IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 March 2026, accepted 1 April 2026, date of publication 6 April 2026, date of current version 9 April 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3681267

RESEARCH ARTICLE

ADODN: Attentive Dropout-Based Occlusion-Aware Deep Network for Facial Landmark Detection

MUHAMMAD SADIQ[1], JUNHAO WU[2], YU GENG[1], MOHAMMAD SULTAN MAHMUD[3], (Senior Member, IEEE), AMAR KHELLOUFI[1], HUA ZHENG[4], YUNSHENG ZHANG[1], AND JUNWEI LIANG[1], (Member, IEEE)

[1]Shenzhen University of Information Technology, Shenzhen 518172, China
[2]Department of Computer Science, College of Mathematics and Computer Science, Shantou University, Shantou 515063, China
[3]School of Artificial Intelligence, Shenzhen Technology University, Shenzhen 518118, China
[4]Guangzhou University of Software, Guangzhou 510990, China

Corresponding authors: Muhammad Sadiq (sadiq@suit-sz.edu.cn) and Yunsheng Zhang (zhangys@suit-sz.edu.cn)

This work was supported in part by the Science and Technology Ph.D. Research Startup Project under Grant SZIIT2023KJ016; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515110070; in part by the International Cooperation Research Project of Shenzhen Science and Technology Innovation Commission under Grant GJHZ20220913143013024; in part by the ''Chunhui Plan'' Project of the Chinese Service Center for Scholarly Exchange, Ministry of Education, under Grant HZKY20220106; in part by Guangdong Provincial Natural Science Foundation under Grant 2026A1515011449; in part by Shenzhen Science and Technology Program under Grant RCBS20221008093252092 and Grant 20220820003203001; and in part by Guangdong Basic and Applied Basic Research.

ABSTRACT This paper introduces an attentive dropout-based occlusion-adaptive deep network (ADODN) for robust facial landmark detection under challenging conditions, including occlusions, extreme poses, and illumination variations. While convolutional neural networks have achieved high accuracy in Facial Landmark Detection (FLD), their performance often degrades under partial occlusions due to over-reliance on a few highly discriminative facial regions. ADODN addresses this limitation through three complementary modules: 1) a geometry-aware module that captures spatial relationships and structural priors among facial components; 2) an attentive dropout module that stochastically alternates drop masks and importance maps to encourage balanced feature learning from both dominant and subtle facial cues; and 3) a low-rank learning module that regularizes the regression representation by exploiting inter-feature correlations to recover occlusion-missing information. Unlike deterministic reweighting schemes, the attentive dropout mechanism improves robustness by randomly suppressing prominent responses during training, which mitigates feature over-dependence and promotes holistic structural inference. The resulting framework remains end-to-end and does not require auxiliary classifiers or multi-stage training. Extensive evaluations on challenging benchmarks (300W, COFW) show that ADODN achieves competitive and consistent performance, especially under occlusion-heavy settings. For example, ADODN attains 2.80 NRMSE on the 300W Common set and 5.81 on the Challenging set, improving upon recent baselines including ODN, AODN, and RHT-R. We also report parameter efficiency relative to prior ODN-style designs, supporting efficient training and inference.

INDEX TERMS Attention, facial landmark detection, occlusion, attentive dropout, low-rank regularization.

I. INTRODUCTION

Facial Landmark Detection (FLD) aims to precisely localize key points on facial components such as eyes, nose, mouth, and chin, and it serves as a fundamental step in facial analysis

The associate editor coordinating the review of this manuscript and approving it for publication was Katherine VanDenburgh.

applications including face recognition, expression analysis, and 3D face modeling [1]. Despite substantial progress, accurate landmark localization remains challenging under real-world conditions involving occlusions, extreme poses, and illumination variations [2].

Traditional FLD approaches can be broadly categorized into template-based and regression-based methods. Template-based approaches [3], [4], [5], [6] employ statistical models such as Active Shape Models (ASM) and Active Appearance Models (AAM) to represent facial structure, but they often fail under severe occlusions due to strong global-shape assumptions. Regression-based methods [7], [8], [9] learn mappings from image features to landmark coordinates, yet remain sensitive to spatial distortions and partially missing facial evidence.

The advent of deep learning, particularly Convolutional Neural Networks (CNNs), has substantially improved FLD performance [10]. However, many CNN-based FLD pipelines still exhibit a critical limitation: they tend to over-emphasize a few highly discriminative facial regions, and performance degrades when these regions are partially occluded. Attention-based mechanisms [11], [12], [13], [14] alleviate this issue by focusing on informative regions, but they are often deterministic reweighting schemes and may not sufficiently encourage learning from subtle but structurally important cues. Moreover, erasing-style designs may introduce additional training complexity (e.g., auxiliary branches or staged optimization), which can affect reproducibility.

To address these limitations, we propose **A**ttentive **D**ropout-based **O**cclusion-adaptive **D**eep **N**etwork (ADODN). ADODN builds upon the ODN backbone [10] but introduces a different occlusion-learning mechanism: a stochastic attentive dropout strategy that alternates between drop masks and attention-derived importance maps. This stochastic suppression mitigates feature over-dependence and encourages holistic structural inference under unpredictable occlusion patterns. As illustrated in Fig. 1, real-world occlusions exhibit diverse and irregular patterns that require adaptive handling.

ADODN integrates three modules in a single end-to-end pipeline. First, a geometry-aware module encodes structural priors and spatial relations among facial components. Second, an attentive dropout module stochastically balances learning from dominant and non-dominant cues via drop masks and importance maps. Third, a low-rank learning module regularizes the regression representation by exploiting inter-feature correlations to recover occlusion-missing information. We further report ablations and a brief sensitivity/stability analysis in Sec. V to support reproducibility.

Extensive evaluations on challenging benchmarks show that ADODN achieves competitive and consistent performance, especially on occlusion-heavy settings. For example, ADODN attains NRMSE scores of 2.80 on the 300W Common set and 5.81 on the Challenging set, improving upon recent baselines including ODN, AODN, and RHT-R. We additionally report parameter efficiency relative to prior ODN-style designs to support efficient training and inference.

