Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Frequency-Aware Gradient Correction for Unified Facial Landmark Detection

Shun Ren
China Three Gorges University
Qingjia Li
China Three Gorges University
Hang Sun
China Three Gorges University
Xu Wu
Nanyang Technological University
Beihang Song
Ministry of Emergency Management of China
Jun Wan
junwan2014@whu.edu.cn

Zhongnan University of Economics and Law

Research Article

Keywords: Unified Facial Landmark Detection, Heatmap Regression, Frequency Learning, Multi-Dataset Training, Gradient Correction, Transformer.

Posted Date: August 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10481504/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Frequency-Aware Gradient Correction for Unified Facial Landmark Detection

Shun Ren[1], Qingjia Li[1], Hang Sun[1], Xu Wu[4], Beihang Song[2], Jun Wan[3,4*]

[1]College of Computer and Information Technology, China Three Gorges University, Yichang, 443002, Hubei, China.
[2]National Institute of Natural Hazards, Ministry of Emergency Management of China, Beijing, 100085, China.
[3]School of Information Engineering, Zhongnan University of Economics and Law, Wuhan, 430073, Hubei, China.
[4]School of Computer Science and Engineering, Nanyang Technological University, Singapore, 639798, Singapore.

*Corresponding author(s). E-mail(s): junwan2014@whu.edu.cn;
Contributing authors: renshun@ctgu.edu.cn; 1990760379@qq.com;
sunhang0418@whu.edu.cn; N2309016A@e.ntu.edu.sg; bhs605906@whu.edu.cn;

Abstract

Facial landmark detection (FLD) across multiple datasets remains challenging due to heterogeneous annotation protocols, data distributions, and inconsistent supervision, which often lead to representation bias and gradient conflicts in unified models. In this paper, we propose Frequency-Aware Gradient Correction for Unified Facial Landmark Detection (FGC-UFLD), a unified framework that jointly addresses representation learning and optimization conflicts from both frequency and gradient perspectives. Specifically, we introduce two frequency-aware representation modules: a Diagonal-Compensated Wavelet Module (DiC-Wave) and a Progressive Frequency-Aligned Module (PFAM), which together exploit high-frequency information to progressively refine backbone representations for robust facial structural modeling. In addition, we propose an Anchor-Guided Gradient Correction (AGGC) strategy that explicitly models inter-dataset gradient interactions, mitigating gradient conflicts and enabling stable and consistent optimization under mixed-batch multi-dataset training. Extensive experiments on four widely used benchmarks demonstrate that FGC-UFLD achieves performance comparable to state-of-the-art methods, while relying on a single unified model. The code is publicly available at: https://github.com/17762682759/FGC-UFLD.

Keywords: Unified Facial Landmark Detection , Heatmap Regression, Frequency Learning , Multi-Dataset Training, Gradient Correction , Transformer.

1 Introduction

Facial Landmark Detection (FLD), also known as Face Alignment, aims to accurately localize predefined semantic keypoints (e.g., eye corners, nose tip, and mouth corners) within a facial image. As a core task in computer vision, FLD plays a central role in downstream applications such as facial expression classification [1], 3D face reconstruction [2], and human-computer interaction [3].

Existing deep learning-based FLD methods are typically categorized into two classes: coordinate regression methods [4] and heatmap regression methods [5]. The former directly predicts landmark coordinates with efficient computation. The latter generates probability heatmaps for each landmark, enabling stronger spatial consistency and better performance in challenging environments. Recently, Transformers [6] have

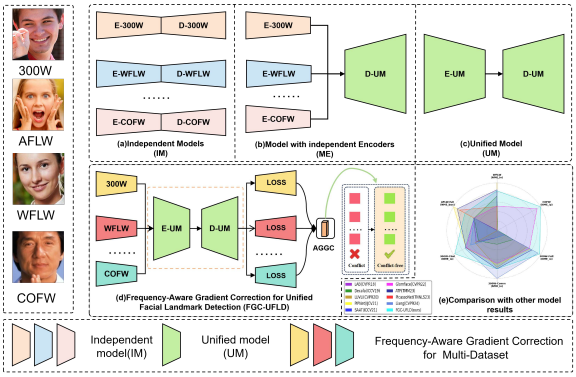


Fig. 1 (a) Independent Model (IM): separate E-300W, E-WFLW, and E-COFW. (b) Independent Encoders (ME): independent encoders per dataset. (c) Unified Model (UM): shared E-UM and D-UM. (d) FGC-UFLD: frequency-guided fusion and gradient correction within a unified framework. (e) Comparison with SOTA FLD methods on 300W, COFW, WFLW, and AFLW.

been adopted for FLD, modeling global relationships and long-range dependencies to handle complex facial geometry. However, their high computational cost and reliance on large-scale data remain challenging.Despite significant progress in heatmap regression, coordinate regression, and Transformer-based approaches, performance bottlenecks remain under extreme poses, illumination variations, and severe occlusions. Key challenges include: 1) scarcity of high-quality annotated data due to expensive and subjective manual annotation and limited generalization; 2) difficulty in structural modeling under complex conditions, where large pose and illumination variations disrupt local appearance and degrade global geometric consistency. Although datasets annotate different numbers of landmarks, they share semantically consistent keypoints. Effectively leveraging cross-dataset overlapping landmarks improves accuracy, while dataset-specific landmarks provide additional structural cues. Our goal is to learn unified representations and consistent facial structures for robust and generalizable FLD.However, most FLD methods are trained independently on single datasets (IM), limiting their ability to learn unified transferable representations across heterogeneous data sources (Fig. 1 (a)) [7]. Similar challenges in other dense prediction tasks have inspired unified frameworks, such as multi-encoder shared-decoder architectures and unified models (Fig. 1 (b)(c)) [8], which rely on implicit feature routing (e.g., MoE [9]) and fail to explicitly coordinate gradients across datasets.

