

# Topology-Guided Semantic Face Center Estimation for Rotation-Invariant Face Detection

Hathai Kaewkorn, Lifang Zhou, Weisheng Li, Chengjiang Long, *Member, IEEE*

**Abstract**—Face detection accuracy significantly decreases under rotational variations, including in-plane (RIP) and out-of-plane (ROP) rotations. ROP is particularly problematic due to its impact on landmark distortion, which leads to inaccurate face center localization. Meanwhile, many existing rotation-invariant models are primarily designed to handle RIP, they often fail under ROP because they lack the ability to capture semantic and topological relationships. Moreover, existing datasets frequently suffer from unreliable landmark annotations caused by imperfect ground truth labeling, the absence of precise center annotations, and imbalanced data across different rotation angles. To address these challenges, we propose a topology-guided semantic face center estimation method that leverages graph-based landmark relationships to preserve structural integrity under both RIP and ROP. Additionally, we construct a rotation-aware face dataset with accurate face center annotations and balanced rotational diversity to support training under extreme pose conditions. Next, we introduce a Hybrid-ViT model that fuses CNN spatial features with transformer-based global context and employ a center-guided module for robust landmark localization under extreme rotations. In order to evaluate center quality, we further design a hybrid metric that combines topological geometry with semantic perception for a more comprehensive evaluation of face center accuracy. Finally, experimental results demonstrate that our method outperforms state-of-the-art models in cross-dataset evaluations. Code: [https://github.com/Catster111/TCE\\_RIFD](https://github.com/Catster111/TCE_RIFD).

**Index Terms**—Rotation-Invariant Face Detection, Face Center Estimation, Rotation-Aware Dataset, Center Evaluation Metric

## I. INTRODUCTION

**F**ACE detection is a fundamental task in computer vision, which supports a wide range of applications such as surveillance, biometric authentication, mobile face unlocking, augmented reality, and multi-angle video conferencing. In controlled environments, modern CNN-based detectors such as RetinaFace [1], YOLO-FaceV2 [2], SCRFD [3], YuNet [4] achieve high accuracy. However, real-world scenarios introduce substantial challenges due to the frequent appearance of faces at diverse orientations, caused by variations in camera

This research was supported in part by the National Natural Science Foundation of China (62276039 and 62331008), the Guizhou University Key Laboratory of Advanced Manufacturing Technology MOE (KY[2022]376), and the Chongqing Science Health Joint Medical Scientific Research Project (2024ZDXM022). (Corresponding author: Lifang Zhou.)

Lifang Zhou is with the Key Laboratory of Image Cognition and the College of Software, Chongqing University of Posts and Telecommunications, Chongqing 400065, China, and also with the KeyLaboratory of Advanced Manufacturing Technology, Ministry of Education, Guizhou University, Guizhou 550025, China (e-mail: zhoulf@cqupt.edu.cn).

Hathai Kaewkorn and Weisheng Li are with the Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: l201910012@stu.cqupt.edu.cn; liws@cqupt.edu.cn).

Chengjiang Long is with the META Reality Labs, Burlingame, CA 94010 USA (e-mail: cjfykx@gmail.com).

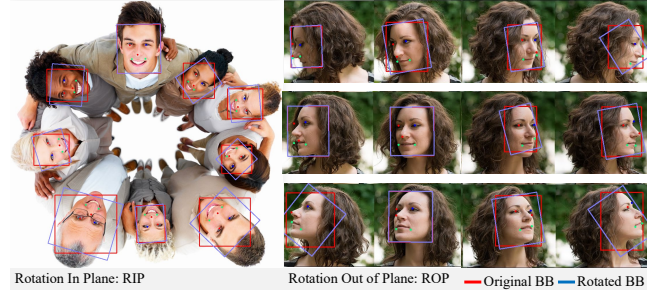


Fig. 1. Face detection under rotational variations. Left: Examples of RIP, where faces are rotated within the image plane. Right: Examples of ROP, where faces are rotated in 3D space. The red (upright) and blue (rotated) bounding boxes highlight the limitations of standard detection methods under rotation. Accurate orientation estimation is crucial for rotation-invariant multimedia understanding, robust pose estimation, expression analysis, and facial alignment in unconstrained settings. RIFD addresses this by jointly detecting faces and predicting their rotation angles.

angle, user posture, and occlusion. This leads to the critical problem of Rotation-Invariant Face Detection (RIFD), which aims to robustly detect and localize faces under non-frontal and unconstrained conditions [5].