The remainder of this paper is organized as follows: Section II provides grouped background and related work. Section III details the proposed ADODN architecture. Section IV presents the mathematical optimization framework. Section V provides experimental validation with ablation and discussion, and Section VI concludes with key findings and future directions.

TABLE 1. Main notation used in ADODN.

Symbol	Meaning
I	Input face image
$\mathcal{Z}$	Backbone feature tensor (e.g., $\mathcal{Z} \in \mathbb{R}^{C\times H\times W}$)
$\mathcal{A}$	Attention / importance response map
$\mathcal{M}_d$	Attentive dropout mask (dropping dominant responses)
$\mathcal{M}_i$	Importance-guided mask (highlighting informative responses)
$\mathcal{M}$	Final mask applied to features
$\odot$	Element-wise (Hadamard) product
$\|\cdot\|_*$	Nuclear norm (sum of singular values)
$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$	SVD of a matrix

II. RELATED BACKGROUND AND CONTEXT

This section reviews closely related work and positions ADODN with respect to existing occlusion-robust facial landmark detection methods. For clarity and reproducibility, we also summarize the main symbols and operators used throughout the paper in Sec. II-A.

A. NOTATION AND OPERATORS

Table 1 summarizes the key symbols used throughout the paper. We denote vectors and matrices in bold lowercase and uppercase letters, respectively. Feature tensors are denoted by calligraphic letters. The Hadamard product is denoted by $\odot$.

B. RELATED WORK AND POSITIONING

Facial landmark detection serves as a cornerstone technology in computer vision, enabling applications including facial recognition, expression analysis, affective computing, and human–computer interaction systems. The primary objective of FLD is to localize predefined anatomical keypoints (typically 68 or 29 landmarks) that define facial components such as eyes, nose, mouth, and contours [2].

Despite substantial advances in deep learning-based approaches, occlusion remains one of the most challenging factors in real-world FLD deployment. Occlusions manifest as *external* occlusions caused by accessories (glasses, masks) and objects (hands, hair), and as *internal* occlusions resulting from extreme poses, expressions, or self-occlusion. Because occlusions are irregular and unpredictable, robust FLD requires learning strategies that do not collapse when a small set of dominant cues becomes unavailable. As illustrated in Fig. 1, these effects can significantly degrade landmark localization accuracy.

FIGURE 1. Representative occlusion examples from the COFW dataset [15], demonstrating diverse occlusion patterns including hair, hands, accessories, and external objects. These cases highlight the need for robust occlusion-adaptive learning in facial landmark detection.

1) EXPLICIT OCCLUSION/VISIBILITY MODELING

Early methods addressed occlusion by explicitly estimating visibility or occlusion patterns and using them to guide regression. Wu et al. [2], [16] developed a supervised regression framework that updates landmark visibility probabilities, while Xing et al. [17] introduced occlusion dictionaries to represent common occlusion patterns. Liu et al. [18] combined regression with statistical shape constraints to infer missing landmarks via geometric consistency. These works highlight the importance of coupling occlusion reasoning with structural priors.

2) CNN-BASED FLD AND OCCLUSION-AWARE MODULES

CNNs enabled end-to-end learning of appearance–shape relationships, but standard CNN features remain vulnerable to occlusion-induced corruption. Zhu et al. proposed the Occlusion-adaptive Deep Network (ODN) [10], which introduced an occlusion estimation pathway and a low-rank learning module for feature recovery. Building on ODN, AODN [12] incorporated attention mechanisms to emphasize informative regions, improving robustness under partial occlusion. However, both ODN and AODN mainly rely on deterministic reweighting, which can still over-concentrate learning on highly discriminative regions and leave the regressor under-trained on subtle but structurally important cues.

3) ERASING/DROPOUT-STYLE ROBUSTNESS AND PRACTICALITY

Feature erasing and attention-guided dropping [19], [20] encourage exploration of alternative evidence, but some designs rely on auxiliary classifiers or staged optimization, which can increase training complexity and affect reproducibility. This motivates designing occlusion-robust learning strategies that remain end-to-end and lightweight.

4) ROBUSTNESS UNDER REALISTIC CORRUPTION PIPELINES (CONTEXT)

Some robustness studies focus on real-world distortion pipelines (e.g., compression, re-recording/screen-shooting) that can indirectly affect downstream face analysis. For example, motion cues have been exploited to detect compressed deepfake videos [21], and screen-shooting resistant watermarking has been studied under severe distortion-time distortions [22], [23], [24]. While these works do not perform landmark localization, they reinforce the importance of robustness-aware modeling under practical corruption processes. In parallel, structured regularization has been explored via graph-based objectives such as aligning feature- and prediction-induced graphs [25]; in contrast, ADODN targets occlusion-robust landmark regression by combining intra-face geometry-aware context, stochastic attentive dropout, and low-rank regularization in a single end-to-end FLD pipeline.

5) MOTIVATION AND POSITIONING OF ADODN

The key insight motivating ADODN is that effective occlusion handling requires balanced feature learning across both dominant and subtle facial regions while maintaining computational efficiency. Rather than treating occlusion only as a binary visibility problem, ADODN treats it as a feature-learning optimization issue: the regressor should remain effective when dominant regions are suppressed or missing. Accordingly, ADODN integrates a geometry-aware module for structural context, a stochastic attentive dropout module that alternates between dropping and highlighting responses, and a low-rank learning module for correlation-driven recovery.

As illustrated in Fig. 2, ADODN maintains end-to-end training and avoids auxiliary classifiers or multi-stage optimization. Compared with ODN/AODN, the main difference is the stochastic attentive dropout mechanism, which reduces persistent reliance on the same salient regions and encourages more complete structural utilization under occlusion. We validate this design using ablations, sensitivity analysis, and distribution-level evaluation (CED curves) in the experimental section.