To address these challenges, we propose FGC-UFLD, comprising DiC-Wave, PFAM, and AGGC. The first two modules incorporate frequency-domain cues to enhance representations and reinforce structural constraints across scales, while AGGC explicitly models gradient interactions to mitigate conflicts and stabilize multi-dataset optimization. By integrating frequency-aware learning with gradient-level coordination, FGC-UFLD achieves robust and accurate facial heatmap regression under challenging conditions. The main contributions are summarized as follows:

- We propose Anchor-Guided Gradient Correction (AGGC) to decouple conflicting gradients in heterogeneous datasets, achieving stable multi-task optimization for unified facial landmark detection.
- We introduce DiC-Wave and PFAM, two frequency-aware modules that leverage high-frequency cues for backbone feature refinement and progressive multi-scale fusion, enabling robust facial structure modeling.
- Extensive experiments on four benchmark datasets, including AFLW [10], WFLW [4], COFW [11], and 300W [12], demonstrate that FGC-UFLD achieves performance comparable to state-of-the-art FLD methods while relying on a single unified model.

2 Related Work

Early FLD methods, including Active Shape Models and Active Appearance Models, relied on statistical modeling and template matching with limited robustness. With deep learning, FLD has shifted to coordinate- and heatmap regression-based approaches, evolving from single-task to unified multi-task frameworks.

Coordinate Regression. Coordinate-based regression methods directly predict facial landmark coordinates, offering high efficiency and end-to-end localization. Early methods like PFLD [13] use lightweight ConvNets with auxiliary pose estimation. Recent approaches adopt Transformers for long-range modeling: SLPT [14] uses sparse attention on local patches, while DTLD [15] employs a dual-branch Transformer for global propagation. Others, including PicassoNet [16] and LDDMM [17], focus on network optimization and deformation modeling. Despite their efficiency, coordinate-based methods often struggle with spatial uncertainty and structural consistency under challenging conditions.

Heatmap Regression. Heatmap-based regression predicts a likelihood map per landmark, preserving spatial information and enabling explicit modeling of facial structure and uncertainty. Classical architectures like Stacked Hourglass [18] and HRNet [19] employ multi-scale feature fusion to capture global and local context.

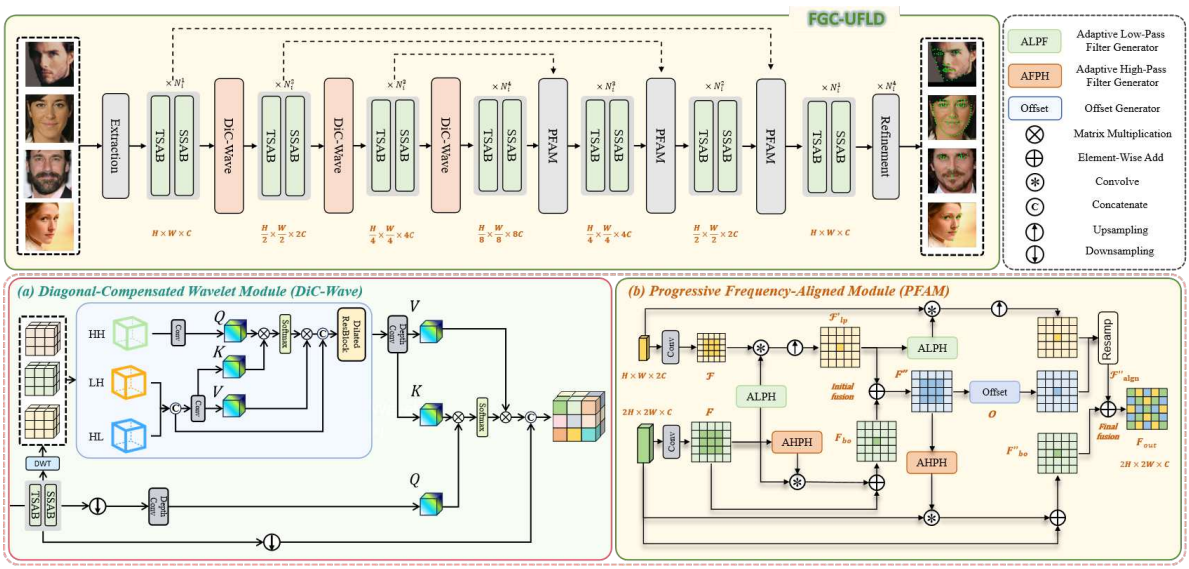


Fig. 2 FGC-UFLD uses a convolutional backbone to extract base features, enhanced by DiC-Wave for structural awareness and fused by PFAM with multi-scale high-frequency details, then a heatmap regression head outputs unified landmark coordinates for cross-dataset detection.

LAB [4] incorporates facial boundary heatmaps as additional supervision for shape constraints, while AWing [20] applies an adaptive loss to balance foreground and background. Several methods have been proposed to enhance robustness under challenging conditions. LUVLi [21] models uncertainty to suppress unreliable landmarks, and ADNet [22] strengthens structural modeling via anisotropic direction loss and point-edge heatmaps. Building upon this paradigm, we incorporate frequency-domain modeling and explicit structural constraints to further enhance representation learning and localization accuracy.