Rotational variations in face detection can be broadly categorized into in-plane rotation (RIP) and out-of-plane rotation (ROP), as illustrated in Fig. 1. RIP refers to lateral tilts within the image plane and primarily affects spatial alignment. It is often addressed through data augmentation or orientation-aware CNN modules [7]. In contrast, ROP involves complex 3D transformations such as yaw, pitch, and roll, which introduce self-occlusion, perspective distortion and significant landmark deformations [6]. These factors severely impair landmark-based face center estimation, which is essential for accurate detection, alignment, and high-level semantic analysis. Conventional methods relying on fixed landmark priors or purely geometric computations are particularly unreliable under extensive ROP conditions.

Recent RIFD methods adopt diverse strategies to enhance robustness. Methods such as MTCNN [8], PCN [5], ASCN [9], and MTCNet [10] leverage cascaded architectures, angle prediction, or spatial normalization to mitigate RIP effects. Despite these improvements, existing methods remain constrained by rigid 2D spatial alignments, which heavily rely on data augmentation, and also suffer from error propagation across cascaded stages. In contrast, handling ROP remains more challenging due to its severe impact on landmark deformation. RP-Net [7] addresses this by converting rotation into translation through polar transformation. However, the effectiveness of this method strongly depends on accurate face center estimation. Even minor localization errors can cause

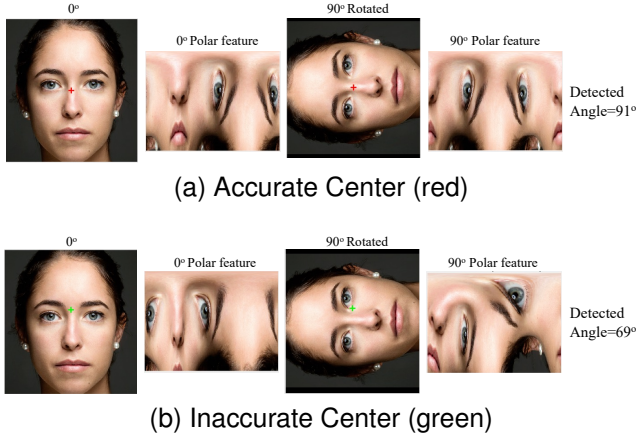


Fig. 2. Demonstrates the importance of accurate face center estimation for consistent feature representation in the polar domain. (a) With precise center estimation, the rotation angle is detected accurately ( $91^\circ$ ). (b) Inaccurate center estimation leads to feature misalignment and errors in the detected rotation angle ( $69^\circ$ ).

severe feature misalignment and fragmentation in the polar domain, as demonstrated in Fig. 2.

Accurate face center estimation is crucial for bounding box construction, feature normalization, and spatial transformations. Traditional methods often approximate the center by averaging landmark positions, which works reasonably well under frontal conditions. However, under ROP, 3D deformations and self-occlusions distort landmark geometry, which results in the estimated center to displace from its true semantic location [11]. This misalignment not only degrades detection accuracy but also compromises the alignment of polar or spatially normalized inputs. As a result, the reliability of subsequent landmark localization is significantly reduced. To address these issues, it is essential to move beyond simple geometric averaging. Instead, we need semantically informed, topologically robust methods that can consistently handle diverse rotation scenarios.

The development and evaluation of RIFD methods inevitably suffer from bottlenecks due to dataset limitations. Accurate face center annotations are especially critical because they directly affect the reliability of face detection and localization under rotation. Common datasets, such as 300W [12], AFLW [13], and AFLW2000-3D [14], primarily focus on frontal or near-frontal faces. These datasets often rely on synthetic augmentations, which may introduce artifacts and lack sufficiently detailed annotations for extreme head poses. Moreover, most datasets approximate face centers using bounding box geometry rather than semantically meaningful references. This undermines the reliability of center-based methods [15] and complicates fair comparison across approaches. To address this shortcoming, we propose a new rotation-aware dataset that provides accurate face center annotations and extensive coverage of challenging poses.