III. DROPOUT-BASED OCCLUSION-AWARE DEEP NETWORK (ADODN)

This section presents the architectural details of the proposed Attentive Dropout-based Occlusion-adaptive Deep Network

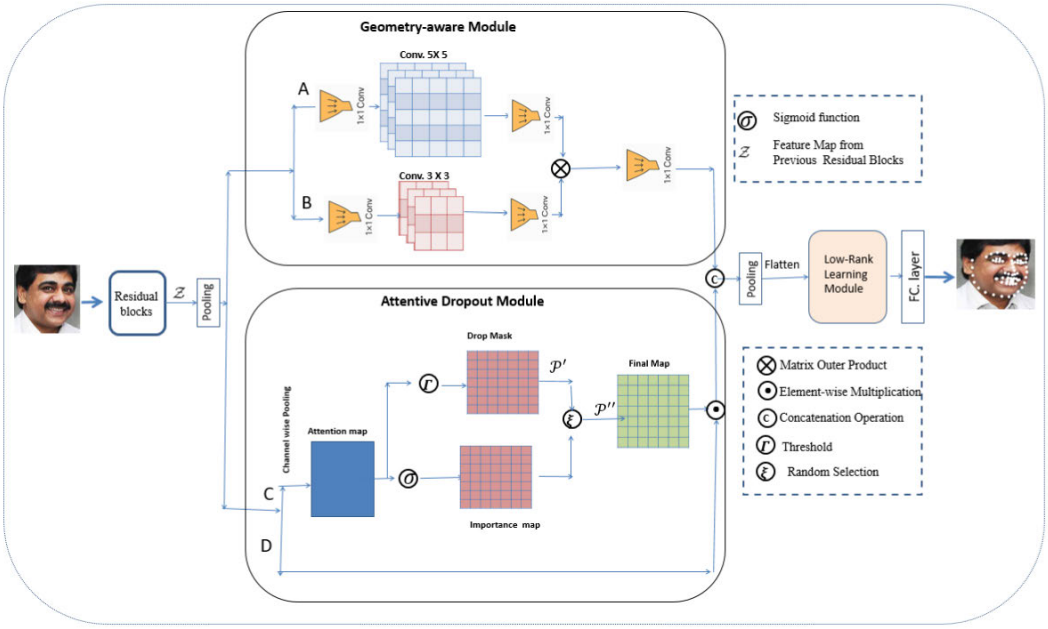

FIGURE 2. Architecture of the proposed Attentive Dropout-based Occlusion-adaptive Deep Network (ADODN), illustrating the integration of geometry-aware, attentive dropout, and low-rank learning modules.

(ADODN). As illustrated in Fig. 2, ADODN builds upon a ResNet-18 backbone and integrates three modules that work synergistically to improve robustness under occlusion.

A. OVERALL ARCHITECTURE

ADODN processes an input image I_i through a modified ResNet-18 backbone, where the final residual unit is replaced by the proposed occlusion-adaptive framework. The backbone produces feature maps $\mathcal{Z} \in \mathbb{R}^{H\times W\times C}$, which are fed into two parallel branches: a geometry-aware branch and an attentive dropout branch. The geometry-aware branch encodes structural relations among facial components, while the attentive dropout branch stochastically modulates feature responses to prevent persistent over-reliance on a few dominant regions. The outputs from both branches are concatenated and passed to the low-rank learning module, which regularizes the regression representation by exploiting inter-feature correlations for occlusion-missing information recovery. Finally, a fully connected regression layer outputs landmark coordinates.

B. GEOMETRY-AWARE MODULE

Inspired by dorsal-stream-like spatial processing, the Geometry-Aware Module captures long-range dependencies and structural relationships among facial components. Conventional convolutions operate with local receptive fields, which can be insufficient for modeling global geometric constraints when facial evidence is partially missing. To address this, we explicitly model pairwise spatial interactions to encode symmetry, proximity, and structural priors.

The module contains two pathways (Pathway-A and Pathway-B). Each pathway is preceded and followed by a 1×1 convolution to increase nonlinearity without expanding receptive fields. Pathway-A uses a 3×3 convolution and Pathway-B uses a 5×5 convolution for multi-scale feature extraction. Geometric relations are encoded through a channel-wise outer product:

$$\mathcal{G} = \mathrm{Conv}_{1\times1}(\mathcal{F}_A \otimes \mathcal{F}_B), \tag{1}$$

where $\mathcal{F}_A$ and $\mathcal{F}_B$ are features from Pathway-A and Pathway-B, respectively, $\otimes$ denotes the channel-wise outer product, and $\mathcal{G}$ is the resulting geometric feature representation.

Computational complexity. Let $S = HW$ be the number of spatial positions. The outer-product interaction has time complexity $\mathcal{O}(C_g S^2)$ and memory complexity $\mathcal{O}(S^2)$, where C_g is the effective channel width used in the geometry-aware block. In our implementation, this module is placed at the late stage of the backbone (small H, W) and uses reduced channel width, making the overhead tractable while providing global structural context.

C. ATTENTIVE DROPOUT MODULE

The Attentive Dropout Module addresses a key limitation of deterministic attention: when attention repeatedly amplifies a small set of salient regions, the regressor may under-utilize subtle but structurally important cues, and performance degrades when the salient regions are occluded. We therefore introduce a stochastic exploration–exploitation mechanism that alternates between dropping highly activated regions and preserving discriminative cues.

The module contains two complementary sub-networks. Pathway-C is a lightweight bottleneck structure (1×1–3×3–1×1) with reduced channels to estimate occlusion-sensitive responses, while Pathway-D is a residual branch that stabilizes feature propagation. This design remains end-to-end and avoids auxiliary classifiers or multi-stage optimization.

1) ATTENTION MAP

Given input features $\mathcal{Z}$, we compute a single-channel attention map $\mathcal{A} \in \mathbb{R}^{H \times W}$ using channel-wise average pooling:

$$\mathcal{A} = \frac{1}{C} \sum_{c=1}^{C} |\mathcal{Z}_{:,:,c}|. \tag{2}$$

This map estimates spatial response intensity and is used to construct both dropping and highlighting masks.