Unified and Multi-Task Frameworks. Recent studies have extended heatmap regression to unified and multi-task frameworks for handling multiple datasets or related facial analysis tasks in a single model. ATF [23] proposes a weakly supervised paradigm that exploits inter-dataset commonalities for unified learning across different annotation protocols, while FaceXFormer [24] introduces a unified Transformer architecture supporting ten facial analysis tasks within a single encoder-decoder framework. From an optimization perspective, multi-task learning often suffers from task imbalance and gradient conflicts. Existing methods such as uncertainty weighting [25], GradNorm [26], and DWA [27] belong to the category of dynamic loss weighting strategies. Although they can balance task losses, they fail to explicitly resolve gradient direction conflicts, leading to unstable optimization in large-scale joint training. To address this, we propose Anchor-Guided Gradient Correction (AGGC), which explicitly models gradient interactions and dynamically decouples conflicts in the optimization space. Through fine-grained, conflict-aware updates, AGGC improves training stability and achieves higher landmark localization accuracy across heterogeneous datasets.

3 Method

The proposed FGC-UFLD (Fig. 2) adopts X-Restormer [28] as the backbone, where each Transformer block consists of a channel self-attention branch (TSAB) [28] and a shifted-window self-attention branch (SSAB) [28] to model channel-wise semantics and localized spatial structures.

Given an input facial image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, a convolutional backbone extracts multi-scale features. In the downsampling encoding stage, features are processed by Transformer blocks and then fed to the Diagonal Compensated Wavelet Module (DiC-Wave) for encoding spatial and frequency-domain information. In the upscaling decoding stage, features are refined by Transformer blocks and passed through the Progressive Frequency-Aligned Module (PFAM) for progressive frequency-gated enhancement and alignment. Finally, Anchor-Guided Gradient Correction (AGGC) is incorporated to alleviate gradient conflicts from multi-source datasets, enabling end-to-end joint training for improved FLD. The detailed designs of DiC-Wave, PFAM, and AGGC are presented in the following subsections.

Diagonal-Compensated Wavelet Module (DiC-Wave). High-frequency components in multi-scale facial representations encode critical fine-grained cues, such as sharp edges and local texture variations that are essential for

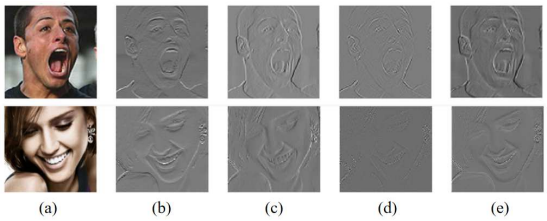


Fig. 3 (a) Original image. (b) Vertical sub-band F_{LH}. (c) Horizontal sub-band F_{HL}. (d) Diagonal sub-band F_{HH}. (e) Refined diagonal response F_{HH}^{ref}. DiC-Wave uses cross-directional wavelet attention to enhance sparse textures and improve high-frequency representations for FLD.

accurate FLD. Subtle structural changes around key regions (e.g., eye corners, mouth contours, and nasal boundaries) are predominantly characterized by high-frequency responses. However, directly incorporating raw high-frequency information remains challenging, as such responses are typically sparse and highly localized, making them prone to structural inconsistency or optimization instability when naively injected into deep features. To address this, we propose a Diagonal-Compensated Wavelet Module (DiC-Wave) that constructs a stable and discriminative high-frequency representation via cross-directional wavelet attention to complement global structural semantics. Let F_{in} be the input feature to DiC-Wave from the corresponding encoder Transformer block. The discrete wavelet transform (DWT) is then applied to explicitly disentangle directional high-frequency components:

$$[F_{LH}, F_{HL}, F_{HH}] = \text{DWT}(F_{in}) \tag{1}$$

where F_{LH} and F_{HL} capture vertical and horizontal edge responses, respectively, while F_{HH} encodes diagonal textures that are informative for landmark corners and junctions. Despite its importance, the diagonal sub-band F_{HH} is inherently sparse and unevenly distributed, making it difficult to exploit directly and prone to optimization instability or loss of structural detail. In contrast, the horizontal and vertical sub-bands provide denser and more reliable directional cues, serving as natural structural supports for refining diagonal high-frequency representations. Motivated by this, DiC-Wave introduces a cross-attention mechanism that explicitly enhances and regularizes the diagonal sub-band by leveraging complementary information from horizontal and vertical directions. Specifically, DiC-Wave treats the diagonal component F_{HH} as the query, with the vertical F_{LH} and horizontal F_{HL} sub-bands as complementary keys and values, enabling the model to infer missing or weak diagonal responses from more stable directional structures. Formally, the cross-attention operation is defined as:

$$Q = W_q F_{HH}, \tag{2}$$

$$K = W_k[F_{LH}, F_{HL}], \tag{3}$$

$$V = W_v[F_{LH}, F_{HL}] \tag{4}$$

$$F_{HH}^{\text{ref}} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right) V, \tag{5}$$

where W_q, W_k, and W_v are learnable projection matrices, d_h denotes the channel dimension, and F_{HH}^{ref} represents the refined diagonal high-frequency textures. This enables diagonal high-frequency responses to be adaptively reconstructed and enriched using structurally consistent cues from orthogonal directions, alleviating sparsity-induced instability.

F_{HH}^{ref} is then fused with F_{LH} and F_{HL}, refined through dilated residual blocks and channel-wise fusion, to produce F_{HF}^{ref}(as shown in Fig. 3(e)), which preserves directional consistency across sub-bands while enriching local details and long-range structural context. F_{HF}^{ref} further guides multi-scale feature aggregation. Following the same cross-attention formulation, we treat the downsampled input feature $F_{in}\downarrow$ as the query and F_{HF}^{ref} as the key–value pair, yielding the frequency-guided refined feature F_{in}^{ref}.

In this way, DiC-Wave produces a compact and expressive high-frequency representation via cross-directional wavelet attention, which directs multi-scale Transformer blocks to extract more informative features and effectively enhances FLD performance.