To overcome these limitations, we introduce a comprehensive approach that addresses critical issues related to face center estimation, dataset inadequacies, and evaluation metrics. Our approach not only improves center localization under extreme poses but also enhances downstream detection and

alignment. Our main contributions are summarized as follows:

- We propose a Topology-based Center Estimation (TCE) method that identifies semantically consistent face centers using stable graph-based landmark relationships. This method provides ground-truth for face center annotations in our new dataset.
- We present the Rotation-Aware Face (RAF) dataset, the first to offer high-quality center annotations with comprehensive  $360^\circ$  rotation coverage. Our augmentation strategy enriches rotational variations while preserving the natural relationship between the face and its background.
- We design Face Center Quality Assessment (FCQA), a hybrid metric that evaluates both structural and feature-level consistency. This ensures that face-center predictions maintain both the expected geometry and visual coherence under rotation.
- We develop a Hybrid-ViT model that integrates CNN and transformer components, which is supported by a center-guided module to ensure accurate polar transformation and improved landmark localization.

## II. RELATED WORK

### A. Face Detection

Face detection is a fundamental task in computer vision and serves as the first step for many face-related applications, including face recognition, alignment, tracking, and analysis. CNN-based approaches have significantly advanced face detection accuracy in recent years and can be broadly categorized into anchor-based and anchor-free paradigms [16].

Anchor-based methods leverage multi-scale anchor grids and feature pyramids to localize faces. S3FD [18] and DSFD [17] improved anchor design and matching strategies for better scale coverage. Subsequent works, such as ASFD [19] and SRN [20], adopt neural architecture search and optimized sample balancing. RetinaFace [1] enhanced this paradigm by incorporating landmark supervision to assist with small and occluded face detection. SCRFD [3] investigate positive sample redistribution and computation allocation across scales and network components under a fixed budget. Despite their high accuracy, these detectors rely on fixed anchor templates, which limit flexibility under large pose variations. In contrast, anchor-free methods simplify detection by directly predicting face centers or bounding boxes. FCOS [22] and CenterFace [21] eliminate anchor dependencies and regress bounding boxes using heatmaps or offset predictions. More recently, lightweight detectors such as YuNet [4] have optimized backbone and neck design for deployment on resource-constrained devices. Although both paradigms perform well on benchmarks, most CNN-based detectors remain sensitive to head rotation and lack mechanisms for modeling orientation or consistent center points. This limits their effectiveness in RIFD.

### B. Rotation-Invariant Face Detection

To explicitly address rotation, several RIFD-specific methods have been proposed. PCN [5] adopts a coarse-to-fine

strategy to progressively narrow the search space for head rotation. ASCN [9] introduces angle regression modules to normalize in-plane rotation. MTCNet [10] extends these ideas with context-aware cascaded refinement. However, these approaches are limited in handling full 360-degree rotations, which rely on complex training pipelines, and often suffer from error propagation due to their multi-stage design. RP-Net [7] converts angular variation into translational shifts via polar transformation, which simplify the detection task. However, this approach depends heavily on accurate face center estimation, as even minor errors can distort polar inputs and lead to fragmented features and incorrect predictions.

*Transformer-based architectures* have demonstrated strong object detection capabilities by leveraging self-attention and long-range dependency modeling [23]. Several efforts have adapted transformers for rotation handling, including FPN-Former [24], RotaTR [25], and RotTrans [26], which introduce rotation-invariant semantics, dynamic anchors, and feature-level invariance constraints, respectively. Although these methods are effective for general objects, they still overlook facial structural priors. Transformer-based face detection under rotation remains underexplored, particularly in modeling landmark topology for robust center estimation. To address this gap, we build on Vision Transformers (ViTs) [27], which naturally encode global spatial relationships and holistic structures. Unlike CNNs, ViTs are better able to accommodate non-local distortions, which are induced by extreme head poses.