2) DUAL STOCHASTIC STRATEGY

We construct two masks from $\mathcal{A}$ and apply one of them at each training iteration. The drop mask is

$$\mathcal{M}_d = \mathbb{I}(\mathcal{A} > \tau_d),$$

which suppresses highly activated regions to encourage learning from alternative evidence. The importance map is

$$\mathcal{M}_i = \sigma(\mathcal{A}),$$

which preserves discriminative responses through a smooth sigmoid mapping. Unlike deterministic reweighting, this stochastic switching reduces persistent dependence on the same salient region and improves robustness under unpredictable occlusions.

At each iteration, we sample one mechanism with probability p:

$$\mathcal{F}_{out} = \begin{cases} \mathcal{F}_{in} \odot \mathcal{M}_d, & \text{with probability } p, \\ \mathcal{F}_{in} \odot \mathcal{M}_i, & \text{with probability } 1-p, \end{cases} \tag{3}$$

where $\odot$ denotes element-wise multiplication. Unless otherwise stated, we use $p = 0.5$ as a balanced default, and we report a sensitivity study for p and τ_d in Sec. V. To mitigate early-stage instability, we enable the drop-mask branch after a short warm-up period. (See Sec. V.)

3) REGULARIZATION

We apply L_1 regularization to the single-channel maps $\mathcal{P}'$ and $\mathcal{P}''$:

$$\mathcal{L}_{reg} = \eta' \|\mathcal{P}'\|_1 + \eta'' \|\mathcal{P}''\|_1, \tag{4}$$

with $\eta' = \eta'' = 1 \times 10^{-6}$ in our experiments.

D. LOW-RANK LEARNING MODULE

The Low-Rank Learning Module addresses feature incompleteness caused by occlusion and stochastic dropping. We treat the low-rank constraint as a principled regularizer that encourages correlated facial attributes to share compact structure, which helps recover missing evidence for landmark regression. Given the concatenated feature vector $\mathcal{X}$, we optimize:

$$\min_{\mathcal{M}, W_{fc}} \frac{1}{N} \sum_{i=1}^{N} \|\breve{S}_i - W_{fc}^T \mathcal{M}^T \mathcal{X}_i\|_F^2 + \beta \|\mathcal{M}\|_* + \gamma \|\mathcal{M}\|_F^2, \tag{5}$$

where $\breve{S}_i$ denotes ground-truth landmarks, W_{fc} denotes regression-head parameters, $\|\cdot\|_*$ is the nuclear norm enforcing low rank, and β, γ are regularization coefficients. The optimization and the nuclear-norm subgradient used in backpropagation are given in Sec. IV.

E. MODULE INTEGRATION AND BIOLOGICAL INSPIRATION

ADODN is inspired by the dual-stream theory of human visual processing [26]. The geometry-aware module corresponds to dorsal-stream-like spatial reasoning, the attentive dropout module corresponds to ventral-stream-like appearance processing, and the low-rank learning module integrates these cues for robust regression. This analogy is used only to motivate the design intuition; the technical contribution is the explicit integration of geometry-aware context with stochastic attentive dropout and low-rank regularization in a single end-to-end FLD pipeline.

All modules are trained end-to-end using backpropagation, and the total loss comprises landmark regression and regularization terms. The optimization details are elaborated in Section IV.

IV. OPTIMIZATION OF PROPOSED METHODOLOGY

This section details the mathematical formulation and optimization strategy for training the proposed ADODN framework. We refine the formulation to improve rigor, particularly for the nuclear-norm term, and we present the corresponding subgradient computations used in end-to-end backpropagation.

A. OBJECTIVE FUNCTION FORMULATION

Overall objective. Given training samples $\{(I_i, \breve{S}_i)\}_{i=1}^N$, we minimize an empirical regression risk together with structured regularization that encourages sparsity and low-rank correlation. The complete objective is:

$$\begin{aligned} \min_{\mathcal{W}_c, \mathcal{W}_{fc}, \mathcal{M}} & \frac{1}{N} \sum_{i=1}^{N} \left(\left\| \breve{S}_i - S_i \right\|_F^2 + \eta' \left\| \mathcal{P}_i' \right\|_1 + \eta'' \left\| \mathcal{P}_i'' \right\|_1 \right) \\ & + \beta \|\mathcal{M}\|_* + \gamma \|\mathcal{M}\|_F^2 \\ & + \alpha \|\mathcal{W}_c\|_F^2 + \lambda \left\| \mathcal{W}_{fc} \right\|_F^2, \end{aligned} \tag{6}$$

where $S_i = \mathcal{F}_{\text{ADODN}}(I_i; \mathcal{W}_c, \mathcal{W}_{fc}, \mathcal{M})$ denotes the landmark predictions generated by the network, and $\breve{S}_i = \{s_1, s_2, \ldots, s_L\}$ denotes the ground-truth landmarks with L points per face. Compared with the earlier draft, the

sparsity terms are explicitly included inside the sample-wise summation to maintain a consistent empirical-risk form.

B. LOSS TERM INTERPRETATION

The first term in Equation 6 is the landmark regression loss, which penalizes the Frobenius distance between predicted and ground-truth landmark coordinates. The attentive dropout module produces single-channel maps $\mathcal{P}_i'$ and $\mathcal{P}_i''$, and the L_1 penalties encourage sparsity to prevent diffuse activation. The structural matrix $\mathcal{M}$ is regularized by the nuclear norm $\|\mathcal{M}\|_*$ (promoting low rank) together with $\|\mathcal{M}\|_F^2$ (stabilizing optimization). Finally, $\|\mathcal{W}_c\|_F^2$ and $\|\mathcal{W}_{fc}\|_F^2$ apply standard weight decay to convolutional and regression-head parameters.

C. GRADIENT COMPUTATION AND OPTIMIZATION

Because the nuclear norm and L_1 norm are non-smooth, ADODN is optimized using subgradient-based backpropagation.

