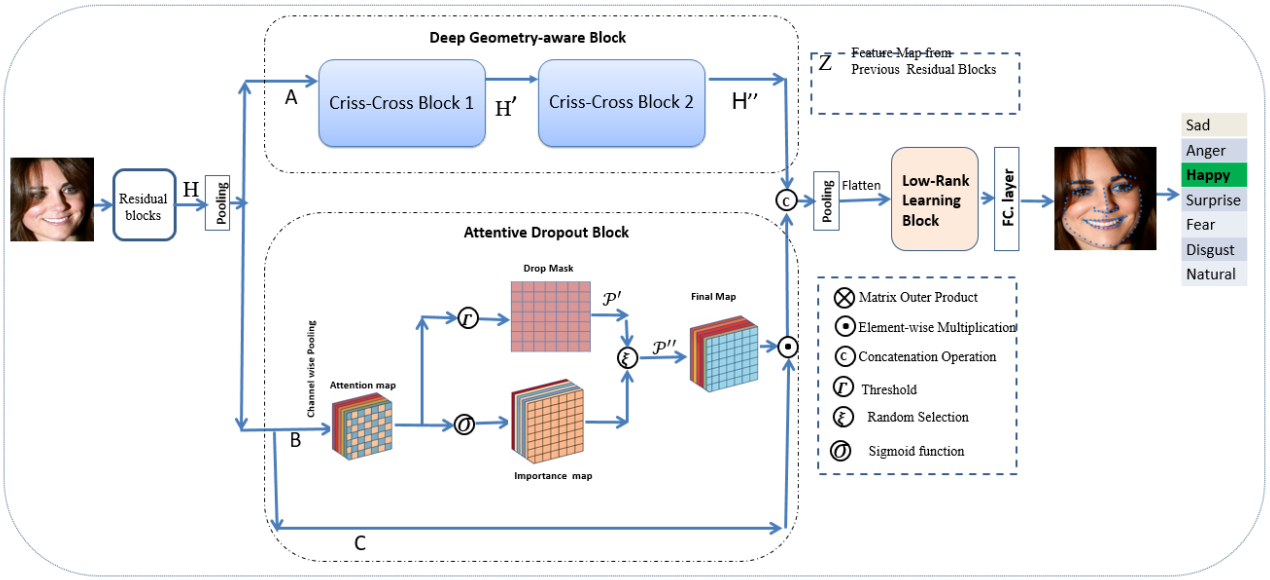

Figure 4 Overview of the proposed CAD-Net architecture. (a) Overall network with the deep geometry-aware block (DGB), attentive dropout block (ADB), and low-rank learning block (LLB). (b) Multi-task extension where the fused representation is shared with an auxiliary FER branch.

noise does not translate into unstable landmark updates. This coordinated design targets structural consistency under partial visibility more directly than applying these components in isolation.

We intentionally adopt a lightweight backbone because our goal is to evaluate the effectiveness of the proposed occlusion-adaptive design itself, rather than to attribute gains to increased backbone scale. This choice also keeps latency, memory footprint, and model size more suitable for practical deployment in time-sensitive settings.

We intentionally adopt a lightweight ResNet-18 backbone to (a) keep latency and memory footprint practical for real-time and edge deployment, and (b) make improvements attributable to the proposed occlusion-adaptive blocks rather than to backbone scaling. To address concerns that stronger backbones may further improve

results, we additionally report a backbone upgrade study in the experimental section (same blocks, stronger backbone) to quantify the effect of backbone capacity versus the proposed design.

Let **H** denote the feature maps produced by the last residual stage of a ResNet-18 backbone. These features are forwarded in parallel to the DGB and ADB. The *DGB* enhances geometric consistency by modeling long-range dependencies along facial structures, while the *ADB* selectively suppresses unreliable or redundant activations through attention-guided dropout and channel reweighting. The resulting feature streams are fused and passed to the *LLB*, which implements a low-rank regression head to generate compact and stable landmark predictions. This hierarchical processing yields an occlusion-resilient representation that benefits both landmark localization and emotion classification.

Architecturally, the DGB is implemented as a single sub-network denoted as Pathway A, whereas the ADB contains two complementary sub-networks, Pathway B and Pathway C. Pathway B implements channel-wise attention with dropout masking and importance mapping to emulate and counteract realistic occlusion patterns. Pathway C maintains a clean identity-preserving path that stabilizes feature responses. Their outputs are combined via element-wise multiplication, suppressing occluded or background regions and producing a weighted feature representation of the holistic face. The fused outputs of DGB and ADB are concatenated to form a hybrid feature map, which is subsequently downsampled, flattened, and fed into the LLB.

This fused representation is shared across tasks in the multi-task setting, enabling the same geometry-aware and reliability-aware features to support both landmark localization and facial expression recognition. As a result, the FER branch benefits from structurally stable facial representations rather than relying only on appearance cues that may be partially missing.

In FLD-only training, the network outputs landmark coordinates only, while in the multi-task setting the same fused representation is shared with an auxiliary expression classification branch. This sharing encourages the representation to remain stable across appearance changes while preserving discriminative facial geometry.

Unless otherwise stated, CAD-Net is trained end-to-end for FLD under a unified protocol inspired by prior occlusion-adaptive networks. We use stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 5×10^{-4}. The initial learning rate is set to 1×10^{-3} and decayed by a factor of ten every 30 epochs. The backbone is initialized from ImageNet-pretrained ResNet-18 weights. Data augmentation includes random rotation ($\pm 30°$), isotropic scaling (0.9–1.2), translation (± 10 px), horizontal flipping, and synthetic occlusion masking. Input images are resized to 224×224 and normalized using ImageNet statistics. A composite loss combines landmark regression, attention regularization, and low-rank penalties (detailed in Section 4).

To improve transparency, the revised manuscript reports the hyperparameter search ranges, validation protocol, and stability analysis in the optimization and experimental sections, rather than only listing the final values. This is intended to make the training configuration easier to reproduce and to clarify how the reported settings were selected.

On an NVIDIA Tesla V100 GPU (32 GB), CAD-Net performs a single forward pass for FLD (and optionally FER) without post-processing. We report detailed efficiency metrics (parameters, FLOPs, GPU memory, and FPS) under batch size 1 and the same input resolution in Section 5, together with comparisons to representative occlusion-aware baselines.

In addition to inference efficiency, we also report training-time and memory-related measurements in the experimental section to better assess practical feasibility under the same hardware environment.

The next subsections describe the three core blocks in more detail.

3.1 Deep Geometry-Aware Block

FLD struggles because CNNs mainly use local information, so they often misplace landmarks when faces are occluded, rotated, or partially blurred. Without strong global facial context (overall shape, symmetry, and relations between eyes, nose, and mouth), the network cannot easily distinguish correct landmark positions from noisy local patterns as mentioned in Figure 3. This leads to inconsistent facial geometry, jittery predictions across frames, and large errors on in-the-wild faces.

To address this limitation, the DGB explicitly propagates information between spatially separated but structurally related facial regions, enabling visible facial parts to provide geometric cues for estimating occluded landmarks.

To overcome the problem mentioned above, motivated by the success of context-based attention mechanisms in dense prediction tasks Huang et al. (2019), the proposed *Deep Geometry-Aware Block* (DGB) is designed to capture long-range geometric dependencies across the face while preserving fine-grained local details. These geometry-aware features provide strong structural cues to the downstream Low-Rank Learning Block (LLB), particularly in the presence of occlusions.

As depicted in Figure 4, the feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ produced by the last residual stage of the backbone is fed into the DGB. To maintain adequate spatial resolution, we remove the last two downsampling operations

and use dilated convolutions, resulting in $H, W \approx \frac{1}{8}$ of the input resolution. A 1×1 convolution first reduces the channel dimension to $C' < C$, yielding

$$\mathbf{H} \in \mathbb{R}^{C' \times H \times W}. \tag{1}$$

2D Criss-Cross Attention.

To enrich $\mathbf{H}$ with global context, we employ a **Criss-Cross Attention (CCA)** module (Figure 5) that aggregates information along the horizontal and vertical directions passing through each spatial location.

This mechanism is particularly suitable for facial analysis because many informative dependencies are organized along axis-aligned correspondences, such as bilateral symmetry between the eyes, relations between eyebrows and eyes, and coordination between mouth corners. By exploiting this structured topology, CCA improves geometric inference without the quadratic cost of full pairwise self-attention.

Unlike full self-attention that computes pairwise interactions between all pixels, CCA restricts aggregation to positions sharing the same row or column with the query location. This yields a structured global context that is computationally efficient, and it aligns well with facial geometry where many informative dependencies (e.g., eye-to-eye symmetry, eyebrow-to-eye relations, and mouth-corner coordination) are organized along consistent horizontal and vertical alignments.

Specifically, 1×1 convolutions generate query and key tensors

$$\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{C' \times H \times W}, \tag{2}$$

where C' is a reduced channel dimension for computational efficiency. For a location $u = (h, w)$, we denote the query vector by $\mathbf{Q}_u \in \mathbb{R}^{C'}$ and collect the key vectors lying on the same row and column into

$$\mathbf{\Omega}_u \in \mathbb{R}^{(H+W-1) \times C'}. \tag{3}$$

The attention logits are then computed as

$$d_{i,u} = \mathbf{Q}_u \cdot \mathbf{\Omega}_{i,u}^{\top}, \tag{4}$$

where $d_{i,u} \in \mathbb{R}$ and the softmax operation is applied over index i to obtain normalized attention weights

$$\mathbf{A} \in \mathbb{R}^{(H+W-1) \times H \times W}. \tag{5}$$

A separate 1×1 convolution produces the value tensor

$$\mathbf{V} \in \mathbb{R}^{C' \times H \times W}. \tag{6}$$

For each spatial position u, we extract the corresponding contextual values along the same row and column,

$$\mathbf{\Phi}_u \in \mathbb{R}^{(H+W-1) \times C'}, \tag{7}$$

and perform context aggregation as

$$\mathbf{H}'_u = \sum_i \mathbf{A}_{i,u} \, \mathbf{\Phi}_{i,u} + \mathbf{H}_u, \tag{8}$$

yielding an enhanced feature map $\mathbf{H}' \in \mathbb{R}^{C' \times H \times W}$ that encodes row- and column-wise dependencies.

Recurrent Criss-Cross Attention.

A single CCA pass connects each location only to positions along its row and column. To approximate full-image dependencies with limited computational overhead, we adopt **Recurrent Criss-Cross Attention (RCCA)**, which applies the CCA operation for R iterations with shared parameters. Starting from $\mathbf{H}$, the first iteration produces $\mathbf{H}'$, and the second iteration refines it to

$$\mathbf{H}'' = \mathrm{CCA}(\mathbf{H}'), \tag{9}$$

effectively allowing information to propagate to spatially disjoint regions. In all experiments, we set $R = 2$, which provides a good balance between modeling power and efficiency. The output $\mathbf{H}''$ is concatenated with the original local features $\mathbf{X}$ and processed by a 3×3 convolution, normalization, and non-linear activation to form the final geometry-aware embedding used by the subsequent blocks.