Progressive Frequency-Aligned Module (PFAM). Although DiC-Wave produces multi-scale and multi-frequency representations, a vanilla decoder with naïve upsampling often blurs high-frequency cues, introduces spatial misalignment, and disperses heatmap energy under complex facial deformations and illumination variations. To address this, we propose *Progressive Frequency-Aligned Module (PFAM)*, which performs frequency-aware decoding by explicitly aligning cross-scale features and progressively enhancing high-frequency details.

At the l-th decoding stage, the encoder's high-resolution feature and the decoder's low-resolution feature are projected to a unified channel dimension via 1×1 convolution, yielding $\mathbf{F}$ and $\mathcal{F}$. The content-aware reassembly operator (Carafe [29]) is denoted as $\mathcal{C}(\mathbf{X}, \mathbf{W}, K)$, which aggregates unfolded patches of $\mathbf{X}$ with location-specific weights $\mathbf{W}$ and a kernel size K. $\uparrow$ denotes upsampling to the high-resolution.

High-pass enhancement. To recover fine details suppressed by downsampling, we generate an

adaptive high-pass kernel $\mathbf{W}_{hp}$ from $\mathbf{F}$ using an Adaptive High-Pass Filter (AHPF). The high-pass response is extracted via content-aware filtering and reinjected in a residual form:

$$\begin{aligned} \mathbf{F}_{hp} &= \mathcal{C}(\mathbf{F}, \mathbf{W}_{hp}, K_{hp}), \\ \mathbf{F}_{bo} &= \mathbf{F} + \big(\mathbf{F} - \mathbf{F}_{hp}\big). \end{aligned} \tag{6}$$

This formulation explicitly boosts high-frequency components while preserving the original semantic content, and $\mathbf{F}_{bo}$ denotes the boosted high-frequency high-resolution features.

Low-pass fusion. In parallel, we extract low-frequency responses from both branches using an Adaptive Low-Pass Filter (ALPF) , producing a low-pass kernel $\mathbf{W}_{lp}$. We then upsample the low-resolution low-frequency feature with $\mathbf{W}_{lp}$ and fuse it with the high-resolution low-frequency cue:

$$\mathcal{F}'_{lp} = \mathbf{F}_{lp} + \mathcal{C}(\mathcal{F}_{lp}, \mathbf{W}_{lp}, K_{lp})\uparrow, \tag{7}$$

where $\mathbf{F}_{lp}$ and $\mathcal{F}_{lp}$ are low-pass responses extracted from $\mathbf{F}_{bo}$ and $\mathcal{F}$, respectively. The fused low-frequency representation $\mathcal{F}'_{lp}$ provides a stable structural reference for subsequent refinement. Then, using the fused low-pass guidance, we integrate high-frequency cues from both branches to obtain an initial fused feature:

$$\mathbf{F}'' = \mathbf{F}_{hp} + \mathcal{C}(\mathcal{F}_{hp}, \mathbf{W}'_{lp}, K'_{lp})\uparrow, \tag{8}$$

where $\mathbf{F}_{hp}$ and $\mathcal{F}_{hp}$ denote high-pass responses extracted from $\mathbf{F}$ and $\mathcal{F}$, and $\mathbf{W}'_{lp}$ is the low-pass kernel refined from $\mathbf{F}'_{lp}$.

Content-adaptive alignment. Directly fusing high-resolution and upsampled low-resolution features may suffer from spatial misalignment. We therefore predict pixel-wise offsets and resample the low-resolution feature before fusion:

$$\mathbf{O} = \big(\mathcal{A} + \mathrm{PU}(\mathbf{A})\big) \odot \sigma\big(\mathcal{B} + \mathrm{PU}(\mathbf{B})\big) + \mathbf{P}_{init}, \tag{9}$$

$$\mathcal{F}_{algn} = \big(\mathcal{C}(\mathcal{F}, \mathbf{W}'_{lp}, K'_{lp}),\ \mathbf{O}\big)\uparrow \tag{10}$$

$$\mathbf{F}_{out} = \mathbf{F}''_{bo} + \mathcal{F}_{algn}, \tag{11}$$

where $(\mathcal{A}, \mathcal{B})$ and $(\mathbf{A}, \mathbf{B})$ are estimated from low- and high-resolution features, respectively. $\mathrm{PU}(\cdot)$ denotes pixel unshuffle, and $\sigma(\cdot)$ is the sigmoid function. $\mathcal{F}_{algn}$ is the aligned low-frequency enhanced feature, and $\mathbf{F}''_{bo}$ is the high-frequency enhanced feature, whose adaptive high-pass kernel $\mathbf{W}$ is generated from $\mathbf{F}''$. The final output $\mathbf{F}_{out}$ is obtained by fusing $\mathbf{F}''_{bo}$ with $\mathcal{F}_{algn}$. Through progressive frequency-gated enhancement and alignment, PFAM produces sharper heatmap peaks and more concentrated energy distributions.

Anchor-Guided Gradient Correction (AGGC). In multi-dataset facial landmark detection with a unified heatmap head, joint optimization across heterogeneous datasets is essential for learning robust and transferable structural representations. However, common strategies such as alternating updates and sequential training suffer from inherent limitations—they prevent gradients from different datasets from being simultaneously computed and coordinated, and frequent task switching leads to discontinuous optimization trajectories, causing unbalanced learning and performance degradation. To address these issues, we adopt a mixed-batch training strategy. In each iteration, samples from all datasets are concatenated into a single batch and forwarded simultaneously, then partitioned by dataset identity,with individual heatmap losses computed per subset. This ensures gradients from all datasets are jointly available within the same backpropagation step, laying the foundation for effective multi-task gradient analysis and correction.

Problem Formulation. Despite the stability brought by mixed-batch training, heterogeneity in annotation protocols, keypoint definitions, and data distributions inevitably introduces *gradient conflicts* during joint optimization, which may lead to oscillatory updates and negative transfer between datasets when using a unified prediction head.