1) NUCLEAR NORM SUBGRADIENT

Let the singular value decomposition be $\mathcal{M} = U\Sigma V^T$, and let $r = \text{rank}(\mathcal{M})$ with U_r and V_r denoting the singular vectors corresponding to the non-zero singular values. A commonly used subgradient is:

$$G_* \in \partial\|\mathcal{M}\|_* \quad \text{with a common choice} \quad G_* = U_r V_r^T. \tag{7}$$

More generally, the subdifferential set satisfies:

$$\partial\|\mathcal{M}\|_* = \left\{ U_r V_r^T + W \;:\; U_r^T W = 0,\; WV_r = 0, \right. \\ \left. \|W\|_2 \le 1 \right\}. \tag{8}$$

In implementation, we use the SVD-based form in Equation 7 (handled by automatic differentiation) to backpropagate through the nuclear-norm regularizer.

2) L_1 NORM SUBGRADIENT

For the sparsity-inducing L_1 term, the element-wise subgradient is:

$$\frac{\partial\,\|\mathcal{P}\|_1}{\partial P_k} = \begin{cases} +1 & P_k > 0, \\ -1 & P_k < 0, \\ [-1, +1] & P_k = 0, \end{cases} \tag{9}$$

where P_k is the k-th element of $\mathcal{P} \in \{\mathcal{P}', \mathcal{P}''\}$.

D. END-TO-END TRAINING PROCEDURE

All modules form a directed acyclic computation graph, allowing gradients from Equation 6 to flow through the geometry-aware, attentive dropout, and low-rank learning components. We update $\{\mathcal{W}_c, \mathcal{W}_{fc}, \mathcal{M}\}$ jointly using Adam. Unless otherwise stated, we train with batch size 32 and an initial learning rate 1×10^{-3}, decayed by a factor of 10 when validation error plateaus, with early stopping based on validation performance. This optimization setup preserves end-to-end differentiability while ensuring the low-rank and sparsity regularizers are treated with theoretically correct subgradient computations.

V. EXPERIMENTAL DETAILS

This section presents an evaluation of the proposed ADODN framework across benchmark datasets and experimental settings. We report results under standard conditions and under occlusions/large pose variations, and we include ablation, sensitivity, and stability analyses to support reproducibility.

A. DATASETS AND EVALUATION PROTOCOL

We conduct experiments on two widely adopted facial landmark detection benchmarks.

1) 300W DATASET

The 300W dataset [27] comprises 3,837 facial images annotated with 68 landmarks, compiled from AFW, HELEN, and LFPW. We follow the standard protocol using 3,148 images for training and 689 for testing. The test set is partitioned into the Common subset (554 images), the Challenging subset (135 images from IBUG), and the Full set (689 images). We use these splits to separately analyze performance on moderate versus difficult conditions.

2) COFW DATASET

The Caltech Occluded Faces in the Wild (COFW) dataset [15] focuses on occlusion, containing 1,345 training and 507 testing images. We adopt the 68-landmark re-annotation strategy from [28] for consistency with 300W evaluation.

3) CROSS-DATASET EVALUATION

To assess generalization, we employ a cross-dataset protocol where models are trained on 300W and tested on COFW without fine-tuning, following [10], [12]. This setting evaluates robustness to unseen occlusion patterns and dataset bias.

B. IMPLEMENTATION DETAILS

1) TRAINING CONFIGURATION

All models are implemented in PyTorch and trained on NVIDIA RTX 3090 GPUs. We use Adam with initial learning rate 1×10^{-3}, reduced by a factor of 10 when validation loss plateaus. Training employs batch size 32 and early stopping with patience of 50 epochs. To reduce training instability for attentive dropout, we enable the drop-mask branch after a short warm-up period while keeping other settings unchanged across baselines for fairness.

2) DATA AUGMENTATION

We apply standard augmentation including random rotation ($\pm 30°$), scaling ($0.8\times$ to $1.2\times$), translation ($\pm 10\%$), horizontal flipping (50% probability), and color jittering (brightness, contrast, saturation). All images are resized to 224×224 pixels.

3) HYPERPARAMETER SETTINGS

Critical hyperparameters are selected via grid search. We set the reduction ratio $r = 8$ for attention modules, low-rank regularization $\beta = 1 \times 10^{-6}$, weight decay factors $\alpha = \gamma = \lambda = 1 \times 10^{-5}$, and sparsity coefficients $\eta' = \eta'' = 1 \times 10^{-6}$. For attentive dropout, we use $p = 0.5$ as the default switching probability and evaluate sensitivity to p and τ_d in the ablation subsection.

4) EVALUATION METRICS

We use two standard evaluation metrics. To avoid overstating small margins, we interpret mean NRMSE together with distribution-level evidence from CED curves.

(i) Normalized Root Mean Square Error (NRMSE):

$$\text{NRMSE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\check{S}_i - S_i\right\|_2}{L \cdot \Omega_i}, \tag{10}$$

where Ω_i is inter-ocular distance for normalization.

(ii) Cumulative Error Distribution (CED): plots the percentage of test images against normalized error.

TABLE 2. Comparison of NRMSE ($\times 10^{-2}$) on 300W Challenging subset.

Method	Year	NRMSE
ODN [10]	2019	6.67
ADN [29]	2019	6.60
RetinaFace [30]	2020	6.83
AODN [12]	2022	6.38
RHT-R [13]	2023	5.88
ADODN (Ours)	**2026**	**5.81**

TABLE 3. Comparison of NRMSE ($\times 10^{-2}$) on 300W Common and Full sets.

Method	Year	Common	Full
ODN [10]	2019	3.56	4.17
ADN [29]	2019	3.52	4.14
LGSA [11]	2020	3.36	3.94
3FabRec [31]	2020	3.36	3.82
AODN [12]	2022	3.27	3.76
RHT-R [13]	2023	2.87	3.46
POPos [14]	2025	3.06	3.38
ADODN (Ours)	**2026**	**2.80**	**3.35**

Statistical validation. In addition to average NRMSE, we compute per-image NRMSE and apply paired bootstrap confidence intervals and a paired permutation test against the strongest baseline to assess whether small differences are statistically reliable. These tests are paired because all methods are evaluated on the same fixed test set.