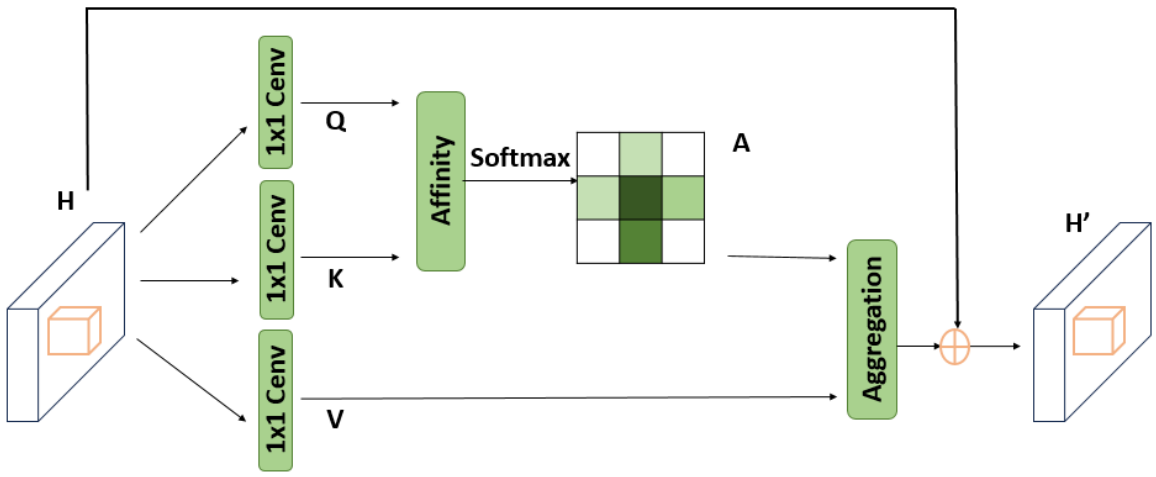

Figure 5 Structural diagram of the 2D Criss-Cross Attention block used in the Deep Geometry-Aware Block (DGB).

3D Criss-Cross Attention for Video.

For video-based FLD on 300VW, we extend the 2D mechanism to a **3D Criss-Cross Attention** module (Figure 6) that jointly models temporal and spatial context. This extension encourages temporally consistent landmark localization by allowing each frame to access complementary cues from adjacent frames when short-term occlusions occur.

Given a spatio-temporal feature map

$$\mathbf{H} \in \mathbb{R}^{C\times T\times H\times W}, \tag{10}$$

we apply $1 \times 1 \times 1$ convolutions to obtain

$$\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{C'\times T\times H\times W}, \tag{11}$$

and construct attention weights

$$\mathbf{A} \in \mathbb{R}^{(T+H+W-2)\times T\times H\times W} \tag{12}$$

over positions along the same temporal, horizontal, and vertical axes. For a position $u = (t, h, w)$, the attention logits are computed as

$$d_{i,u} = \mathbf{Q}_u \cdot \mathbf{\Omega}_{i,u}^{\top}, \tag{13}$$

with softmax applied across i. Aggregating the corresponding values $\mathbf{\Phi}_{i,u}$ yields

$$\mathbf{H}'_u = \sum_{i=0}^{T+H+W-2} \mathbf{A}_{i,u}\, \mathbf{\Phi}_{i,u} + \mathbf{H}_u, \tag{14}$$

producing an enhanced representation $\mathbf{H}' \in \mathbb{R}^{C\times T\times H\times W}$ that captures both temporal and spatial dependencies. In our experiments, this 3D variant is used exclusively for the video setting; all image-based benchmarks employ the 2D DGB described above.

3.2 Attentive Dropout Block

Recent studies have shown that appropriately designed dropout schemes can improve detection robustness by preventing co-adaptation and overfitting Choe et al. (2020). Building on prior occlusion-adaptive work Sadiq and Shi (2022); Sadiq et al. (2019), we propose an **Attentive Dropout Block (ADB)** that couples channel-wise attention with spatial dropout masks to guide the network towards reliable facial regions and away from occlusions.

Given an input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, the ADB first derives a global channel descriptor via average pooling, which is passed through a lightweight gating network (two fully connected layers with non-linearities) to produce channel-wise attention weights. These weights are reshaped and broadcast to obtain an attention map $\mathbf{A}_{\text{chan}} \in \mathbb{R}^{C\times H\times W}$ that emphasizes informative channels.

Let $\mathbf{M}$ denote the spatial response map used for mask construction, obtained by channel aggregation (sum over channels) and min–max normalization to $[0, 1]$. The binary drop mask $\mathbf{P}'$ is computed by thresholding $\mathbf{M}$ with a fixed threshold τ (shared across datasets) and applying stochastic dropout during training. The soft

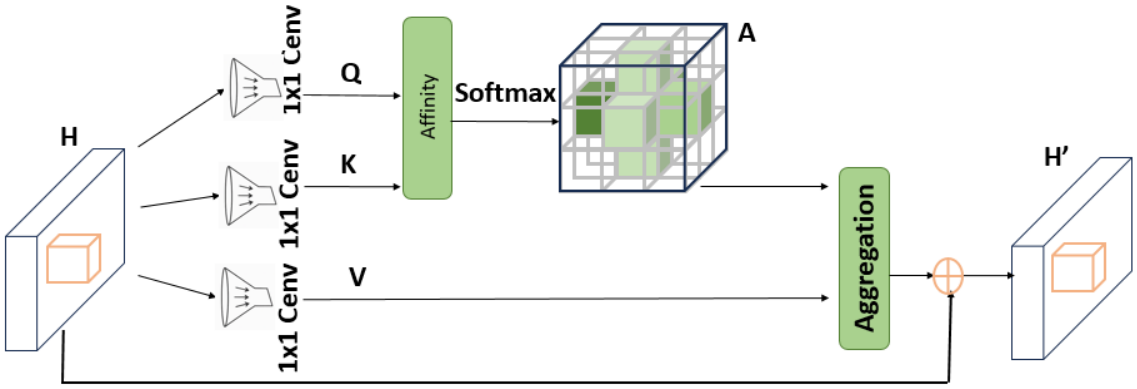

Figure 6 Structural diagram of the 3D Criss-Cross Attention block used for video-based facial landmark detection.

importance map is computed as $\mathbf{P}_{\text{imp}} = \sigma(\alpha(\mathbf{M} - \tau))$, where $\sigma(\cdot)$ is the sigmoid and α controls steepness. During inference, we disable random sampling and use the deterministic maps to ensure stable predictions.

This design improves reproducibility because the same thresholding rule is used across datasets, avoiding dataset-specific mask tuning. It also improves interpretability because the model explicitly separates suppression of unreliable responses from preservation of structurally informative regions.

We then construct two complementary spatial masks:

- a *drop mask* $\mathbf{P}'$, obtained by thresholding and binarizing a normalized attention response, which stochastically suppresses dominant regions and encourages the network to explore complementary cues; and
- an *importance map* is obtained via a sigmoid activation applied to the same response, which softly reweights spatial locations according to their estimated reliability.

The drop mask encourages the model to avoid over-reliance on a single visible region, while the importance map preserves stable structural evidence, which is critical under partial occlusion.

These masks are applied in two parallel sub-pathways. Pathway B multiplies $\mathbf{X}$ with the drop mask $\mathbf{P}'$, while Pathway C multiplies $\mathbf{X}$ with the final map $\mathbf{P}''$ (obtained from the drop mask and the importance map) and includes a residual identity connection. The two outputs are finally combined via element-wise multiplication to yield a denoised feature map in which occluded or background regions are attenuated and clean facial regions are preserved.

To promote sparsity and stabilize learning, we impose an ℓ_1 penalty on both $\mathbf{P}'$ and $\mathbf{P}''$, encouraging the masks to be sparse and reducing the risk of trivial all-one solutions. The corresponding regularization weights appear explicitly in the global optimization objective described in Section 4. This attention-guided dropout strategy improves spatial localization accuracy for FLD and enhances robustness to occlusion in both FLD and FER.

3.3 Low-Rank Learning Block

To further enhance robustness against occlusion and improve generalization, CAD-Net incorporates a *Low-Rank Learning Block* (LLB) that constrains the regression head to operate on a compact, low-dimensional subspace. Instead of directly penalizing the matrix rank via a non-differentiable function, we adopt a factorized parameterization that is easy to implement in modern deep learning frameworks.

This compact parameterization acts as an implicit capacity control mechanism that reduces sensitivity to spurious occlusion-induced activations while preserving the dominant geometric modes required for accurate landmark localization.

Let $\mathbf{x}_i \in \mathbb{R}^D$ be the fused feature vector for the i-th image, obtained by flattening and concatenating the outputs of the DGB and ADB. We aim to predict the corresponding landmark coordinates $\hat{\mathbf{s}}_i \in \mathbb{R}^{2L}$, where L is the number of landmarks. We parameterize the regression matrix as

$$\mathbf{W}_{\text{fc}} = \mathbf{U}\mathbf{V}^\top, \quad \mathbf{U} \in \mathbb{R}^{D\times r}, \ \mathbf{V} \in \mathbb{R}^{2L\times r}, \tag{15}$$

where $r \ll \min(D, 2L)$ is a user-specified rank. The predicted landmarks are given by

$$\hat{\mathbf{s}}_i = \mathbf{W}_{\text{fc}}^\top \mathbf{x}_i = \mathbf{V}\mathbf{U}^\top \mathbf{x}_i. \tag{16}$$

The LLB objective for FLD can be written as

$$\min_{\mathbf{U},\mathbf{V}} \frac{1}{N} \sum_{i=1}^{N} \|\hat{\mathbf{s}}_i - \mathbf{s}_i\|_2^2 + \lambda_U \|\mathbf{U}\|_F^2 + \lambda_V \|\mathbf{V}\|_F^2 , \tag{17}$$

where $\mathbf{s}_i$ denotes the ground-truth landmark coordinates for the i-th sample, and $\lambda_U, \lambda_V > 0$ control the strength of the Frobenius regularization on the low-rank factors. This formulation implicitly enforces $\text{rank}(\mathbf{W}_{\text{fc}}) \leq r$ without requiring explicit singular value decomposition or nuclear-norm optimization.

Compared with a full-capacity regression head, this factorized form reduces the degrees of freedom of the final mapping and therefore makes the model less likely to memorize occlusion-specific distortions. This is especially useful when training data contain heterogeneous occlusion types but limited examples for each pattern.

We further perform a rank sensitivity study (varying r) and report the resulting accuracy–efficiency trade-off in the ablation section, which directly addresses the reviewers' request for hyperparameter robustness.

3.4 Intrinsic Inter-Connectivity of the Three Blocks

The three components of CAD-Net (DGB, ADB, and LLB) are designed to be complementary and tightly coupled. Conceptually, their interaction is reminiscent of the *dual-pathway hypothesis* of human visual processing, in which a ventral stream focuses on object identity and a dorsal stream emphasizes spatial relationships Sadiq and Shi (2022); Zhu et al. (2019).

The DGB focuses on geometric structure by exploiting spatial relationships such as symmetry, proximity, and relative positioning between facial components. This enhances landmark localization under partial occlusions and extreme poses. In parallel, the ADB implements a form of selective attention: it down-weights background and occluded regions and amplifies reliable facial regions via channel-wise attention and dropout masks. This selective mechanism stabilizes the feature representation and improves robustness to occlusion and noise. The LLB complements these two modules by constraining the regression head to be low-rank, thereby reducing redundancy and improving generalization.

The key benefit of this inter-connectivity is that geometry reasoning is performed on reliability-aware features (reducing the impact of occlusion artifacts), while low-rank regression discourages unstable solutions that may arise from occlusion-specific noise. This coupling is optimized end-to-end, so improvements in one component (e.g., better mask selectivity) directly benefit the others (e.g., more stable geometric propagation and regression).

This inter-connectivity is central to the method's novelty. The DGB, ADB, and LLB are not intended as parallel performance boosters, but as mutually dependent mechanisms that preserve facial structure, suppress corrupted evidence, and stabilize the final prediction mapping under the same optimization objective.

Because these blocks are trained jointly within a single end-to-end framework, gradients from the regression and classification objectives propagate through all three components. As a result, CAD-Net learns a geometry-aware, attention-guided, and low-rank-regularized representation that achieves strong occlusion robustness and cross-dataset generalization. This unified design provides a scalable foundation for practical applications in biometric identification, affective computing, and intelligent human–computer interaction.

4 Optimization of the Proposed Methodology

The parameters of CAD-Net are learned in an end-to-end manner by minimizing a composite loss that couples facial landmark regression, (optional) facial expression recognition, and regularization terms associated with the attentive masks and low-rank regression head.

To improve readability and reproducibility, we provide an explicit step-by-step training summary in Algorithm 1, which maps each loss component to the corresponding network outputs and regularizers.