For heatmap-based landmark detection, we define the regression loss for each dataset $\mathcal{D}_i$ as:

$$\mathcal{L}_i = \frac{1}{N} \sum_{n=1}^{N} \|H_i(n) - H_i^*(n)\|_2^2, \tag{12}$$

where H_i and H_i^* denote the predicted and ground-truth heatmaps, respectively, N is the number of landmarks in this dataset. The overall objective is a weighted combination of all datasets:

$$\mathcal{L}_{total} = \sum_{i=1}^{M} \lambda_i \mathcal{L}_i, \tag{13}$$

where M denotes the number of datasets, and λ_i is a non-negative scalar that controls the relative contribution of the i-th dataset to the overall optimization objective.

Let θ denote the network parameters. For a mixed batch involving M datasets, the gradient corresponding to the m-th dataset is:

$$g_m = \nabla_\theta \mathcal{L}_m. \tag{14}$$

Anchor-Guided Gradient Correction. To explicitly resolve gradient conflicts under mixed-batch training, we propose Anchor-Guided Gradient Correction (AGGC), which performs pairwise conflict checks ($g_i^\top g_j < 0$) to stabilize coordination in heterogeneous settings. At each backpropagation step, one task is randomly selected as the anchor with gradient $\bar{g}$, serving as a temporary reference direction that introduces stochasticity and avoids fixed-task bias. For each remaining gradient g_i, a conflict is detected if $g_i^\top \bar{g} < 0$. If conflicted, the component inconsistent with the anchor direction is removed via orthogonal projection:

$$g_i \leftarrow g_i - \frac{g_i^\top \bar{g}}{\|\bar{g}\|^2}\bar{g}. \tag{15}$$

This correction preserves task-specific gradient information orthogonal to the anchor while removing destructive interference along the anchor direction. Using WFLW as the fixed anchor provides a stable reference, preventing bias toward other datasets and encouraging consistent alignment across tasks.

After correction, the final update direction $\hat{g}$ is obtained by averaging all task gradients, and the model parameters are updated as:

$$\theta \leftarrow \theta - \eta\hat{g}. \tag{16}$$

Combining mixed-batch training with anchor-guided gradient correction thus suppresses gradient conflicts and ensures stable convergence in multi-dataset joint learning.

4 Experiments

Experiments are conducted on four benchmark datasets—300W [12], COFW [11], WFLW [4], and AFLW [10]—with comparisons to state-of-the-art FLD methods. Ablation studies assess individual network components.

300W (68 landmarks) [12]: Contains 3,148 training and 689 testing images. The training set comprises AFW, LFPW, and HELEN subsets, while the test set includes LFPW, HELEN, and IBUG. Following the standard protocol, the test set is divided into a common subset (554 images) and a challenging subset (135 images).

WFLW (98 landmarks) [4]: Contains 10,000 face images annotated with 98 landmarks, split into 7,500 for training and 2,500 for testing. Provides multiple attribute labels—occlusion, pose, expression, illumination, blur, and makeup—for comprehensive evaluation under diverse conditions.

Table 1 Comparison against SOTA approaches on the 300W dataset, where the normalized mean error (NME) is measured relative to the inter-ocular distance.

Method	Common	Challen.	Full
LAB (CVPR18) [4]	2.98	5.19	3.49
AWing (ICCV19) [20]	2.72	4.52	3.07
ODN (CVPR19) [30]	3.56	6.67	4.17
DeCaFA (ICCV19) [31]	2.93	5.26	3.39
LUVLi (CVPR20) [21]	2.76	5.16	3.23
SAAT (ICCV21) [32]	2.82	5.03	3.25
ADNet (ICCV21) [22]	2.53	4.58	2.93
PIPNet (IJCV21) [33]	2.78	4.89	3.19
SLPT (CVPR22) [14]	2.75	4.90	3.17
GlomFace (CVPR22) [34]	2.79	4.87	3.20
DTLD (CVPR22) [35]	2.59	4.50	2.96
ATF (TMM22) [23]	2.75	4.89	3.17
PicassoNet (TNNLS23) [16]	3.03	5.81	3.58
Lite-HRNet (ICIP23) [36]	3.97	6.89	4.54
LDDMM-Face (CVPR24) [17]	3.07	5.40	3.53
ORFormer (WACV25) [37]	2.53	4.43	2.90
FGC-UFLD (ours)	**2.75**	**4.56**	**3.08**

COFW (29 landmarks) [11]: Comprises 1,345 training and 507 testing images with 29 landmarks. The training set includes 845 images from LFPW and 500 additional heavily occluded images. Test images feature significant occlusions, large pose variations, and diverse expressions.

AFLW (19 landmarks) [10]: Contains 25,993 annotated faces with 19 landmarks, covering large head poses with yaw angles from –120° to +120° and pitch angles from –90° to +90°. AFLW-Full uses 20,000 images for training and 4,386 for testing, while AFLW-Frontal selects 1,165 near-frontal faces from the test subset for frontal-view evaluation.

Evaluation Metrics: Normalized Mean Error (NME) is used to assess facial alignment accuracy. NME_{ip} (inter-pupil distance) is used for COFW, NME_{io} (inter-ocular distance) for 300W and WFLW, and NME_{box} (bounding box size) for AFLW. Failure Rate (FR) [4] is also reported for COFW. For an image with N landmarks, NME is defined as:

$$\text{NME} = \frac{1}{N}\sum_{i=1}^{N}\frac{\|p_i - p_i^*\|_2}{d} \times 100\%, \tag{17}$$

where p_i and p_i^* are the predicted and ground-truth coordinates of the i-th landmark, $\|\cdot\|_2$ is the Euclidean norm, and d is the normalization factor.