*Dataset limitations* remain a key challenge to progress in RIFD. Publicly available datasets primarily emphasize frontal or near-frontal faces and lack precise annotations for semantic face centers. 300W [12] consists mostly of well-aligned facial images with 68-point landmarks, but offers limited pose diversity and excludes extreme orientations. AFLW [13] includes a broader range of head poses, but its sparse 21-point annotations are insufficient for detailed topological reasoning. AFLW2000-3D [14] builds upon AFLW with 3D landmark regression, yet suffers from coarse alignment inaccuracies under occlusion or profile views. Synthetic augmentations are commonly used to simulate rotation, but they introduce artifacts from cropping and padding, which can degrade detection quality [26]. Moreover, Transformer-based models require large-scale and balanced datasets to fully leverage their long-range modeling capabilities. Providing this kind of methods is trained on relatively small or imbalanced datasets, they tend to underperform due to overfitting and limited pose diversity [27]. To improve model generalization under real-world variations, we construct a dataset specifically designed to address limitations in annotation quality and pose diversity. This dataset serves as a more robust foundation for rotation-invariant learning.

### III. PROPOSED METHOD

We propose a topology-aware face detection framework designed to robustly address rotation-induced variances through a semantic center estimation approach (see Fig. 3). The core idea integrates Topology-guided Center Estimation (TCE) to determine stable semantic face centers and a Hybrid Vision

Transformer (Hybrid-ViT) trained on the Rotation-Aware Face (RAF) dataset for visual center prediction. The Face Center Quality Assessment (FCQA) metric ensures structural consistency and precise spatial alignment. Semantic centers then serve as anchors for a Polar Transformation Module (PT-Module) to enhance landmark localization accuracy under arbitrary rotations.

#### A. Topology-Guided Face Center Estimation

Reliable face localization under both RIP and ROP rotations requires a robust semantic reference point. Unlike simple geometric centers, which are prone to distortion from head rotation and occlusion, semantic centers provide greater stability and alignment consistency. To address this, we propose a topology-aware framework that integrates graph-based landmark reasoning with a hybrid deep learning model.

##### 1) Landmark Connectivity Graph:

a) *Notation and weights.*: Let

$$\ell = \{V_{LE}, V_{RE}, V_N, V_{LL}, V_{RL}\}$$

denote the set of facial landmarks (left eye, right eye, nose, left lip, right lip). We construct a connectivity graph

$$G = (V, E, w),$$

where  $V = \ell$  are the vertices and  $E$  the anatomical connections:

$$E = \{(V_{LE}, V_N), (V_{RE}, V_N), (V_{LL}, V_N), (V_{RL}, V_N), \\ (V_{LE}, V_{RE}), (V_{LL}, V_{RL})\}.$$

For each edge  $(i, j) \in E$ , the weight is defined as

$$w_{ij} = \|V_i - V_j\|_2.$$

All-pairs shortest-path distances on  $G$  encode topological relationships among landmarks (see Fig. 4).

b) *Missing or occluded landmarks*: If a landmark is missing or occluded, its corresponding node is removed from  $V$ , and the edge set  $E$  and weights  $w$  are restricted accordingly. Shortest-path distances are then computed on the largest remaining connected subgraph, which preserves the topological relationships under partial annotations.

2) *Semantic Center Definition*: Let  $S \subseteq V$  denote the subset of key landmarks used to define the semantic center. We estimate the semantic center  $c_s \in \mathbb{R}^2$  as the weighted geometric median:

$$c_s = \arg \min_{c \in \mathbb{R}^2} \sum_{i \in S} w_i \|c - V_i\|_2, \quad (1)$$

where each  $w_i$  encodes the relative importance of landmark  $V_i$ . In practice,  $w_i$  is initialized using topology-aware weighting derived from the Landmark Connectivity Graph.