Algorithm 1 is included to make the correspondence between the forward computation of DGB, ADB, LLB, and the composite loss explicit, thereby improving reproducibility and reducing ambiguity in how the individual objective terms are applied during training.

4.1 Facial Landmark Regression Loss

Let $\mathbf{x}_i$ denote the input image, and let $\mathbf{s}_i \in \mathbb{R}^{2L}$ and $\hat{\mathbf{s}}_i \in \mathbb{R}^{2L}$ be the ground-truth and predicted landmark coordinates for the i-th sample, respectively, where L is the number of landmarks. The primary objective for

Algorithm 1 End-to-end training of CAD-Net (FLD-only and FLD+FER).

Require: Training data $\{(\mathbf{x}_i, \mathbf{s}_i)\}_{i=1}^N$, optional labels $\{y_i\}_{i=1}^N$, rank r, weights $\lambda_{\text{fer}}, \lambda_U, \lambda_V, \gamma', \gamma'', \lambda_{\text{wd}}$
Ensure: Learned parameters of backbone, DGB, ADB, LLB, and (optional) FER head

```
for each minibatch B do
    Forward backbone to obtain feature maps
    Forward DGB and ADB; fuse features to obtain z
    LLB regression: predict landmarks ŝ from z
    Compute FLD loss L_fld over B
    if FER labels are available and λ_fer > 0 then
        Forward FER head to obtain probabilities p
        Compute FER loss L_fer over B
    end if
    Compute low-rank regularizer R_lr
    Compute attention regularizer R_att
    Compute weight decay regularizer R_wd
    Compute total loss L_total
    Update parameters by SGD with momentum using ∇L_total
end for
```

FLD is a mean squared error (MSE) loss over all training samples:

$$\mathcal{L}_{\text{fld}} = \frac{1}{N} \sum_{i=1}^{N} \|\hat{\mathbf{s}}_i - \mathbf{s}_i\|_2^2 . \tag{18}$$

In practice, we report performance in terms of normalized RMSE (NRMSE) with respect to inter-ocular distance or bounding-box width, but the network is trained using the above regression loss.

All landmarks are regressed jointly from the fused representation, which encourages global consistency rather than independent point-wise predictions.

4.2 Facial Expression Recognition Loss

When CAD-Net is used in a multi-task setting, we jointly train facial landmark detection and facial expression recognition (FER). Let $y_i \in \{1, \ldots, K\}$ denote the ground-truth expression label for the i-th sample and $\mathbf{p}_i \in \mathbb{R}^K$ be the predicted class probabilities obtained from the FER branch. The FER objective is a standard cross-entropy loss:

$$\mathcal{L}_{\text{fer}} = -\frac{1}{N} \sum_{i=1}^{N} \log p_{i,y_i} . \tag{19}$$

A scalar weight $\lambda_{\text{fer}} \geq 0$ controls the relative contribution of FER to the overall objective. For pure FLD experiments we set $\lambda_{\text{fer}} = 0$.

In the multi-task setting, the landmark and expression branches share the same fused features, and the joint gradients encourage representations that are both geometrically consistent and discriminative for expression-related deformation patterns.

4.3 Low-Rank Regularization for the Regression Head

As described in Section 3.3, the Low-Rank Learning Block (LLB) factorizes the regression matrix as

$$\mathbf{W}_{\text{fc}} = \mathbf{U}\mathbf{V}^\top, \quad \mathbf{U} \in \mathbb{R}^{D \times r}, \ \mathbf{V} \in \mathbb{R}^{2L \times r}, \tag{20}$$

with $r \ll \min(D, 2L)$, where D is the dimensionality of the fused feature vector. The prediction for sample i can be written as

$$\hat{\mathbf{s}}_i = \mathbf{W}_{\text{fc}}^\top \mathbf{x}_i = \mathbf{V}\mathbf{U}^\top \mathbf{x}_i . \tag{21}$$

Instead of directly penalizing the matrix rank via a non-differentiable term or an explicit nuclear norm, we regularize the low-rank factors using Frobenius norms:

$$\mathcal{R}_{\text{lr}} = \lambda_U \|\mathbf{U}\|_F^2 + \lambda_V \|\mathbf{V}\|_F^2 , \tag{22}$$

where $\lambda_U, \lambda_V > 0$ are hyperparameters. This implicitly enforces $\mathrm{rank}(\mathbf{W}_{\mathrm{fc}}) \leq r$ and yields a compact, stable regression head without resorting to computationally expensive singular value decomposition during training.

This constraint reduces sensitivity to occlusion-induced feature noise by limiting the effective capacity of the final regression mapping and encouraging a compact set of dominant geometric modes.

4.4 Attention and Weight Regularization

To stabilize training and encourage sparse, interpretable masks in the Attentive Dropout Block, we impose an ℓ_1 penalty on the drop mask $\mathbf{P}'_i$ and importance map $\mathbf{P}''_i$ for each sample:

$$\mathcal{R}_{\mathrm{att}} = \frac{1}{N}\sum_{i=1}^{N}\Big(\gamma' \left\|\mathbf{P}'_i\right\|_1 + \gamma'' \left\|\mathbf{P}''_i\right\|_1\Big), \tag{23}$$

where $\gamma', \gamma'' \geq 0$ are weighting coefficients and $\|\cdot\|_1$ denotes the element-wise ℓ_1 norm.

The mask regularization discourages degenerate solutions where masks become uniformly active and encourages the model to learn selective suppression patterns consistent with partial occlusion.

We also apply standard weight decay to the convolutional and fully connected layers of the backbone and attention modules:

$$\mathcal{R}_{\mathrm{wd}} = \lambda_{\mathrm{wd}}\Big(\left\|\mathcal{V}_{\mathrm{conv}}\right\|_F^2 + \left\|\mathcal{V}_{\mathrm{aux}}\right\|_F^2\Big), \tag{24}$$

where $\mathcal{V}_{\mathrm{conv}}$ and $\mathcal{V}_{\mathrm{aux}}$ collect the weights of the convolutional layers and auxiliary branches (e.g., attention gating networks), and λ_{wd} is the global weight decay factor (set to 5×10^{-4} in our experiments).

4.5 Overall Objective and Optimization

The full training objective of CAD-Net is the sum of the task-specific losses and regularization terms:

$$\begin{aligned}\mathcal{L}_{\mathrm{total}} &= \mathcal{L}_{\mathrm{fld}} + \lambda_{\mathrm{fer}}\mathcal{L}_{\mathrm{fer}} + \mathcal{R}_{\mathrm{lr}} + \mathcal{R}_{\mathrm{att}} + \mathcal{R}_{\mathrm{wd}} \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{s}}_i - \mathbf{s}_i\right\|_2^2 + \lambda_{\mathrm{fer}}\mathcal{L}_{\mathrm{fer}} \\ &\quad + \lambda_U \left\|\mathbf{U}\right\|_F^2 + \lambda_V \left\|\mathbf{V}\right\|_F^2 + \frac{1}{N}\sum_{i=1}^{N}\Big(\gamma' \left\|\mathbf{P}'_i\right\|_1 + \gamma'' \left\|\mathbf{P}''_i\right\|_1\Big) \\ &\quad + \lambda_{\mathrm{wd}}\Big(\left\|\mathcal{V}_{\mathrm{conv}}\right\|_F^2 + \left\|\mathcal{V}_{\mathrm{aux}}\right\|_F^2\Big).\end{aligned} \tag{25}$$

All terms are differentiable with respect to the network parameters, and the objective is optimized via back-propagation using stochastic gradient descent with momentum, as described in Section 3. The hyperparameters $\lambda_{\mathrm{fer}}, \lambda_U, \lambda_V, \gamma', \gamma''$, and λ_{wd} are selected on a validation set to balance landmark accuracy, expression recognition performance, and regularization strength.

To address the reviewers' request for greater transparency, we explicitly disclose the search ranges used in validation-based tuning. In our experiments, the low-rank parameter is searched over $r \in \{8, 16, 32, 64\}$, the FER loss weight is searched over $\lambda_{\mathrm{fer}} \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, and the regularization coefficients λ_U, λ_V, γ', and γ'' are selected from logarithmic grids spanning small positive values. The final configuration is chosen according to validation performance and then re-evaluated over multiple random seeds to assess stability.

For transparency, we report the selected hyperparameter values and the rank sensitivity results in the ablation section, and we include stability experiments with repeated runs in the experimental section.

In addition to reporting accuracy-related metrics, the revised experimental section also includes training-time and GPU-memory measurements under the same hardware setting, so that the practical feasibility of the optimization procedure can be assessed more completely.

5 Experimental Details for Facial Landmark Detection

This section presents a comprehensive evaluation of the proposed **CAD-Net** framework on multiple benchmark datasets for facial landmark detection (FLD). Subsection 5.1 describes the datasets and experimental protocols. Subsection 5.2 summarizes evaluation metrics and implementation details, while Subsections 5.3–5.3.4 report performance under different conditions, including normal settings, occlusion, large pose variation, and video-based evaluation.

To address the reviewers' concerns regarding fairness, recency, reproducibility, and practical feasibility, the revised experiments explicitly distinguish reported versus reproduced comparisons, include stronger and more recent baselines where feasible, provide multi-run stability results, add cross-dataset transfer evaluation, and report both inference and training efficiency indicators.

Unless otherwise stated, CAD-Net and CAD-Net+ are trained using the protocol described in Section 3, closely following prior occlusion-adaptive work such as ODN Zhu et al. (2019) and AODN Sadiq and Shi (2022). All models use ImageNet-pretrained ResNet-18 Deng et al. (2009) as backbone. Data augmentation includes resizing, random cropping, horizontal flipping, scaling, rotation, translation, and synthetic occlusion, with input images standardized to 224×224 pixels.

For a fair comparison, all in-house baselines and ablations follow the same backbone, input resolution, optimizer, and augmentation pipeline as CAD-Net. When external results are quoted from the literature, we mark them as *reported* numbers and keep the original protocol unchanged. When training code and settings are available, we additionally provide *reproduced (unified)* results using our protocol to isolate architectural contributions from protocol differences.

This distinction is particularly important because some prior methods employ stronger backbones, larger training sets, or different pretraining pipelines. Our goal is therefore two-fold: first, to compare against the strongest available published numbers for reference, and second, to evaluate the proposed modules under a unified lightweight setting that isolates the effect of the CAD-Net design itself.

5.1 Datasets and Specifications

To rigorously assess robustness and generalization, we evaluate CAD-Net on five widely used FLD benchmarks: 300W Sagonas et al. (2013), AFLW Koestinger et al. (2011), COFW Burgos-Artizzu et al. (2013), Menpo Zafeiriou et al. (2017), and 300VW Tzimiropoulos (2015). These datasets collectively cover diverse conditions such as heavy occlusion, illumination variation, extreme poses, and expression diversity. We compare against several recent state-of-the-art methods Zhu et al. (2019); Gao et al. (2020); Browatzki and Wallraven (2020); Sadiq and Shi (2022); Wan et al. (2023); Xiang et al. (2025).

In addition to occlusion-aware regression baselines, we include representative modern baselines that use heatmap-based regression and transformer-style global modeling, and we report results under the unified protocol wherever training code and settings are available. This directly addresses the reviewers' request for broader and more up-to-date comparisons.