The Failure Rate (FR) measures the proportion of predictions with NME exceeding a threshold τ. For a dataset with N_s samples:

$$\text{FR} = \frac{1}{N_s}\sum_{i=1}^{N_s}\mathbb{I}\big(\text{NME}(Y_i, \hat{Y}_i) > \tau\big), \tag{18}$$

Table 2 Results on the WFLW subset compared with SOTA methods; NME normalized by inter-ocular distance. Symbols: ∘ – heatmap regression; ◇ – coordinate regression (% omitted).

Method	Testset	Pose	Expression	Illumination	Make-Up	Occlusion	Blur
∘ LAB(CVPR18) [4]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
∘ Wing(CVPR18) [38]	4.99	8.75	5.36	4.93	5.41	6.37	5.81
∘ DeCaFA(ICCV19) [31]	4.62	8.11	4.65	4.41	4.63	5.74	5.38
∘ HRNet(CVPR19)[19]	4.60	–	–	–	–	–	–
∘ AWing(ICCV19)[20]	4.36	–	–	–	–	–	–
◇ PIPNet(IJCV21)[33]	4.31	–	–	–	–	–	–
◇ MMDN(TNNLS22)[39]	4.87	7.71	4.79	4.61	4.72	6.17	5.72
◇ GlomFace(CVPR22)[34]	4.81	8.71	–	–	–	5.14	–
◇ ATF(TMM23)[23]	4.50	7.54	4.65	4.45	4.20	5.30	5.19
◇ PicassoNet(TNNLS23)[16]	4.82	8.61	5.14	4.73	4.68	5.91	5.56
◇ HRNet(ICIP23)[36]	5.58	9.79	6.13	5.44	5.87	6.57	6.05
◇ LDDMM-Face(CVPR24)[17]	4.63	8.21	5.00	4.53	4.31	5.37	5.22
◇ **FGC-UFLD (ours)**	**4.52**	**7.62**	**4.78**	**4.55**	**4.45**	**5.65**	**5.13**

where $\mathbb{I}(\cdot)$ is the indicator function. The threshold is set to $\tau = 0.10$.

Implementation Details: Input images are $256 \times 256 \times 3$. Data augmentation includes: (1) random scaling (1.0–1.25), (2) in-plane rotation ($-30°$ to $+30°$, 60% samples), and (3) horizontal flipping (0.5 probability). FGC-UFLD is trained for 200k iterations on a single NVIDIA RTX 4090 GPU. The initial learning rate is 1.0×10^{-4}, decayed by 0.5 every 40k iterations.

4.1 Normal Condition Evaluations

Under standard imaging conditions (no significant occlusion, extreme pose, illumination variation, or motion blur), we evaluate FGC-UFLD on the 300W-Common and 300W-Full subsets. As reported in Table 1, FGC-UFLD achieves NME_{io} of 2.75% on Common and 3.08% on Full, comparable to state-of-the-art heatmap- and coordinate-regression methods. This demonstrates the effectiveness of unified landmark mapping in learning consistent structural representations across heterogeneous datasets.

4.2 Occlusion Robustness Evaluation

Occlusion remains a major challenge in facial landmark detection, as partially occluded regions disrupt the continuity of semantic features and hinder accurate inference of complete facial structures. To evaluate the robustness of FGC-UFLD under different conditions, we conduct experiments on the occlusion subsets of the WFLW dataset and the COFW dataset, where facial regions experience significant occlusion or challenging pose/illumination variations. As reported in Table 3, on the COFW dataset, FGC-UFLD achieves an NME_{ip} of 4.82% and an FR10_{ip} of 0.39%, demonstrating strong robustness under severe occlusion. On the occlusion subset of WFLW (Table 2), the proposed approach attains an NME_{io} of 5.65%, reflecting reliable landmark localization performance in the presence of partial occlusion. Overall, these results suggest that FGC-UFLD maintains stable and accurate facial landmark detection across diverse and challenging scenarios involving occlusion.

Table 3 Comparisons with state-of-the-art methods on the COFW dataset. The normalized mean error (NME) is computed with respect to the inter-pupil distance. (% omitted)

Method	NME_{ip}	FR
ODN (CVPR19) [30]	5.30	–
AWing (ICCV19) [20]	4.94	0.99
MMDN (TNNLS22) [39]	5.01	1.78
SLPT (CVPR22) [14]	4.79	1.18
STAR (CVPR23) [40]	4.62	0.79
DSLPT-R50 (TPAMI23) [41]	4.81	1.18
CIT-v2 (IJCV24) [15]	4.93	1.58
LDDMM-Face (CVPR24) [17]	4.54	1.58
ORFormer (WACV25) [37]	4.46	-
FGC-UFLD (ours)	**4.82**	**0.39**

Table 4 Comparison with state-of-the-art (SOTA) methods on the AFLW dataset. The normalized mean error (NME) is computed relative to face size.

Method	Testset
∘ LAB (CVPR18) [4]	1.85
∘ HRNet (CVPR19) [19]	1.57
∘ AWing (ICCV19) [20]	1.53
∘ LUVLi (CVPR20) [21]	2.28
◇ PIPNet (IJCV21) [33]	1.48
◇ DTLD (CVPR22) [35]	1.38
◇ ATF (TMM23) [23]	1.55
◇ PicassoNet (TNNLS23) [16]	1.59
◇ **FGC-UFLD (ours)**	**1.69**

4.3 Large Pose Robustness Evaluation

Large pose variations and exaggerated facial expressions introduce significant geometric distortions and semantic ambiguities, challenging

facial landmark detection. To evaluate robustness, we test on AFLW full test set, WFLW pose/expression subsets, and 300W Challenge. On 300W Challenge, FGC-UFLD achieves NME_{io} of 4.56%, ranking second among compared methods [16, 36]. On WFLW subsets, our model attains NME_{io} of 7.62 and 4.78, surpassing most prior approaches [16, 36, 42]. On AFLW full set, FGC-UFLD achieves NME_{diag} of 1.69%, comparable to representative methods [4, 21]. The strong performance stems from frequency-aware structural modeling and anchor-guided gradient correction, enhancing pose-robust representation and stabilizing multi-dataset optimization under severe geometric variations.