C. COMPARATIVE ANALYSIS

1) PERFORMANCE ON 300W DATASET

Table 3 reports NRMSE on the 300W Common and Full sets. ADODN achieves competitive performance and improves upon recent strong baselines on these splits. We rely on both the mean NRMSE and CED curves to characterize overall behavior.

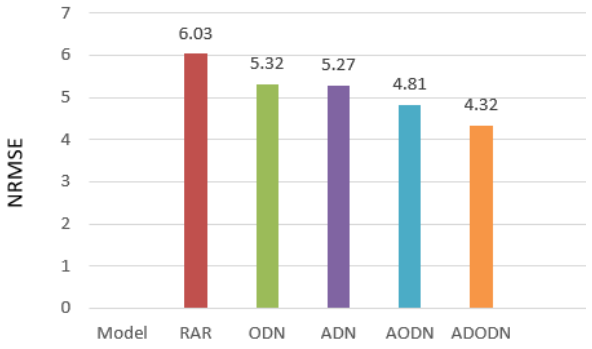

FIGURE 3. Cross-dataset evaluation on COFW (train on 300W, test on COFW without fine-tuning). Lower NRMSE indicates stronger generalization to unseen occlusion patterns.

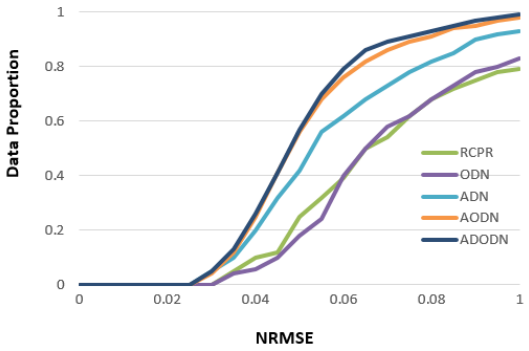

FIGURE 4. CED curve on the 300W Challenging subset. A curve closer to the upper-left indicates that a larger fraction of test images achieves lower normalized error.

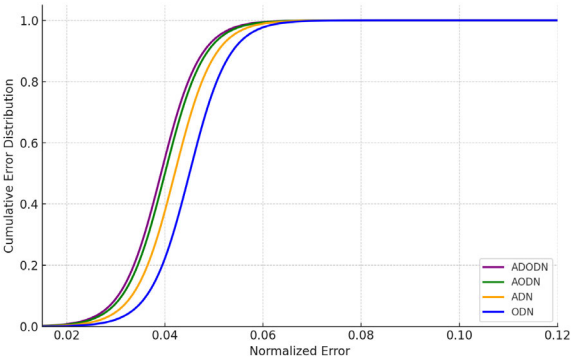

FIGURE 5. CED curve on the 300W Full set. The curve complements mean NRMSE results reported in Table 3.

2) ROBUSTNESS UNDER SEVERE OCCLUSIONS

Table 2 reports results on the 300W Challenging subset. The improvements are modest in absolute value but align with

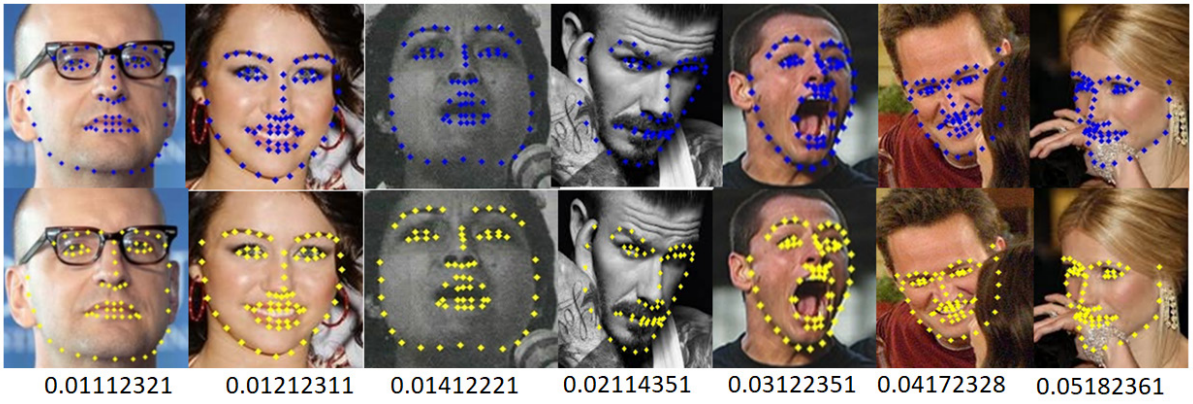

FIGURE 6. Qualitative results on 300W. Ground truth (blue) and predictions (yellow) are shown under occlusions, extreme poses, and lighting variations.

our design goal: improved robustness when discriminative regions are partially unavailable. Figure 4 shows the CED curve; the curve-level improvement indicates that gains are not limited to only a small subset of easy samples.

3) CROSS-DATASET GENERALIZATION

Figure 3 presents cross-dataset results on COFW. When trained on 300W and tested on COFW without fine-tuning, ADODN maintains strong generalization, suggesting improved robustness to unseen occlusion patterns.

D. ABLATION STUDIES

We conduct ablation studies on the 300W Challenging subset to quantify the contribution of each module. Table 4 reports results for different module combinations. Overall, geometry-aware modeling and low-rank regularization improve robustness, and attentive dropout further strengthens performance under occlusion by reducing over-dependence on dominant regions.

Sensitivity and stability of attentive dropout. We vary the switching probability p and the drop threshold τ_d around the default setting to assess sensitivity. Across a reasonable range of values, performance trends remain stable, and we use $p = 0.5$ as a balanced choice. To mitigate instability, we enable the drop-mask branch after a short warm-up and keep the importance-map branch active throughout training.