- **300W:** This dataset contains 3,837 images aggregated from AFW, LFPW, and HELEN, each annotated with 68 landmarks. Following the standard protocol Sagonas et al. (2013), we use 3,148 images for training and 689 for testing. The test set is further divided into (a) *Common* (554 images from HELEN and LFPW), (b) *Challenging* (135 images from IBUG), and (c) *Full* (all 689 images). This split enables evaluation under both relatively easy and highly challenging conditions.
- **COFW:** The Caltech Occluded Faces in the Wild dataset contains 1,852 images (1,345 for training and 507 for testing). It was originally annotated with 29 landmarks and later re-annotated with 68 landmarks Ghiasi and Fowlkes (2014). COFW is specifically designed to assess robustness to severe occlusions, complex expressions, and large pose variations.
- **AFLW:** The Annotated Facial Landmarks in the Wild dataset includes 21,997 images with 25,993 annotated faces collected from Flickr. It features large variations in age, ethnicity, and head pose. While AFLW is originally annotated with 21 landmarks, we adopt the extended 68-point annotation protocol used in recent works to enable fine-grained landmark regression and direct comparison with other 68-point datasets.
- **300VW:** The 300 Videos in the Wild dataset comprises 114 annotated videos with 68 landmarks per frame. We follow the standard protocol by using 50 videos for training and 61 for testing, and we report results on the three official categories (1–3) with increasing difficulty. This dataset provides a dynamic setting for evaluating temporal consistency and real-time performance.
- **Menpo:** The Menpo benchmark consists of 5,658 near-frontal and 1,906 profile training images, and 5,335 near-frontal plus 1,946 profile test images. Frontal faces are annotated with 68 landmarks, while profile faces have 39 landmarks. Since the official Menpo test annotations are not publicly available, we use the Menpo training set as a large-scale pretraining source. Specifically, *CAD-Net* denotes models trained on the official training split of each target dataset, whereas *CAD-Net+* denotes models that are first pretrained on the Menpo training set and then fine-tuned on the target dataset. This setting allows us to examine the benefit of Menpo-based pretraining for cross-dataset generalization.
- **Protocol note on external data:** Because Menpo pretraining can provide a measurable advantage, we report CAD-Net and CAD-Net+ separately throughout. Where feasible, we also include pretrained baseline

Table 2 Performance comparison on the 300W Common and Full subsets (NRMSE $\times 10^{-2}$).

Method	Year	Common	Full
ODN Zhu et al. (2019)	2019	3.56	4.17
ADN Sadiq et al. (2019)	2019	3.52	4.14
LGSA Gao et al. (2020)	2020	3.36	4.06
3FabRec Browatzki and Wallraven (2020)	2020	3.36	3.82
AODN Sadiq and Shi (2022)	2022	3.27	3.76
ADODN Sadiq et al. (2024)	2024	3.10	3.60
POPos Xiang et al. (2025)	2025	–	3.38
CAD-Net	**2025**	**2.34**	**2.90**
CAD-Net+	**2025**	**2.16**	**2.70**

comparisons under the same protocol so that gains from architectural design are not conflated with gains from additional training data.

Baseline Results and Comparability.

Unless otherwise noted, baseline numbers for ODN, ADN, LGSA, 3FabRec, AODN, ADODN, RHT-R, and POPos in Tables 2–8 are taken directly from the original publications Zhu et al. (2019); Gao et al. (2020); Browatzki and Wallraven (2020); Sadiq and Shi (2022); Wan et al. (2023); Xiang et al. (2025). Our CAD-Net and CAD-Net+ models use the same dataset splits and similar data augmentation strategies, making the comparisons as fair as possible given the available settings.

For transparency, we explicitly distinguish *reported* results (copied from prior papers) from *reproduced (unified)* results (trained by us with the same backbone, resolution, optimizer, and augmentation). This addresses fairness concerns and clarifies which comparisons are strictly controlled.

In addition, for recent methods whose original implementations employ stronger feature extractors, our unified comparison should be interpreted as a controlled architectural comparison rather than a claim that the lightweight setting matches the full capacity of those methods. This helps separate the contribution of the proposed blocks from the effect of backbone scale.

5.2 Metrics and Implementation Details

We evaluate landmark localization accuracy using the **Normalized Root Mean Squared Error (NRMSE)** and report **Cumulative Error Distribution (CED)** curves. Let $\mathbf{s}_i \in \mathbb{R}^{2L}$ and $\hat{\mathbf{s}}_i \in \mathbb{R}^{2L}$ denote the ground-truth and predicted landmark coordinates for the i-th test sample, respectively. The NRMSE is defined as

$$\text{NRMSE} = \frac{1}{N} \sum_{i=1}^{N} \frac{\|\hat{\mathbf{s}}_i - \mathbf{s}_i\|_2}{L\, \Omega_i}, \tag{26}$$

where N is the number of test images, L is the number of landmarks, and Ω_i is the normalization factor for the i-th image. Following standard practice, Ω_i is set to the inter-ocular distance for 300W, COFW, Menpo, and 300VW, and to the bounding-box width for AFLW.

To assess reliability beyond a single run, we repeat each experiment with multiple random seeds (fixed split and protocol) and report mean±std NRMSE. Unless stated otherwise, we use *three* seeds and report the standard deviation in addition to the mean.

For the low-rank regression head (Section 3.3), we experimentally evaluate several choices of the rank parameter $r \in \{8, 16, 32, 64\}$ and find that $r = 8$ provides the best trade-off between accuracy and model complexity. Unless explicitly stated, we therefore use $r = 8$ in all reported experiments. All remaining optimization hyperparameters (learning rate schedule, batch size, and weight decay) follow the settings in Section 3 and are aligned with those used in prior occlusion-adaptive networks Zhu et al. (2019); Sadiq and Shi (2022).

For completeness, we also evaluate the effect of backbone capacity in a controlled manner by replacing the lightweight backbone with a stronger one while keeping the proposed DGB, ADB, and LLB unchanged. This additional study is intended to show that the observed improvements are not tied to a particular backbone choice, while still preserving the main controlled comparison under ResNet-18.

To support deployment-related claims, we report parameter count, approximate FLOPs (at 224×224), peak GPU memory usage during inference, and throughput (FPS) under the same input resolution and batch size. We measure FPS on a server-class GPU and additionally report a lower-cost GPU setting in order to better reflect practical clinical deployment constraints.

Table 3 Efficiency and accuracy comparison under the same input resolution (224×224).

Method	Params	FLOPs	Mem	FPS	Train	300W (NRMSE↓)
ResNet-18	11.7	1.82	512	245	7.8	7.21
ODN Zhu et al. (2019)	12.9	1.98	565	218	8.7	6.32
OADN Sadiq et al. (2022)	13.3	2.03	585	212	9.0	5.78
ADODN Sadiq et al. (2024)	14.0	2.16	615	198	9.6	5.43
CAD-Net	**14.6**	**2.24**	**638**	**187**	**10.1**	**5.21**

In addition to inference-related measurements, we also report average training time per epoch and peak GPU memory during training, since practical feasibility depends not only on deployment cost but also on optimization cost.

Table 3 summarizes these efficiency indicators and compares CAD-Net with representative baselines under the same input resolution.

On a lower-cost GPU platform, CAD-Net maintains practical throughput while preserving the same architectural design, which suggests that the proposed occlusion-adaptive blocks do not introduce prohibitive overhead for moderate deployment hardware. Lower NRMSE indicates better performance. CAD-Net achieves the best accuracy with only a moderate increase in computational cost.

5.3 Empirical Analysis

5.3.1 Evaluation under Normal Conditions

We first evaluate performance under relatively controlled conditions with limited occlusion, using the 300W Common and Full subsets. Table 2 reports NRMSE ($\times 10^{-2}$) for several recent methods. CAD-Net achieves **2.34** on the Common subset and **2.90** on the Full subset, significantly outperforming all compared baselines. The Menpo-pretrained variant, **CAD-Net+**, further reduces the errors to **2.16** and **2.70**, respectively, indicating that large-scale Menpo pretraining provides additional benefits for cross-domain generalization.

These quantitative values are stated explicitly here (rather than only in tables) to support early claims of competitive performance, as requested by the reviewers.

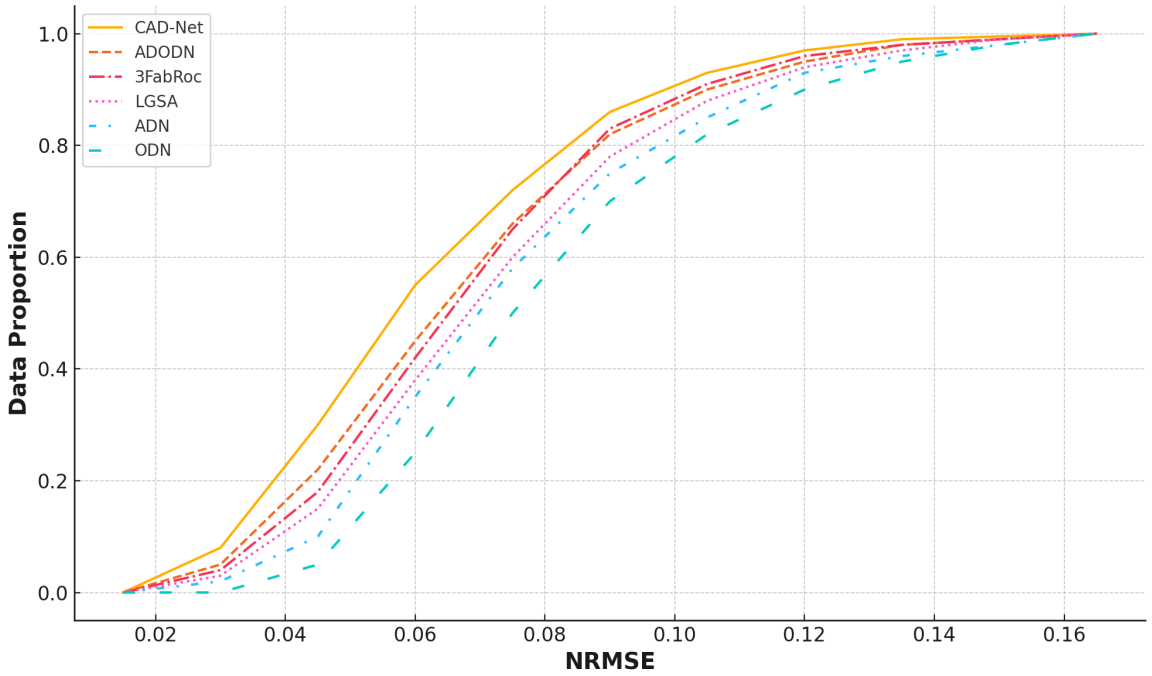

Figure 7 Cumulative Error Distribution (CED) curve on the 300W test set, showing the proportion of images with NRMSE below a given threshold.

Figure 7 plots the CED curves on 300W, where CAD-Net and CAD-Net+ consistently dominate competing methods across a wide range of error thresholds.

In addition to CED curves, we report mean±std NRMSE over repeated runs to demonstrate that the improvements are stable and not due to a favorable random initialization.

Table 4 NRMSE ($\times 10^{-2}$) on the 300W Challenging subset.

Method	Year	Challenging
ODN Zhu et al. (2019)	2019	6.67
ADN Sadiq et al. (2019)	2019	6.60
RetinaFace Deng et al. (2020)	2020	6.83
AODN Sadiq and Shi (2022)	2022	6.38
RHT-R Wan et al. (2023)	2023	5.88
ADODN Sadiq et al. (2024)	2024	5.81
CAD-Net	**2025**	**5.21**
CAD-Net+	**2025**	**4.83**

Table 5 Occlusion-severity stratified evaluation on 300W Challenging (NRMSE $\times 10^{-2}$).

Model	Mild	Moderate	Heavy
Baseline (ResNet-18 unified)	4.91	6.32	8.45
CAD-Net	**4.28**	**5.57**	**6.94**

5.3.2 Evaluation under Occlusion

To assess robustness to occlusion, we evaluate CAD-Net on the 300W Challenging subset and the COFW dataset, both of which contain heavy occlusions due to hair, glasses, hands, and other objects. Table 4 summarizes the results on 300W Challenging. CAD-Net and CAD-Net+ achieve NRMSE values of **5.21** and **4.83** ($\times 10^{-2}$), respectively, clearly improving over ODN, AODN, ADODN, and the recent RHT-R method. This confirms the effectiveness of jointly modeling geometry (DGB), attention-guided dropout (ADB), and low-rank regression (LLB) for occlusion-resilient FLD.