4.4 Blur Robustness Evaluation

Facial images captured under low-light or blurry conditions often suffer from a severe loss of visual details, posing significant challenges for landmark detection. To evaluate the robustness of FGC-UFLD under such adverse scenarios, we conduct experiments on the illumination and blur subsets of the WFLW dataset. As shown in Table 2, FGC-UFLD achieves an NME_{io} of 4.55 on the WFLW illumination subset and 5.13 on the blur subset, outperforming most current methods [4, 16, 36, 38, 39]. These results indicate that FGC-UFLD can maintain accurate landmark localization even when visual quality is severely degraded. This robustness stems from the model's frequency-guided feature fusion, which enhances fine-grained structural details while preserving global semantic consistency under challenging imaging conditions.

4.5 Ablation Study

The ablation studies are conducted to evaluate the contributions of individual components in FGC-UFLD, including the Diagonal-Compensated Wavelet Module (DiC-Wave), the Progressive Frequency-Aligned Module (PFAM), and Anchor-Guided Gradient Correction (AGGC), as well as the effects of multi-dataset joint training.

Impact of DiC-Wave, PFAM and AGGC. To evaluate the contribution of each module, we conduct ablation experiments on the baseline Transformer (Trans) by incorporating DiC-Wave, PFAM, and AGGC individually and jointly, yielding (Trans+DiC-Wave), (Trans+DiC-Wave+PFAM), and (Trans+DiC-Wave+PFAM+AGGC). All models are evaluated on challenging subsets across multiple datasets. As reported in Table 5, (Trans+DiC-Wave+PFAM+AGGC) achieves the highest accuracy, confirming the complementary effects of the three modules. DiC-Wave injects high-frequency structural information to enhance feature representations and improve adaptability to diverse facial geometries. PFAM progressively fuses multi-scale features to capture fine-grained details. AGGC mitigates inter-task gradient conflicts for stable and consistent learning. Their combined integration yields the most accurate FLD performance across datasets.

Table 5 Ablation study of DiC-Wave, PFAM, and AGGC on the 300W Challenging Subset.

Method	Base	DiC-Wave	PFAM	$\mathbb{L}_{AGGC}$	NME_{io}
Trans (baseline)	✓				4.72
Trans+ $\mathbb{L}_{AGGC}$	✓			✓	4.63
Trans+DiC-Wave	✓	✓			4.65
Trans+DiC-Wave+PFAM	✓	✓	✓		4.60
FGC-UFLD(ours)	✓	✓	✓	✓	**4.56**

5 Conclusion

We present FGC-UFLD, a unified facial landmark detection framework designed to handle diverse facial variations and challenging imaging conditions. By integrating DiC-Wave and PFAM for high-frequency refinement with Anchor-Guided Gradient Correction (AGGC), the model preserves critical structural cues while ensuring stable learning across datasets with different annotation protocols. DiC-Wave and PFAM strengthen fine-grained representations by capturing and enhancing high-frequency details, improving landmark localization under extreme poses, occlusions, and blur. AGGC mitigates heterogeneous gradient conflicts, maintaining global geometric consistency during training. Extensive experiments on 300W, AFLW, COFW, and WFLW demonstrate robust and accurate performance across diverse conditions, highlighting its generalization and practical applicability.

Author contributions: Shun Ren and Jun Wan supervised the work; Qingjia Li wrote and reviewed the manuscript; Hang Sun and Beihang Song handled validation and experiments; Jun Wan handled funding acquisition and project administration.

Funding This work was jointly supported by the following projects: the Natural Science Foundation of Hubei Province of China (Grant No. 2026AFB642), the National Key Research and Development Program (No. 2024YFD1702005) and the Natural Science Foundation of Hubei Province, China (Grant No. 2024AFB992).

Data availability This study uses publicly available datasets, including 300W, AFLW, COFW, and WFLW.

Declarations

Conflict of interest The authors declare no competing interests.
Ethical approval:Not applicable.

References

[1] So, J.-K., Han, Y.: Facial landmark-driven keypoint feature extraction for robust facial expression recognition. Sensors (Basel, Switzerland) **25** (2025)

[2] Chandran, P., Zoss, G., Gotardo, P., Bradley, D.: Continuous landmark detection with 3d queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16858–16867 (2023)

[3] Vilchis, C., Perez-Guerrero, C., Mendez-Ruiz, M., Gonzalez-Mendoza, M.: A survey on the pipeline evolution of facial capture and tracking for digital humans. Multimedia Systems **29**(4), 1917–1940 (2023)

[4] Wu, W., Yang, S.: Look at boundary: A boundary-aware face alignment algorithm. In: CVPR, pp. 2129–2138 (2018)

[5] Bulat, A., Sanchez, E., Tzimiropoulos, G.: Subpixel heatmap regression for facial landmark localization. CoRR **abs/2111.02360** (2021) 2111.02360

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929** (2020) 2010.11929

[7] Wan, J., Liu, J., Zhou, J., Lai, Z., Shen, L., Sun, H., Xiong, P., Min, W.: Precise facial landmark detection by reference heatmap transformer. IEEE Transactions on Image Processing **32**, 1966–1977 (2023)

[8] Hu, S., Ma, J., Wan, J., Min, W., Jing, Y., Zhang, L., Tao, D.: Clusir: Towards cluster-guided all-in-one image restoration. arXiv preprint arXiv:2512.10948 (2025)