1) REDUCTION RATIO ANALYSIS

We analyze the channel reduction ratio r in the attention modules. The setting $r = 8$ provides the best balance between efficiency and representation capacity, while larger r values over-compress channels and reduce accuracy.

E. QUALITATIVE ANALYSIS

Figure 6 shows qualitative results on challenging samples. ADODN preserves geometric consistency under large pose variations and remains robust when salient regions are partially occluded, which is consistent with the quantitative improvements on the Challenging split and the distribution-level trends in the CED curves.

TABLE 4. Ablation study on 300W Challenging subset (NRMSE $\times 10^{-2}$).

Configuration	NRMSE
BRNet (ResNet-18)	7.21
+ GM + LM	6.90
+ GM + ADM	6.72
+ ADM + LM	6.68
ADODN (GM + ADM + LM)	**5.81**

VI. CONCLUSION

This paper presented ADODN, an attentive dropout-based occlusion-adaptive deep network for facial landmark detection under occlusions, extreme poses, and illumination variations. The central idea is to improve robustness by encouraging balanced feature learning from both dominant and subtle facial cues while maintaining an end-to-end trainable pipeline. ADODN combines geometry-aware modeling for structural context, a stochastic attentive dropout mechanism that alternates between dropping and highlighting responses during training, and low-rank regularization to exploit inter-feature correlations for feature recovery.

On challenging benchmarks, ADODN achieves NRMSE scores of 2.80 on the 300W Common set, 5.81 on the Challenging set, and 3.35 on the Full set, outperforming recent baselines including ODN, AODN, RHT-R, and POPos. We intentionally avoid overstating modest margins and instead support the conclusions using both mean errors and distribution-level evidence from CED curves, together with paired statistical validation on per-image errors. The ablation results further indicate that each module contributes complementary benefits and that the full integration provides the strongest robustness on occlusion-heavy settings.

Limitations and future work. The current implementation relies on a ResNet-18 backbone, and scaling to larger backbones or transformer-based encoders may yield additional gains. Although attentive dropout improves robustness,

it can introduce training instability if applied too aggressively; we mitigate this with a warm-up strategy and provide sensitivity analysis for key hyperparameters. Future work will explore (i) extending ADODN to video-based FLD with temporal consistency, (ii) multi-task learning with related facial analysis tasks, (iii) further efficiency optimization via compression and acceleration for mobile/embedded deployment, and (iv) domain adaptation to improve robustness under larger distribution shifts (e.g., artistic renderings, low-resolution inputs, and extreme lighting).

REFERENCES

[1] I. Kemelmacher-Shlizerman and R. Basri, "3D face reconstruction from a single image using a single reference face shape," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 394–405, Feb. 2011.

[2] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *Int. J. Comput. Vis.*, vol. 127, no. 2, pp. 115–142, Feb. 2019.

[3] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[4] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.

[5] G. Tzimiropoulos, J. Alabort-i-Medina, S. Zafeiriou, and M. Pantic, "Generic active appearance models revisited," in *Proc. Asian Conf. Comput. Vis.*, 2013, pp. 650–663.

[6] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3444–3451.

[7] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.

[8] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment via regressing local binary features," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1233–1245, Mar. 2016.

[9] M. Sadiq, D. Shi, and J. Liang, "A robust occlusion-adaptive attention-based deep network for facial landmark detection," *Int. J. Speech Technol.*, vol. 52, no. 8, pp. 9320–9333, Jun. 2022.

[10] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, "Robust facial landmark detection via occlusion-adaptive deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3481–3491.

[11] P. Gao, K. Lu, J. Xue, L. Shao, and J. Lyu, "A coarse-to-fine facial landmark detection method based on self-attention mechanism," *IEEE Trans. Multimedia*, vol. 23, pp. 926–938, 2021.

[12] M. Sadiq and D. Shi, "Attentive occlusion-adaptive deep network for facial landmark detection," *Pattern Recognit.*, vol. 125, May 2022, Art. no. 108510.

[13] J. Wan, J. Liu, J. Zhou, Z. Lai, L. Shen, H. Sun, P. Xiong, and W. Min, "Precise facial landmark detection by reference heatmap transformer," *IEEE Trans. Image Process.*, vol. 32, pp. 1966–1977, 2023.

[14] C.-Y. Xiang, J.-Y. He, Z.-Q. Cheng, X. Wu, and X. Hua, "POPoS: Improving efficient and robust facial landmark detection with parallel optimal position search," in *Proc. AAAI Conf. Artif. Intell.*, 2025, vol. 39, no. 8, pp. 8602–8610.

[15] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1513–1520.

[16] Y. Wu and Q. Ji, "Robust facial landmark detection under significant head poses and occlusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3658–3666.

[17] J. Xing, Z. Niu, J. Huang, W. Hu, X. Zhou, and S. Yan, "Towards robust and accurate multi-view and partially-occluded face alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 987–1001, Apr. 2018.

[18] Q. Liu, J. Deng, J. Yang, G. Liu, and D. Tao, "Adaptive cascade regression model for robust face alignment," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 797–807, Feb. 2017.

[19] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2214–2223.

[20] J. Choe, S. Lee, and H. Shim, "Attention-based dropout layer for weakly supervised single object localization and semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4256–4271, Dec. 2021.

[21] X. Liao, Y. Wang, T. Wang, J. Hu, and X. Wu, "FAMM: Facial muscle motions for detecting compressed deepfake videos over social networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7236–7251, Dec. 2023.

[22] L. Fu, X. Liao, J. Guo, L. Dong, and Z. Qin, "WaveRecovery: Screen-shooting watermarking based on wavelet and recovery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 4, pp. 3603–3618, Apr. 2025.

[23] Y. Li, X. Liao, and X. Wu, "Screen-shooting resistant watermarking with grayscale deviation simulation," *IEEE Trans. Multimedia*, vol. 26, pp. 10908–10923, 2024.