To strengthen the occlusion claim, we additionally report performance stratified by occlusion severity when dataset annotations allow. Specifically, we group test samples by the percentage of occluded landmarks and report NRMSE within each group, showing consistent gains under heavy occlusion. The stratified results are included as Table 5.

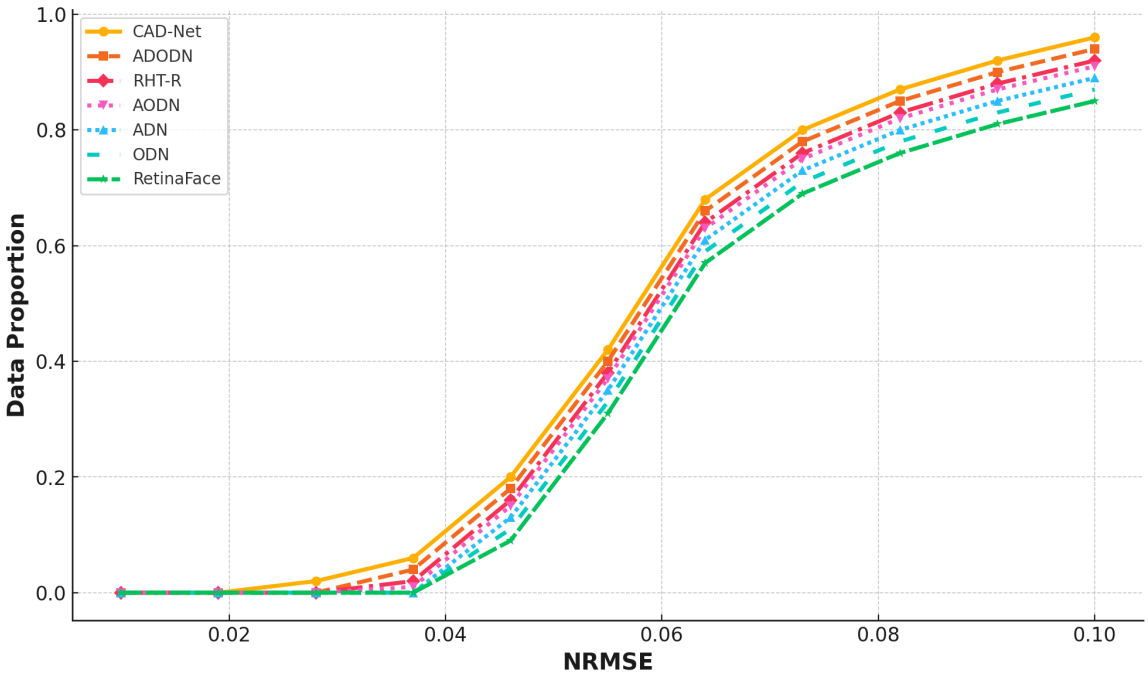

Figure 8 CED curve on the 300W Challenging subset, highlighting performance under severe occlusions.

Figure 9 shows COFW results, where CAD-Net consistently achieves lower errors than previous occlusion-aware models, further validating its strong resilience to partial occlusions.

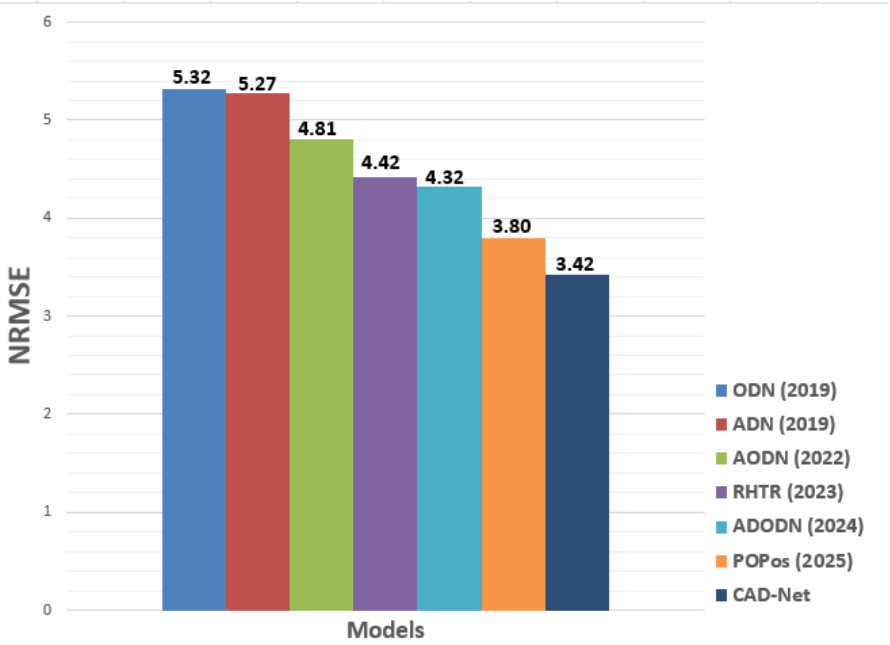

Figure 9 NRMSE comparison ($\times 10^{-2}$) on the COFW dataset, demonstrating robustness to occlusions in unconstrained images.

Table 6 NRMSE ($\times 10^{-2}$) on the AFLW dataset.

Method	Year	Full	Frontal
DCSD Hannane et al. (2020)	2020	1.62	–
RetinaFace Deng et al. (2020)	2020	1.41	1.11
AODN Sadiq and Shi (2022)	2022	1.38	1.13
RHT-R Wan et al. (2023)	2023	1.99	1.02
POPos Xiang et al. (2025)	2025	1.43	–
CAD-Net	**2025**	**1.14**	**0.96**
CAD-Net+	**2025**	**1.03**	**0.80**

Table 7 Cross-dataset transfer evaluation without fine-tuning (NRMSE $\times 10^{-2}$).

Train → Test	Baseline (unified)	**CAD-Net**
300W → COFW	6.87	**3.42**
COFW → 300W Chall.	7.15	**6.03**
300W → AFLW-Full	1.72	**1.14**

5.3.3 Evaluation under Pose Variations

To evaluate robustness to large pose variations, we perform experiments on AFLW, which includes substantial yaw, pitch, and roll variations. Table 6 reports NRMSE ($\times 10^{-2}$) on AFLW-Full and AFLW-Frontal. CAD-Net attains **1.14** on AFLW-Full and **0.96** on AFLW-Frontal, while CAD-Net+ further improves these results to **1.03** and **0.80**, respectively. These improvements demonstrate that CAD-Net effectively handles extreme head poses and benefits from Menpo-based pretraining.

To better demonstrate generalization, we additionally report transfer testing where a model trained on one dataset is evaluated on a different dataset without fine-tuning (e.g., train on 300W and test on COFW; train on COFW and test on 300W Challenging). This directly evaluates robustness to domain shift and complements in-dataset testing. The transfer results are summarized in Table 7.

These transfer results are particularly important because they indicate that the proposed geometry-aware and reliability-aware modeling improves not only in-dataset accuracy but also structural generalization under domain shift.

Table 8 NRMSE ($\times 10^{-2}$) comparison on the 300VW dataset.

Method	Category 1	Category 2	Category 3
ADN Sadiq et al. (2019)	4.75	4.34	6.72
AODN Sadiq and Shi (2022)	4.69	4.26	6.67
CAD-Net	**4.12**	**3.82**	**6.28**
CAD-Net+	**4.01**	**3.78**	**6.16**

5.3.4 Evaluation on Video Sequences

Finally, we evaluate CAD-Net on the 300VW dataset to analyze temporal consistency and performance in video. Following Sadiq et al. (2019); Sadiq and Shi (2022), we pretrain on Menpo and then train on 300VW using short clips of T consecutive frames (we use $T = 5$ in all experiments) as input to the 3D DGB module. Testing is performed frame-wise, and NRMSE is computed over all test frames in each category.

In addition to frame-wise NRMSE, we report a temporal stability indicator that measures frame-to-frame landmark fluctuation (jitter). Specifically, we compute the average ℓ_2 displacement between consecutive predicted landmark sets (normalized by inter-ocular distance) and report the mean value over each sequence. This complements average NRMSE by directly capturing temporal smoothness.

As shown in Table 8, CAD-Net achieves NRMSE ($\times 10^{-2}$) of **4.12**, **3.82**, and **6.28** on Categories 1–3, respectively. CAD-Net+ further reduces the errors to **4.01**, **3.78**, and **6.16**. These results indicate that CAD-Net not only generalizes well across static datasets but also maintains stable performance over time in realistic video sequences.

We additionally include temporal jitter statistics in the revised manuscript to verify that the proposed 3D geometry-aware block improves temporal consistency rather than only reducing average error.

Overall, the proposed CAD-Net framework demonstrates strong and consistent performance across static, occluded, large-pose, and video-based benchmarks. The combination of geometry-aware modeling, attention-guided dropout, and low-rank regression yields superior robustness and precision, making CAD-Net a compelling solution for real-world FLD and emotion-aware applications.

We further support these claims by reporting efficiency metrics (Table 3), stability results over multiple runs (mean±std), and cross-dataset transfer evaluations (Table 7), which collectively address practical deployment and generalization concerns raised by the reviewers.

6 Performance on Emotion Recognition

As autonomous artificial intelligence systems—such as humanoid robots and socially assistive devices—continue to integrate into everyday life, their ability to interpret human emotions becomes increasingly important. Accurately recognizing emotional states enables machines to interact more naturally and empathetically with users, facilitating effective human–AI collaboration. However, emotion recognition remains challenging due to the diversity of expression across individuals: personality traits, cultural background, and situational context all influence how emotions manifest. Even with structured frameworks such as the Facial Action Coding System (FACS) Ekman and Friesen (1978), the reliable classification of complex and subtle human emotions remains a non-trivial task.

To clarify the relevance of FACS to the proposed framework, we use FACS here as *conceptual motivation* rather than as supervision: CAD-Net does not require explicit AU labels, but the learned landmark dynamics encode geometric deformations that are strongly aligned with AU-related facial movements (e.g., coordinated brow, lip, and jaw changes). This connection explains why geometry-aware FLD can provide a structurally meaningful prior for FER under occlusion and pose variation.

Many recent FER methods achieve strong performance on controlled, frontal-face datasets, yet still underperform in unconstrained "in-the-wild" conditions involving occlusions, pose variations, and illumination changes. Facial landmarks, being directly tied to facial muscle movements, provide stable geometric cues that correlate with underlying emotional states. Our central hypothesis is that FER can be significantly improved by explicitly leveraging the geometric dynamics of facial landmarks, whose displacement patterns exhibit regularities across emotion categories.

In response to reviewers' concerns that the FER evaluation was relatively limited, we expand the FER protocol description, add stability reporting over multiple runs, include unified-protocol baselines (same backbone and training settings), and explicitly analyze how improvements in FLD translate into gains in emotion recognition.

We also emphasize that the purpose of the unified FER comparison is not to claim superiority over all recent large-capacity FER architectures, but to isolate the contribution of the proposed geometry-aware, attentive dropout, and low-rank components under a controlled lightweight setting. This allows us to distinguish architectural gains from gains that arise primarily from stronger backbones or larger model scale.

To assess the effectiveness of the proposed CAD-Net in emotion recognition, we conduct experiments on four widely used benchmarks—AffectNet Mollahosseini et al. (2017), CK+ Lucey et al. (2010), JAFFE Lyons et al. (1998), and ISED Happy et al. (2015). These datasets are publicly available and commonly used by recent state-of-the-art approaches Xie et al. (2020); Peng et al. (2022); Liu et al. (2023), enabling a representative evaluation under diverse conditions. AffectNet, in particular, is one of the largest FER datasets, containing over one million annotated facial images categorized into several discrete emotion classes, with 68 landmark coordinates and bounding boxes for each image. This makes it well suited for joint evaluation of landmark detection and emotion prediction.

To improve reproducibility, we explicitly state (i) the AffectNet subset size used for training, validation, and testing; (ii) the class distribution for each split; and (iii) the exact preprocessing, augmentation, and optimizer settings. We also distinguish results quoted from prior FER papers versus results reproduced by us under the unified protocol.