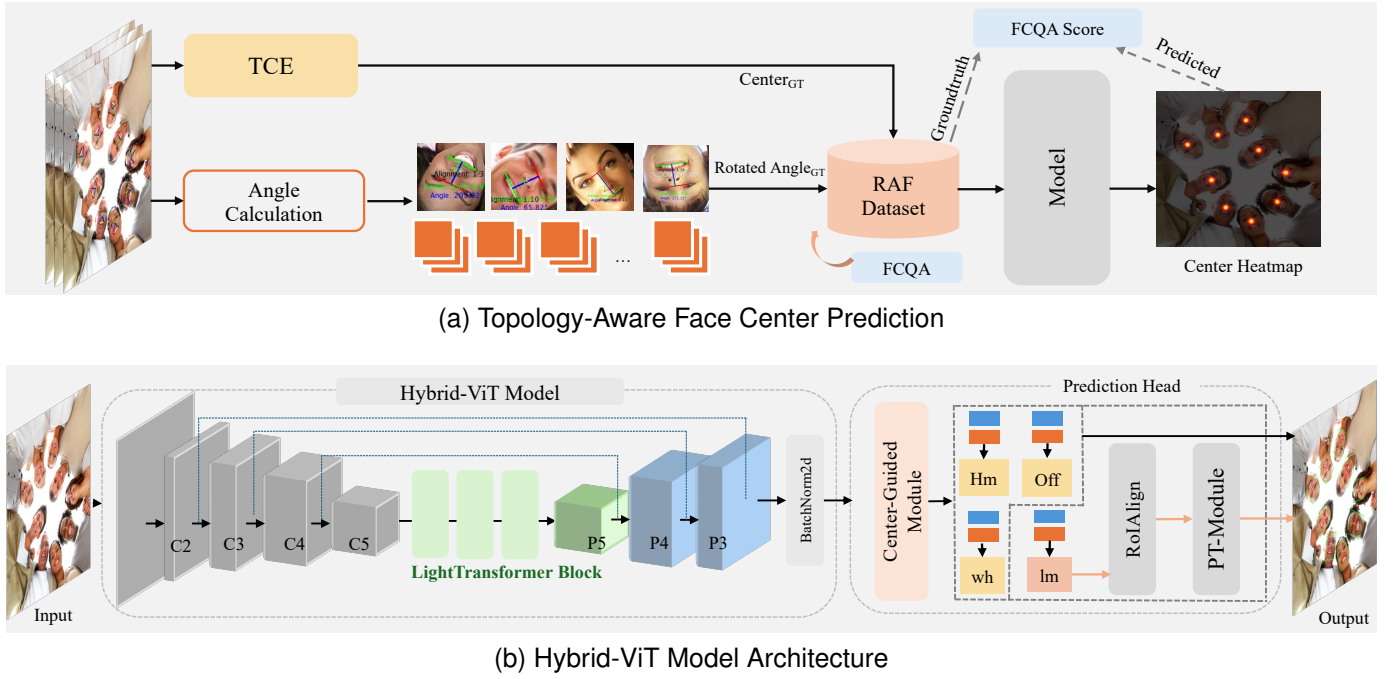


Fig. 3. Illustration of the proposed system. (a) The TCE module produces semantically aligned face centers from five landmarks; these centers, together with RIP angles, are used to construct the RAF dataset. FCQA assesses the consistency of predicted centers. (b) The Hybrid-ViT pipeline combines CNN-based local feature extraction with transformer-based global reasoning (LightTransformer), followed by center localization and CGM-based refinement.

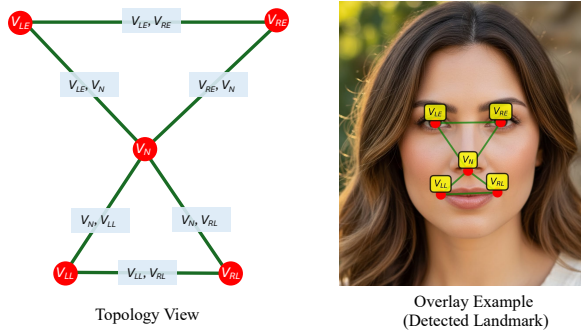


Fig. 4. Landmark Connectivity Graph. Left: a topological view with facial landmarks represented as nodes and anatomical connections as edges. Right: an example overlay on a detected face that demonstrates the graph's encoding of facial structure for topology-based center estimation.