[24] M. Chen, X. Liao, H. Fang, J. Guo, Y. Chen, and X. Wu, "Flexible partial screen-shooting watermarking with provable robustness," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 12, pp. 12152–12166, Dec. 2025.

[25] X. P. Ding, L. Wang, P. Koniusz, and Y. Gao, "Graph your own prompt," in *Proc. 39th Annu. Conf. Neural Inf. Process. Syst.*, 2025, pp. 1–18. [Online]. Available: https://openreview.net/forum?id=7tXGIbrIA5

[26] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action," *Trends Neurosci.*, vol. 15, no. 1, pp. 20–25, Jan. 1992.

[27] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 397–403.

[28] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1899–1906.

[29] M. Sadiq, D. Shi, M. Guo, and X. Cheng, "Facial landmark detection via attention-adaptive deep network," *IEEE Access*, vol. 7, pp. 181041–181050, 2019.

[30] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5202–5211.

[31] B. Browatzki and C. Wallraven, "3FabRec: Fast few-shot face alignment by reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6109–6119.

MUHAMMAD SADIQ is currently an Assistant Professor and a Foreign Technology Expert with Shenzhen University of Information Technology, Shenzhen, China. His research interests include artificial intelligence and deep learning, vehicular network security and intrusion detection systems, facial landmark detection under occlusion, blockchain and federated learning, and cloud/edge security. He has over four years of university teaching experience and more than 18 years of professional experience in cybersecurity and AI, with multiple research grants in facial analysis and cybersecurity training. He has published extensively in high-impact journals and conferences in AI, network security, and computer vision; and actively leads cross-border collaborations with institutions in Pakistan, Saudi Arabia, and other Belt and Road countries.

JUNHAO WU received the bachelor's and master's degrees in computer science and technology from Shenzhen University and the Ph.D. degree in parallel information processing, in 2020. He is currently a Lecturer with the College of Mathematics and Computer Science, Shantou University. His research interests include medical image processing, high-performance computing, machine learning and deep learning algorithms, medical image registration, parallel computing performance analysis, and classification algorithm optimization. His significant academic contributions include publications in top-tier journals, such as IEEE Transactions on Pattern Analysis and Machine Intelligence (2023, CCF A-rated, Q1) and IEEE Transactions on Parallel and Distributed Systems (CCF A-rated).

YU GENG received the B.Eng. degree in optical engineering from Yanshan University, Qinhuangdao, China, in 2006, the master's degree in optical engineering from Zhejiang University, Hangzhou, China, in 2008, and the Ph.D. degree in electrical and computer engineering from The Hong Kong University of Science and Technology, Hong Kong, China, in 2015. He is currently an Associate Professor with Shenzhen University of Information Technology, Shenzhen, China. His research interests include semiconductor devices simulation and fabrication, liquid crystal, and artificial intelligence application in the above areas.

MOHAMMAD SULTAN MAHMUD (Senior Member, IEEE) received the M.Sc. degree from the King Mongkut's University of Technology North Bangkok, Thailand, in 2014, and the Ph.D. degree from Shenzhen University, China, in 2023. He is currently a Research Associate Professor with Shenzhen Technology University, with a broad interest in the theoretical and practical aspects of high-performance computing, particularly in big data intelligent processing and analysis, data mining and machine learning, scalable computing, and ensemble learning. Following the Ph.D. degree, he was a Research Fellow at Shenzhen University, between 2024 and 2025. He is one of the pioneers in the distributed clustering ensemble of big data research. He has published research articles in esteemed journals, such as *Pattern Recognition*, *Information Fusion*, IEEE TRANSACTIONS ON BIG DATA, *Big Data Mining and Analytics*, and *Journal of Big Data*. His research was recognized with the Excellent Paper Award at the Big Data Mining and Analytics, in 2021. He has organized a special issue on "Mixture of Experts (MoE) and Ensemble Learning for Big Data" in *Information Fusion* (Elsevier).

AMAR KHELLOUFI received the B.S. degree (Hons.) in computer science from the Faculty of Sciences and Technology, Ziane Achour University of Djelfa, Djelfa, Algeria, in 2012, the M.S. degree in distributed information systems from the Faculty of Sciences, University of Boumerdés, Boumerdés, Algeria, and the Ph.D. degree in computer and communication engineering from the University of Science and Technology, Beijing, China, in 2024. He is currently a Research Associate with Shenzhen University of Information Technology. His current research interests include the Internet of Things, the Social Internet of Things, AGI, service recommendation, and distributed systems.

HUA ZHENG received the Ph.D. degree from Bournemouth University, U.K., in 2024. He is currently a Lecturer with Guangzhou University of Software, Guangzhou, China. His research interests include self-supervised learning and its applications in graph learning (e.g., graph clustering and KG embedding), foundation models (e.g., LLMs and MLLMs), bioinformatics, and AI pharmaceuticals.

YUNSHENG ZHANG received the M.S. degree in software engineering and the Ph.D. degree in computer software and theory from the University of Electronic Science and Technology of China (UESTC), in 2005 and 2011, respectively. He is currently an Associate Professor with Shenzhen University of Information Technology. During his doctoral studies, he was a Joint-Training Ph.D. Student with the University of Missouri, Columbia, from 2007 to 2009, and a Visiting Scholar at Yale University, in 2010. He conducted postdoctoral research with Tsinghua University (2011–2012) and subsequently worked as a Research Fellow at The Hong Kong Polytechnic University (2016–2018). His research interests include computer vision and signal processing, networks and communications, parallel and distributed computing, and AI agent technology.

JUNWEI LIANG (Member, IEEE) received the Ph.D. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore, in 2021. He is currently an Assistant Professor with Shenzhen University of Information Technology and a Research Fellow with the School of Electrical and Electronic Engineering, NTU. His research interests include cybersecurity, evolutionary computation, and machine learning, with applications in industrial control systems and federated learning. He has published over 30 peer-reviewed articles in top-tier journals (e.g., IEEE TRANSACTIONS) and conferences and leads multiple projects on privacy-preserving intrusion detection for critical infrastructures.