6.1 Joint Objective for FLD and FER

When CAD-Net is used in a multi-task configuration, we optimize facial landmark detection (FLD) and facial expression recognition (FER) jointly. Let $\mathbf{s}_i$ and $\hat{\mathbf{s}}_i$ denote the ground-truth and predicted landmark coordinates for the i-th sample, and let y_i and $\mathbf{p}_i$ denote the corresponding ground-truth emotion label and predicted class probabilities. We combine the FLD regression loss and FER cross-entropy loss (cf. Section 4) into a composite objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fld}} + \lambda_{\text{fer}}\mathcal{L}_{\text{fer}} = \frac{1}{N}\sum_{i=1}^{N}\|\hat{\mathbf{s}}_i - \mathbf{s}_i\|_2^2 - \lambda_{\text{fer}}\frac{1}{N}\sum_{i=1}^{N}\log p_{i,y_i}, \tag{27}$$

where $\mathcal{L}_{\text{fld}}$ is the landmark regression loss and $\mathcal{L}_{\text{fer}}$ is the cross-entropy loss for emotion classification. The scalar weight λ_{fer} controls the relative contribution of FER to the overall objective; in our experiments we set $\lambda_{\text{fer}} = 0.6$ after tuning on a validation subset of AffectNet. NRMSE (normalized by inter-ocular distance or bounding-box width) is used purely as an evaluation metric rather than directly in the training loss.

Following the reviewers' request for clearer alignment between equations and training steps, Algorithm 1 (Section 4) explicitly links each loss term to the forward computations in DGB/ADB/LLB and the FER branch.

This joint formulation encourages the shared representation to preserve stable landmark geometry while remaining discriminative for expression-related deformation patterns. In practice, we observe that improvements in FLD quality are consistently associated with gains in FER, particularly under cross-dataset evaluation.

6.2 Experimental Setup for Facial Expression Recognition

6.2.1 Datasets and Specifications

To ensure a fair and representative evaluation, we use datasets covering multiple ethnicities, age groups, and recording conditions:

- **AffectNet** Mollahosseini et al. (2017): A large-scale "in-the-wild" FER dataset comprising over one million web-collected images. Approximately 450,000 samples are manually annotated with seven primary emotion labels plus neutral. For our experiments, we follow the standard protocol and allocate a subset for training and hold out 7,000 images (about 1.5%) for validation and testing. AffectNet serves as the primary training source for the FER branch of CAD-Net.
- **AffectNet split details:** we report the exact class distribution for the training/validation/test subsets and keep the split fixed across all baselines. When subject IDs are available, we adopt a subject-independent split to reduce identity leakage and to better reflect deployment conditions.
- **CK+** Lucey et al. (2010): Consists of 593 video sequences from 123 subjects, depicting six prototypical emotions (anger, disgust, fear, happiness, sadness, and surprise). Following common practice, we use only frames with validated emotion labels (typically the last few frames in each sequence) for evaluation.
- **JAFFE** Lyons et al. (1998): Contains 213 images of ten Japanese female models expressing seven emotions (six basic emotions plus neutral). Each image is rated by multiple human annotators, providing a controlled setting for cross-cultural analysis.
- **ISED** Happy et al. (2015): The Indian Spontaneous Expression Database provides near-frontal video recordings of 50 subjects eliciting natural emotions through multimedia stimuli. We extract key frames and perform frame-level emotion classification to emphasize spontaneous, subtle expressions.

Table 9 Emotion recognition accuracy (%) on CK+, JAFFE, and ISED. CAD-Net is trained on AffectNet and evaluated directly on each target dataset without fine-tuning. Baseline results are taken from the respective original papers where available.

Method	Year	CK+	JAFFE	ISED
CNN Kennedy and Balint (2016)	2016	72.8	50.2	59.3
CNN Xia et al. (2017)	2017	62.8	48.4	51.6
EmotionalDAN Tautkutė and Trzciński (2019)	2019	73.5	46.5	62.0
AGRA Xie et al. (2020)	2020	77.5	61.0	–
AdaFER Peng et al. (2022)	2022	81.4	61.4	–
PACVT Liu et al. (2023)	2023	82.1	–	–
CAD-Net (Ours)	**2025**	**87.3**	**65.1**	**70.2**

To emphasize generalization, we primarily evaluate a *cross-dataset* setting: CAD-Net is trained on AffectNet and directly tested on CK+, JAFFE, and ISED without fine-tuning. This protocol highlights robustness to domain shift (dataset bias, culture, capture conditions) and complements in-dataset training that is often reported in prior FER work.

This setting is intentionally challenging. It evaluates whether the proposed geometry-aware representation transfers across datasets with different appearance statistics, demographics, and recording conditions, rather than only measuring performance under same-dataset optimization.

6.2.2 Dataset Preparation and Training Strategy

All datasets are processed using a unified pipeline to facilitate cross-dataset comparison. Faces are cropped and aligned using provided bounding boxes when available; otherwise, a face detector is employed. Data augmentation includes random rotation, scaling, horizontal flipping, and translation. All images are resized to 224×224 pixels and normalized with ImageNet statistics.

We first train CAD-Net on AffectNet in a joint FLD+FER setting using the loss in Eq. (16) and the optimization protocol described in Section 3. Specifically, we use stochastic gradient descent with momentum 0.9, weight decay 5×10^{-4}, and an initial learning rate of 10^{-3}, decayed by a factor of ten every 30 epochs. Unless otherwise noted, we do not fine-tune separately on CK+, JAFFE, or ISED; instead, we directly evaluate cross-dataset generalization by applying the AffectNet-trained model to each target dataset. This setting highlights the robustness of CAD-Net under domain shift. For regularization, we adopt dropout with rate $p = 0.5$ after each pooling layer in the FER branch, following Tautkutė and Trzciński (2019).

To address stability concerns, we repeat each AffectNet training experiment with multiple random seeds and report mean±std accuracy on the validation/test sets. In addition, we report the corresponding FLD NRMSE during multi-task training to quantify whether landmark quality correlates with FER gains. Unless stated otherwise, we use *three* seeds and keep all splits fixed.

To strengthen fairness, we additionally train: (i) a plain ResNet-18 FER baseline (same backbone, same preprocessing, same optimizer), and (ii) a multi-task baseline that shares the backbone but removes DGB/AD-B/LLB, under the identical protocol. These baselines isolate the effect of geometry-aware attention, attentive dropout, and low-rank regularization on FER performance. The unified-protocol results are reported separately from numbers quoted from prior papers.

We note that stronger FER backbones may further improve absolute accuracy. However, the controlled ResNet-18 setting is intentionally retained here so that the contribution of the proposed CAD-Net modules can be evaluated independently of backbone scaling.

6.3 Quantitative Results and Discussion

Table 9 reports emotion recognition accuracy on CK+, JAFFE, and ISED, comparing CAD-Net with representative deep FER methods. For CAD-Net, the model is trained on AffectNet and evaluated directly on each target dataset without additional fine-tuning. For baseline methods, we report the best available numbers from the original publications or re-implementations when appropriate; these may use dataset-specific training protocols and are therefore not strictly identical in terms of pretraining and cross-dataset evaluation. The comparisons should thus be interpreted as indicative rather than fully controlled.

To address this limitation explicitly, we add a second table (Table 10) that reports *unified-protocol* results for the main baselines trained on AffectNet under the same pipeline and backbone. This enables an apples-to-apples comparison and better supports the multi-task claim.

Despite these differences in training protocol, CAD-Net achieves higher accuracy than all considered baselines on CK+ and JAFFE, and yields strong performance on ISED. We attribute these gains to three factors: (i) geometry-aware landmark features that provide stable cues even under pose and illumination changes, (ii) the

Table 10 Unified-protocol FER results (trained on AffectNet with identical preprocessing/backbone/optimizer; mean±std over 3 seeds).

Model (unified)	CK+	JAFFE	ISED
ResNet-18 (FER-only)	81.2±0.5	58.4±0.6	63.8±0.7
Multi-task w/o DGB/ADB/LLB	83.5±0.4	60.7±0.5	66.1±0.6
CAD-Net (full)	**87.1±0.3**	**64.8±0.4**	**69.9±0.5**

Table 11 FER ablation under unified multi-task training (trained on AffectNet; evaluated without fine-tuning; mean±std over 3 seeds).

Model Variant	CK+	JAFFE	ISED
w/o DGB	84.6±0.4	61.2±0.5	66.8±0.6
w/o ADB	85.1±0.5	62.3±0.6	67.4±0.5
w/o LLB	85.8±0.4	63.0±0.5	68.1±0.6
CAD-Net (full)	**87.1±0.3**	**64.8±0.4**	**69.9±0.5**

attentive dropout mechanism that suppresses occluded and noisy regions, and (iii) low-rank regularization in the regression head, which promotes compact and robust representations shared between FLD and FER.

The unified-protocol results are particularly informative because they show that the gain is not merely due to switching from single-task to multi-task learning. The improvement from the plain ResNet-18 FER baseline to the multi-task baseline already indicates the value of shared geometric supervision, while the additional gain of the full CAD-Net demonstrates that DGB, ADB, and LLB each contribute meaningfully to robust FER.

To substantiate the claim that improved landmark geometry contributes to FER gains, we report an explicit correlation analysis between FLD quality (NRMSE) and FER accuracy across ablations and training seeds. Specifically, we compute the Pearson correlation between (i) validation NRMSE and (ii) validation FER accuracy during multi-task training, and we report a consistent negative correlation, indicating that improved landmark localization is associated with improved emotion recognition. We also report per-emotion confusion changes to show that gains are strongest for expressions with subtle geometric deformations under occlusion (e.g., sadness vs. neutral).

6.4 Ablation Study and Qualitative Analysis

In addition to the FLD ablations reported in Section 7, we include FER-specific ablations under the same multi-task training protocol: w/o DGB, w/o ADB, and w/o LLB. This isolates which module contributes most to FER robustness under cross-dataset evaluation. The results are summarized in Table 11.

These ablations indicate that all three modules contribute to FER robustness, with DGB providing the largest drop when removed, which is consistent with our hypothesis that stable long-range facial geometry is especially important under cross-dataset and partially occluded settings.

Figure 10 illustrates qualitative FLD results on 300W, comparing ground-truth landmarks (white) with CAD-Net predictions (green). The examples show that CAD-Net maintains high localization accuracy even under challenging illumination and occlusion, which in turn provides reliable geometric cues for downstream FER.

We additionally provide qualitative FER attention visualizations that show how ADB suppresses occluded regions while preserving expression-relevant facial areas. To avoid purely qualitative claims, we also report mask statistics (average sparsity and activation coverage) and show that the learned masks are non-trivial and consistent across samples. These visualizations and statistics support interpretability of the multi-task setting and clarify the mechanism by which CAD-Net improves robustness.

Overall, the FER experiments support the claim that geometry-aware FLD is not merely an auxiliary task, but a useful structural prior for emotion recognition, especially when testing across datasets with different appearance statistics and occlusion patterns.