3) *Topology-Guided Center Estimation (TCE)*: The TCE algorithm produces stable, semantically aligned facial centers by leveraging the graph structure defined above. Each node  $V_i \in \ell$  corresponds to a 2D coordinate, and edges reflect meaningful anatomical relationships weighted by Euclidean distance.

To capture both direct and indirect relationships, we compute the all-pairs shortest-path matrix  $D$  using the Floyd–Warshall algorithm [28]. The geometric median is initialized at the centroid of all landmarks and refined iteratively. At each iteration, the influence of each landmark is weighted inversely by its distance to the current estimate, which reduces the effect of outliers. Iterations continue until convergence or until the maximum iteration threshold is reached.

This topology-guided estimation yields rotation-invariant,

#### Algorithm 1 Topology-Guided Center Estimation (TCE)

- 1: **Input:** Landmark coordinates  $L = \{V_1, \dots, V_n\}$ , tolerance  $\epsilon$ , maximum iterations  $\text{max\_iters}$ , graph  $G = (V, E)$
- 2: **Output:** Geometric median  $g = (c_x, c_y)$
- 3: Initialize distance matrix  $D$  for  $(i, j) \in E$  using Euclidean distances
- 4: Compute all-pairs shortest-path distances  $D_{ij}$  using Floyd–Warshall
- 5: Initialize  $g^{(0)}$  as the centroid of  $L$ ; set  $t = 0$
- 6: **repeat**
- 7:   **Compute weights:**  $w_i = \frac{1}{\|g^{(t)} - V_i\|_2 + \epsilon}$  (direct Euclidean; not a lookup in  $D$ )
- 8:   **Update:**  $g^{(t+1)} = \frac{\sum_{i \in V} w_i V_i}{\sum_{i \in V} w_i}$
- 9:   **Check:** if  $\|g^{(t+1)} - g^{(t)}\|_2 < \epsilon$  then break
- 10:    $t \leftarrow t + 1$
- 11: **until** convergence or  $t \geq \text{max\_iters}$
- 12: **Return:**  $g = g^{(t)}$

*Note.*  $D_{ij}$  represents landmark–landmark geodesic distances used only for topology. The Euclidean distance  $\|g^{(t)} - V_i\|_2$  in Step 7 is computed directly since  $g^{(t)} \notin L$ .

semantically consistent facial centers that improve robustness for RAF annotation and provide reliable supervision for Hybrid-ViT training.

4) *Hybrid Vision Transformer (Hybrid-ViT)*: Figure 3(b) shows that Hybrid-ViT integrates fine-grained spatial features with global self-attention to estimate semantic face centers. The LightTransformer block (Fig. 5) consists of two parallel branches, Multi-Head Self-Attention (MSA) and Efficient

Channel Attention (ECA), each employing a residual connection. Their outputs are fused by element-wise addition, which is followed by a lightweight MLP. This design balances global context modeling and channel-wise reweighting with minimal computational overhead.

a) *CNN Backbone*: Given an input image  $I \in \mathbb{R}^{H \times W \times 3}$ , local feature maps are extracted by MobileNetV4:

$$F_c = f_{\text{CNN}}(I), \quad F_c \in \mathbb{R}^{H' \times W' \times C}, \quad (2)$$

where  $f_{\text{CNN}}$  denotes the backbone network [29].

b) *Tokenization*: The convolutional feature map is reshaped into  $N = H'W'$  tokens:

$$X = \text{reshape}(F_c), \quad X \in \mathbb{R}^{N \times C}. \quad (3)$$

c) *MSA Branch*: Following the Transformer paradigm [23], we normalize and project the token sequence:

$$\begin{aligned} \tilde{X} &= \text{LN}(X), \\ Q_h &= \tilde{X}W_K^{(h)}, \quad K_h = \tilde{X}W_K^{(h)}, \quad V_h = \tilde{X}W_V^{(h)} \in \mathbb{R}^{N \times d}. \end{aligned} \quad (4)$$