[9] Hu, S., Qi, H., Wan, J., et al.: Protoformer: Unified facial landmark detection by prototype transformer. arXiv preprint arXiv:2510.15338 (2025)

[10] Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: 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), pp. 2144–2151 (2011)

[11] Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: ICCV, pp. 1513–1520 (2013)

[12] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: ICCV Workshops, pp. 397–403 (2013)

[13] Watchareeruetai, U., Sommana, B., Jain, S., Noinongyao, P., Ganguly, A., Samacoits, A., Earp, S.W., Sritrakool, N.: Lotr: face landmark localization using localization transformer. IEEE Access **10**, 16530–16543 (2022)

[14] Xia, J., Qu, W., Huang, W., Zhang, J., Wang, X., Xu, M.: Sparse local patch transformer for robust face alignment and landmarks inherent relation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4052–4061 (2022)

[15] Li, Y., Tan, G., Gou, C.: Cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose. International Journal of Computer Vision **132**(4), 1242–1257 (2024)

[16] Wen, T., Ding, Z., Yao, Y., Wang, Y., Qian, X.: Picassonet: Searching adaptive architecture for efficient facial landmark localization. IEEE Transactions on Neural Networks and Learning Systems **34**(12), 10516–10527 (2022)

[17] Yang, H., Lyu, J., Cheng, P., Tam, R., Tang, X.: Lddmm-face: Large deformation diffeomorphic metric learning for cross-annotation face alignment. Pattern Recognition **154**, 110569 (2024)

[18] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision, pp. 483–499 (2016). Springer

[19] Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693–5703 (2019)

[20] Wang, X., Bo, L., Fuxin, L.: Adaptive wing loss for robust face alignment via heatmap regression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6971–6981 (2019)

[21] Kumar, A., Marks, T.K., Mou, W., Wang, Y., Jones, M., Cherian, A., Koike-Akino, T., Liu, X., Feng, C.: Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In: Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition, pp. 8236–8246 (2020)
[22] Huang, Y., Yang, H., Li, C., Kim, J., Wei, F.: Adnet: Leveraging error-bias towards normal direction in face alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3080–3090 (2021)
[23] Lan, X., Hu, Q., Cheng, J.: Atf: An alternating training framework for weakly supervised face alignment. IEEE Transactions on Multimedia **25**, 1798–1809 (2022)
[24] Narayan, K., VS, V., Chellappa, R., Patel, V.M.: Facexformer: A unified transformer for facial analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11369–11382 (2025)
[25] Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR, pp. 7482–7491 (2018)
[26] Chen, Z., Badrinarayanan, V., Lee, C.-Y., Rabinovich, A.: Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: ICML, vol. 80, pp. 794–803 (2018)
[27] Liu, S., Johns, E., Davison, A.J.: End-to-end multi-task learning with attention. In: CVPR, pp. 1871–1880 (2019)
[28] Chen, X., Li, Z., Pu, Y., *et al.*: A comparative study of image restoration networks for general backbone network design. In: ECCV, pp. 74–91. Springer, Cham (2024)
[29] Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D.: Carafe: Content-aware reassembly of features. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 3007–3016 (2019)
[30] Zhu, M., Shi, D., Zheng, M., Sadiq, M.: Robust facial landmark detection via occlusion-adaptive deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3486–3496 (2019)
[31] Dapogny, A., Bailly, K., Cord, M.: Decafa: Deep convolutional cascade for face alignment in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6893–6901 (2019)
[32] Zhu, C., Li, X., Li, J., Dai, S.: Improving robustness of facial landmark detection by defending against adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11751–11760 (2021)
[33] Jin, H., Liao, S., Shao, L.: Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. International Journal of Computer Vision **129**(12), 3174–3194 (2021)
[34] Zhu, C., Wan, X., Xie, S., Li, X., Gu, Y.: Occlusion-robust face alignment using a viewpoint-invariant hierarchical network architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11112–11121 (2022)
[35] Li, H., Guo, Z., Rhee, S.-M., Han, S., Han, J.-J.: Towards accurate facial landmark detection via cascaded transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4176–4185 (2022)
[36] Kato, S., Hotta, K., Hatakeyama, Y., Konishi, Y.: Lite-hrnet plus: Fast and accurate facial landmark detection. In: 2023 IEEE International Conference on Image Processing (ICIP), pp. 1500–1504 (2023). IEEE
[37] Chiang, J.-C., Hu, H.-N., Hou, B.-S., Tseng, C.-Y., Liu, Y.-L., Chen, M.-H., Lin, Y.-Y.: Orformer: occlusion-robust transformer for accurate facial landmark detection. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 784–793 (2025). IEEE
[38] Feng, Z.-H., Kittler, J., Awais, M., Huber, P., Wu, X.-J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2235–2245 (2018)
[39] Wan, J., Lai, Z., Li, J., Zhou, J., Gao, C.: Robust facial landmark detection by multi-order multiconstraint deep networks. IEEE Transactions on Neural Networks and Learning Systems **33**(5), 2181–2194 (2021)
[40] Zhou, Z., Li, H., Liu, H., Wang, N., Yu, G., Ji, R.: Star loss: Reducing semantic ambiguity in facial landmark detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15475–15484 (2023)
[41] Xia, J., Xu, M., Zhang, H., Zhang, J., Huang, W., Cao, H., Wen, S.: Robust face alignment via inherent relation learning and uncertainty estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 10358–10375 (2023)
[42] Gao, P., Lu, K., Xue, J., Lyu, J., Shao, L.: A facial landmark detection method based on deep knowledge transfer. IEEE Transactions on Neural Networks and Learning Systems **34**(3), 1342–1353 (2021)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ArticleTitle21.zip