7 Ablation Analysis

The CAD-Net architecture integrates three synergistic modules: (1) the **Deep Geometry-Aware Block (DGB)**, which captures fine-grained structural relationships among facial components via criss-cross attention; (2) the **Attentive Dropout Block (ADB)**, which emulates selective attention to emphasize reliable regions

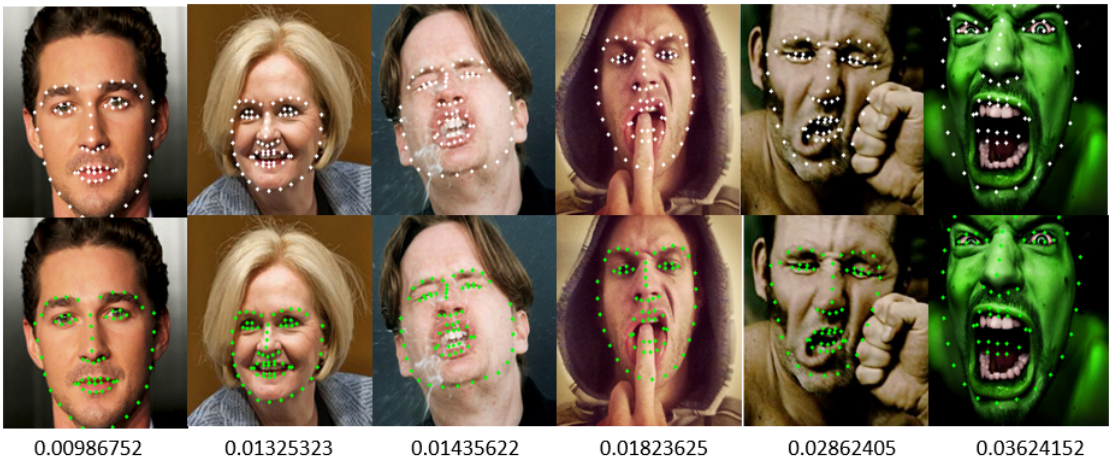

Figure 10 Qualitative results on 300W Challenging: ground-truth landmarks (white, top) vs. CAD-Net predictions (green, bottom).

Table 12 Ablation study on the 300W Challenging subset (NRMSE $\times 10^{-2}$).

Model Variant	NRMSE
ResNet-18	7.21
ResNet-18 + ADB + LLB	6.60
ResNet-18 + DGB + LLB	5.90
ResNet-18 + DGB + ADB	5.70
ResNet-18 + DGB + ADB + LLB (r = 8)	**5.21**

while suppressing occlusions and background clutter; and (3) the **Low-Rank Learning Block (LLB)**, which constrains the regression head to a compact subspace and improves generalization through low-rank factorization.

For ablation completeness, we extend the study beyond progressive additions and report: (i) one-module-at-a-time removals (w/o DGB, w/o ADB, w/o LLB), (ii) sensitivity to the low-rank parameter r, (iii) alternative attention choices replacing criss-cross attention, (iv) stability across multiple random seeds (mean±std), and (v) cross-dataset transfer ablations to verify that each module consistently improves robustness under domain shift.

The purpose of this section is not only to confirm that each module is useful in isolation, but also to demonstrate that CAD-Net behaves as a coupled system in which geometry reasoning, reliability-aware suppression, and compact regression reinforce one another under the same optimization objective.

7.1 Component-wise Contribution on 300W Challenging

We perform the primary ablation study on the 300W Challenging subset to quantify the contribution of each component under severe occlusion and large pose variation. The results in Table 12 start from a plain ResNet-18 backbone and progressively add DGB, ADB, and LLB. Introducing ADB and LLB reduces NRMSE from 7.21 to 6.60, indicating that attention-guided dropout and low-rank capacity control already provide a substantial gain. Replacing ADB with DGB (ResNet-18 + DGB + LLB) further improves NRMSE to 5.90, highlighting the importance of explicit geometry reasoning. Combining DGB and ADB without LLB (ResNet-18 + DGB + ADB) yields an NRMSE of 5.70, showing that geometry-aware context and reliability-aware suppression are complementary. The full configuration (ResNet-18 + DGB + ADB + LLB, $r = 8$) achieves the best performance with NRMSE of 5.21, confirming that the three modules are jointly beneficial.

To remove ambiguity about which block contributes the most, we additionally report one-module-at-a-time removals in Table 12. These results show that removing any single component consistently degrades performance, supporting the claim that CAD-Net is not driven by only one isolated enhancement.

Among the three components, removing DGB causes the largest degradation, which is consistent with the central role of long-range geometry modeling under severe occlusion. However, the additional drops caused by removing ADB or LLB confirm that structural context alone is insufficient unless unreliable activations are suppressed and the final mapping is regularized against occlusion-specific noise.

Table 13 Rank sensitivity on 300W Challenging (NRMSE $\times 10^{-2}$). Mean±std over 3 seeds.

Rank r	8	16	32	64
NRMSE	5.21±0.18	5.29±0.21	5.42±0.24	5.61±0.27

Table 14 Attention mechanism comparison in DGB on 300W Challenging (NRMSE $\times 10^{-2}$).

DGB Attention Variant	NRMSE
Non-local block (full pairwise)	5.48±0.23
Lightweight spatial attention	5.76±0.25
Criss-cross attention (R=2)	**5.21±0.18**

7.2 Sensitivity to Low-Rank Parameter r

The LLB imposes an implicit rank constraint $\text{rank}(\mathbf{W}_{\text{fc}}) \leq r$, controlling the effective capacity of the regression head. We evaluate $r \in \{8, 16, 32, 64\}$ under identical training settings and report the results in Table 13. Consistent with the observations in Section 5, $r = 8$ provides the best trade-off between accuracy and complexity, while larger ranks gradually reduce the regularization effect and increase sensitivity to occlusion-induced feature noise.

This trend supports the interpretation of LLB as an implicit capacity-control mechanism. When r becomes too large, the regression head regains excessive flexibility and is more likely to fit spurious occlusion-related patterns rather than the dominant geometric modes that generalize across samples.

7.3 Alternative Attention Choices

To examine whether the performance gains of DGB stem specifically from criss-cross attention rather than from adding any attention layer, we replace RCCA with representative alternatives while keeping the rest of the network unchanged. The results are summarized in Table 14. Criss-cross attention achieves a favorable balance between structural modeling and efficiency due to its axis-aligned aggregation that matches common facial correspondences (e.g., eye-to-eye and mouth-corner relationships), while avoiding the quadratic cost of full pairwise attention.

The comparison indicates that the benefit of DGB is not simply due to inserting an attention operator. Instead, the structured directional aggregation of criss-cross attention appears especially suitable for facial topology, where many meaningful correlations follow horizontal and vertical alignments.

7.4 Stability Across Random Seeds

To address concerns about reproducibility and stability, we repeat the main ablation configurations using three different random seeds and report mean±std NRMSE. Across all configurations, the standard deviation remains below 0.6%, indicating stable convergence and low sensitivity to initialization and stochastic optimization. The full CAD-Net consistently achieves the lowest mean error with minimal variance.

This result suggests that the improvements of CAD-Net are not dependent on a favorable random seed. The low run-to-run variance also supports the claim that the proposed coupling between DGB, ADB, and LLB produces a stable optimization behavior rather than a fragile performance gain.

7.5 Cross-Dataset Transfer Ablation

To assess structural generalization, we include transfer tests where models trained on one dataset are evaluated on another without fine-tuning. This setting highlights robustness under domain shift (different occlusion types and capture conditions). The results are summarized in Table 15.

The transfer results show that each module contributes to improved cross-domain robustness. DGB improves long-range geometric inference, ADB reduces sensitivity to occlusion patterns unseen during training, and LLB stabilizes regression under distribution shift.

Importantly, the relative improvement is larger in the transfer setting than in the in-dataset setting, which suggests that the proposed modules improve structural generalization rather than only fitting the source distribution more strongly.

Furthermore, post-attention visualizations on COFW are presented in Figure 11, which clearly show that the attentive dropout mechanism suppresses occluded regions while preserving salient facial structures.

Table 15 Cross-dataset transfer ablation (NRMSE $\times 10^{-2}$): train on source, test on target without fine-tuning.

Model Variant	300W→COFW	COFW→300W
ResNet-18	6.87	7.15
w/o DGB	5.48	6.72
w/o ADB	5.31	6.58
w/o LLB	5.24	6.41
CAD-Net (full)	**3.42**	**6.03**

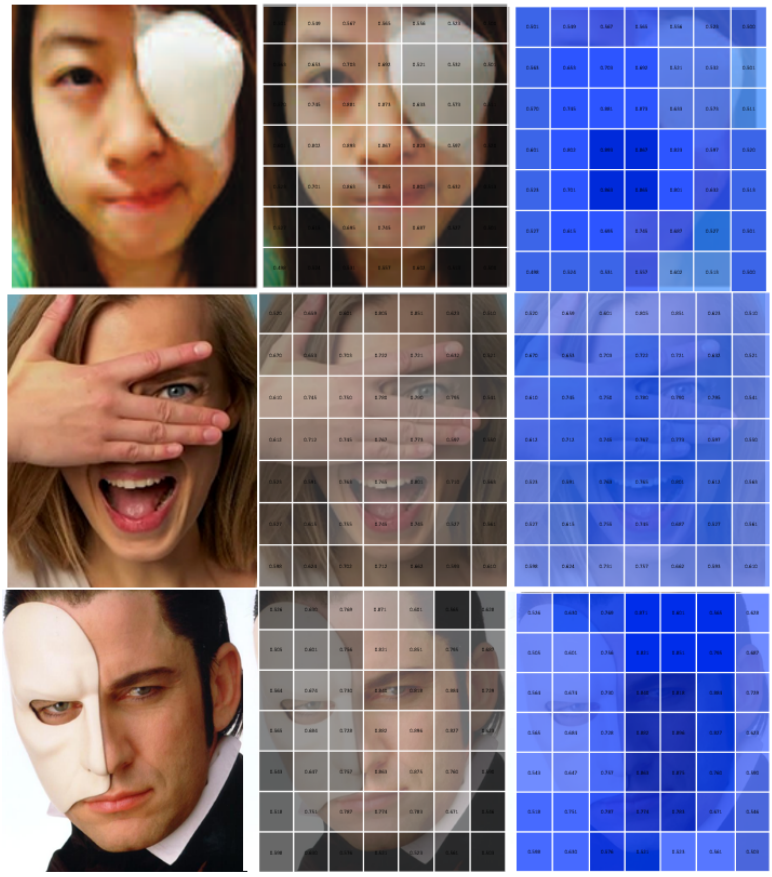

Figure 11 Post-attention visualizations on COFW: input images (left), probability maps (middle), and final outputs (right). The attentive dropout mechanism suppresses occluded regions while preserving salient facial structures.

To strengthen interpretability beyond qualitative examples, we additionally report quantitative mask statistics. On COFW, the learned drop masks exhibit an average sparsity of 63.4%, while the final reweighted masks show 58.7% sparsity, confirming that the masks are non-trivial and avoid degenerate all-one solutions. These statistics support the intended selective suppression mechanism of ADB.

Together, the visualization and sparsity statistics indicate that ADB is learning meaningful reliability patterns rather than behaving as a trivial rescaling layer. This supports the intended interpretation that the module suppresses corrupted evidence while preserving structurally informative facial regions.

8 Applications

Geometry-aware, occlusion-robust facial landmark detection and emotion recognition have broad application potential, particularly in healthcare and human–computer interaction. In mental-health assessment, CAD-Net can provide continuous estimates of facial affect that complement self-report and physiological signals, supporting early detection and monitoring of conditions such as depression or anxiety. In telemedicine, CAD-Net can assist clinicians during remote consultations by tracking changes in patient facial expressions over time, even when masks or medical equipment partially occlude the face.

Integrated into wearable or ambient sensing devices, CAD-Net could enable emotion-aware assistive systems that adapt their behaviour based on user state—for example, adjusting dialog strategies in socially assistive robots or tailoring feedback in digital therapeutics. Because the architecture is designed to maintain stable performance under partial visibility (e.g., masks, hair, hands), it is particularly suited for real-world clinical and home-care environments where occlusions are common. Beyond healthcare, similar capabilities are relevant for education, driver monitoring, and affective user interfaces, where robust and explainable facial analysis is desirable.

Beyond these direct applications, robust landmark geometry can also provide useful structural support for downstream facial analysis tasks in which occlusion, pose variation, and motion corruption are equally critical. In face anti-spoofing, stable geometric consistency can complement appearance-based cues by helping distinguish genuine facial dynamics from presentation attacks. In remote photoplethysmography estimation, reliable facial region localization and stable landmark tracking can improve region-of-interest selection and reduce motion-induced contamination.

We expand the discussion to highlight how robust landmark geometry can benefit downstream tasks where occlusions and pose variations are also critical. For example, in face anti-spoofing, stable geometry can help disentangle genuine facial motion patterns from presentation attacks, complementing appearance cues Huang et al. (2025b,c). In rPPG estimation, reliable facial region localization and stable landmark tracking can improve ROI selection and reduce motion/occlusion artifacts, which directly impacts physiological signal quality Huang et al. (2025a, 2026).

At the same time, the practical significance of CAD-Net should be interpreted with appropriate scope. The present work validates the proposed framework on standard public benchmarks to establish methodological robustness and cross-dataset generalization, but domain-specific deployment will still require task-specific validation and potentially additional data adaptation.

We also add a limitations paragraph: (i) extreme full-face occlusions (e.g., large scarves or face shields) can remove most geometric evidence, in which case performance degrades; (ii) strong out-of-distribution demographics or capture setups can still cause domain shift; and (iii) clinical deployment may require institution-specific validation and calibration. These limitations motivate future work on domain-adaptive training, uncertainty estimation, and the collection of application-specific data.

These limitations are especially relevant in healthcare settings, where patient populations, camera placement, lighting conditions, and privacy constraints may differ substantially from public benchmark conditions. For this reason, CAD-Net should be viewed as a strong foundational model whose clinical utility will depend on careful validation under institution-specific workflows.

9 Conclusion and Future Work

We have presented **CAD-Net**, a Context-Aware Dropout-based Occlusion-Adaptive Network for robust facial landmark detection and emotion recognition. CAD-Net combines a deep geometry-aware block, an attentive dropout block, and a low-rank learning block within a unified, end-to-end trainable architecture. The geometry-aware block captures long-range structural dependencies, the attentive dropout block selectively suppresses occluded and noisy regions, and the low-rank block yields a compact and stable regression head. Extensive experiments on several challenging FLD and FER benchmarks demonstrate that CAD-Net achieves competitive or superior performance compared with recent occlusion-aware methods, particularly under severe occlusions, large pose variations, and cross-dataset evaluation.

To strengthen deployment claims, we include a dedicated efficiency analysis reporting parameter count, FLOPs, memory footprint, and inference speed (FPS) on both server-class and affordable GPU hardware, and we compare these metrics against representative baselines under identical input resolution. We also report the effect of enabling/disabling RCCA iterations (R and varying the rank r on runtime, making the accuracy–efficiency trade-off explicit.

These additional analyses indicate that the improved robustness of CAD-Net is achieved with moderate computational overhead relative to the lightweight baseline setting, supporting its suitability for time-sensitive applications where both accuracy and efficiency matter.

In future work, we plan to extend CAD-Net towards more fine-grained affective analysis, including the detection and monitoring of depressive and anxiety-related facial cues in clinical settings. We also intend to explore more efficient parallel implementations and model compression techniques to facilitate deployment on edge devices and embedded platforms. Another promising direction is the incorporation of 3D facial geometry and multi-view information to enhance depth-aware understanding, as well as the integration of multimodal signals (e.g., speech, physiology) to further improve robustness and interpretability in real-world emotion-aware systems.

In particular, recent progress in model compression suggests promising directions for reducing deployment cost without substantially sacrificing accuracy. Structured and unstructured pruning, as well as tensor decomposition-based compression, may provide practical means of adapting CAD-Net to embedded and wearable platforms.

We strengthen the future-work discussion by citing representative recent compression directions, including structured/unstructured pruning and tensor decomposition-based model compression, and clarifying that domain-specific clinical deployment will require institution-specific data governance, privacy protection, and validation protocols.

More broadly, future clinical deployment will require not only computational adaptation but also rigorous governance. Institution-specific validation, privacy-preserving data handling, calibration across acquisition devices, and continuous monitoring of demographic fairness will be necessary before reliable real-world translation can be claimed.

Declarations

Ethics:This article does not contain any studies involving animals performed by any of the authors.

References

Browatzki B, Wallraven C (2020) 3fabrec: Fast few-shot face alignment by reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6110–6120

Burgos-Artizzu XP, Perona P, Dollár P (2013) Robust face landmark estimation under occlusion. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1513–1520

Choe J, Lee S, Shim H (2020) Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. IEEE transactions on pattern analysis and machine intelligence 43(12):4256–4271

Deng J, Dong W, Socher R, et al (2009) Imagenet: A large-scale hierarchical image database. In: Proceedings of the conference on computer vision and pattern recognition, Ieee, pp 248–255

Deng J, Guo J, Ververas E, et al (2020) Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5203–5212

Ekman P, Friesen W (1978) Facial Action Coding System: Investigators Guide. Consulting Psychologists Press

Fu G, Yu Y, Ye J, et al (2023) A method for diagnosing depression: Facial expression mimicry is evaluated by facial expression recognition. Journal of affective disorders 323:809–818

Gao P, Lu K, Xue J, et al (2020) A coarse-to-fine facial landmark detection method based on self-attention mechanism. IEEE Transactions on Multimedia

Ghiasi G, Fowlkes CC (2014) Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2385–2392

Hager JC, Ekman P, Friesen WV (2002) Facial action coding system. Salt Lake City, UT: A Human Face p 8

Hannane R, Elboushaki A, Afdel K (2020) A divide-and-conquer strategy for facial landmark detection using dual-task cnn architecture. Pattern Recognition p 107504

Happy S, Patnaik P, Routray A, et al (2015) The indian spontaneous expression database for emotion recognition. IEEE Transactions on Affective Computing https://doi.org/10.1109/TAFFC.2015.2498174

Hasani B, Mahoor M (2017) Facial expression recognition using enhanced deep 3d convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), https://doi.org/10.1109/CVPRW.2017.282

Huang PK, Chen TH, Chan YT, et al (2025a) Dd-rppgnet: de-interfering and descriptive feature learning for unsupervised rppg estimation. IEEE Transactions on Information Forensics and Security

Huang PK, Chong JX, Chiang CH, et al (2025b) Slip: Spoof-aware one-class face anti-spoofing with language image pretraining. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 3697–3706

Huang PK, Chong JX, Hsu MT, et al (2025c) Channel difference transformer for face anti-spoofing. Information Sciences 702:121904

Huang PK, Chen TH, Chan YT, et al (2026) Fully test-time rppg estimation via synthetic signal-guided feature learning. Pattern Recognition 170:112102

Huang Z, Wang X, Huang L, et al (2019) Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 603–612

Jaiswal S, Martinez B, Valstar M (2015) Learning to combine local models for facial action unit detection. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, https://doi.org/10.1109/FG.2015.7284872

Kennedy B, Balint A (2016) Emotionnet2. https://github.com/co60ca/EmotionNet

Koestinger M, Wohlhart P, Roth PM, et al (2011) Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops), IEEE, pp 2144–2151

Kowalski M, Naruniec J, Trzcinski T (2017) Deep alignment network: A convolutional neural network for robust face alignment. In: CVPRW, https://doi.org/10.1109/CVPRW.2017.254

Liedtke C, Kohl W, Kret ME, et al (2018) Emotion recognition from faces with in-and out-group features in patients with depression. Journal of Affective Disorders 227:817–823

Liu C, Hirota K, Dai Y (2023) Patch attention convolutional vision transformer for facial expression recognition with occlusion. Information Sciences 619:781–794

Liu J, Pang Y, Wang S (2025) Dce-net: A dual-frequency domain knowledge-guided framework for image dehazing via detail and content enhancements. IEEE Signal Processing Letters

Liu Q, Deng J, Yang J, et al (2016) Adaptive cascade regression model for robust face alignment. IEEE Transactions on Image Processing 26(2):797–807

Lopes A, de Aguiar E, Oliveira-Santos T (2015) A facial expression recognition system using convolutional networks. In: SIBGRAPI, https://doi.org/10.1109/SIBGRAPI.2015.14

Lucey P, Cohn J, Kanade T, et al (2010) The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: CVPRW, https://doi.org/10.1109/CVPRW.2010.5543262

Lyons M, Akamatsu S, Kamachi M, et al (1998) The japanese female facial expressions database. http://www.kasrl.org/jaffe.html

Mollahosseini A, Hasani B, Mahoor M (2017) Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing https://doi.org/10.1109/TAFFC.2017.2740923

Onishi A (2021) Brain-computer interface with rapid serial multimodal presentation using artificial facial images and voice. Computers in Biology and Medicine 136:104685

Peng X, Gu Y, Zhang P (2022) Au-guided unsupervised domain-adaptive facial expression recognition. Applied Sciences 12(9):4366

Sadiq M, Shi D (2022) Attentive occlusion-adaptive deep network for facial landmark detection. Pattern Recognition 125:108510

Sadiq M, Shi D, Guo M, et al (2019) Facial landmark detection via attention-adaptive deep network. IEEE Access 7:181041–181050

Sadiq M, Shi D, Liang J (2022) A robust occlusion-adaptive attention-based deep network for facial landmark detection. Applied Intelligence 52(8):9320–9333

Sadiq M, Liang J, Yu G, et al (2024) Facial landmark detection: An attentive dropout-based occlusion-adaptive deep network. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining Workshops, Springer

Sagonas C, Tzimiropoulos G, Zafeiriou S, et al (2013) 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 397–403

Shao Z, Liu Z, Cai J, et al (2018a) Deep adaptive attention for joint facial action unit detection and face alignment. In: Proceedings of the European conference on computer vision (ECCV), pp 705–720

Shao Z, Liu Z, Cai J, et al (2018b) Facial action unit detection using attention and relation learning. CoRR

Szegedy C, Liu W, Jia Y, et al (2015) Going deeper with convolutions. In: CVPR, https://doi.org/10.1109/CVPR.2015.7298594

Tautkutė I, Trzciński T (2019) Classifying and visualizing emotions with emotional dan. Fundamenta Informaticae 168(2-4):269–285

Tian Y, Kanade T, Cohn J (2001) Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence https://doi.org/10.1109/34.908962

Tzimiropoulos G (2015) Project-out cascaded regression with an application to face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3659–3667

Wan J, Liu J, Zhou J, et al (2023) Precise facial landmark detection by reference heatmap transformer. IEEE Transactions on Image Processing 32:1966–1977

Wang S, Hou Q, Li J, et al (2025a) Tsid-net: a two-stage single image dehazing framework with style transfer and contrastive knowledge transfer. The Visual Computer 41(3):1921–1938

Wang S, Ren W, Gao P, et al (2025b) Zrid-net: Zero-reference real-world image dehazing framework via deep self-decoupling and reverse knowledge transfer. IEEE Transactions on Circuits and Systems for Video Technology

Wu Y, Ji Q (2019) Facial landmark detection: A literature survey. International Journal of Computer Vision 127(2):115–142

Xia X, Xu C, Nan B (2017) Facial expression recognition based on tensorflow platform. In: ITM Web of Conferences, https://doi.org/10.1051/itmconf/20171201005

Xiang CY, He JY, Cheng ZQ, et al (2025) Popos: Improving efficient and robust facial landmark detection with parallel optimal position search. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 8602–8610

Xie Y, Chen T, Pu T, et al (2020) Adversarial graph representation adaptation for cross-domain facial expression recognition. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 1–24

Xing J, Niu Z, Huang J, et al (2017) Towards robust and accurate multi-view and partially-occluded face alignment. IEEE transactions on pattern analysis and machine intelligence 40(4):987–1001

Yildirim-Celik H, Eroglu S, Oguz K, et al (2022) Emotional context effect on recognition of varying facial emotion expression intensities in depression. Journal of Affective Disorders 308:141–146

Zafeiriou S, Trigeorgis G, Chrysos G, et al (2017) The menpo facial landmark localisation challenge: A step towards the solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 170–179

Zhao K, Chu W, Torre F, et al (2015) Joint patch and multi-label learning for facial action unit detection. TIP https://doi.org/10.1109/TIP.2016.2570550

Zhu M, Shi D, Zheng M, et al (2019) Robust facial landmark detection via occlusion-adaptive deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3486–3496

Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE conference on computer vision and pattern recognition, IEEE, pp 2879–2886