Each attention head is computed as

$$\text{head}_h = \text{Softmax}\left(\frac{Q_h K_h^T}{\sqrt{d}}\right) V_h, \quad (5)$$

and concatenated into

$$A = [\text{head}_1 \parallel \dots \parallel \text{head}_H] W_O, \quad A \in \mathbb{R}^{N \times C}. \quad (6)$$

d) *ECA Branch*: To enhance channel selectivity with minimal cost, ECA [30] applies global average pooling (GAP) followed by a 1-D convolution:

$$z = \text{GAP}(F_c), \quad t = \sigma(\text{Conv1D}(z)), \quad t \in \mathbb{R}^C. \quad (7)$$

The reweighted feature map and its projected token form are

$$F_{\text{ECA}} = F_c \otimes t, \quad E = \text{proj}(\text{reshape}(F_{\text{ECA}})), \quad E \in \mathbb{R}^{N \times C}, \quad (8)$$

where  $\otimes$  denotes channel-wise multiplication.

e) *Fusion and MLP*: The outputs of the MSA branch, ECA branch, and skip connection are combined element-wise:

$$U = X + A + E, \quad U \in \mathbb{R}^{N \times C}. \quad (9)$$

The fused representation is then normalized and passed through a two-layer MLP:

$$T_{\text{out}} = \text{MLP}(\text{LN}(U)), \quad T_{\text{out}} \in \mathbb{R}^{N \times C}. \quad (10)$$

$T_{\text{out}}$  serves as the output of the LightTransformer block and the input to the next stage (as illustrated in Fig. 5).

f) *Center Heatmap Prediction*: The transformer output is projected into a spatial heatmap:

$$H = \text{reshape}(T_{\text{out}} W_h), \quad H \in \mathbb{R}^{H' \times W'} \quad (11)$$

g) *Center Localization*: We recover the continuous face center  $\hat{c} = (\hat{x}, \hat{y})$  using the soft-argmax operator:

$$\hat{c} = \sum_{i=1}^{H'} \sum_{j=1}^{W'} \frac{\exp(H_{i,j})}{\sum_{p=1}^{H'} \sum_{q=1}^{W'} \exp(H_{p,q})} (i, j) \quad (12)$$

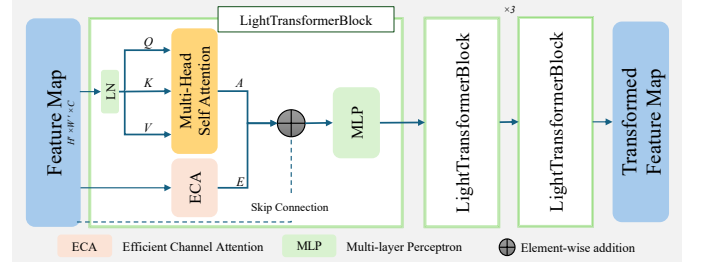


Fig. 5. The proposed LightTransformer architecture enhances feature representation by integrating ECA and a MLP within each block. Element-wise fusion and skip connections facilitate efficient learning of both channel-specific and non-linear transformations. Stacked blocks progressively refine the input feature map.

5) *Center-Guided Module (CGM)*: To refine the predicted center and ensure geometric consistency, the Center-Guided Module (CGM) combines the heatmap center  $\hat{c}$  with the topology-guided estimate  $c^{\text{TCE}}$ . The refined center is

$$c = \hat{c} + \alpha (c^{\text{TCE}} - \hat{c}), \quad c = (x_c, y_c), \quad \alpha \in [0, 1]. \quad (13)$$

At inference,  $c^{\text{TCE}}$  is computed from the model's predicted landmarks (no GT is used). We parameterize  $\alpha$  (either a scalar or  $\text{sigmoid}(g(\cdot))$  from a small gating MLP) and learn it end to end, which implements precision-weighted fusion in practice and reduces variance under large rotations or occlusions.

From the region-of-interest (RoI) features centered at  $c$ , a regression head predicts log-scale parameters:

$$(r_w, r_h) = \psi(\text{RoI}(F_c, c)), \quad (w, h) = s \cdot (e^{r_w}, e^{r_h}), \quad (14)$$

where  $s$  is the feature stride. The final bounding box is then

$$B = \left( x_c - \frac{w}{2}, y_c - \frac{h}{2}, x_c + \frac{w}{2}, y_c + \frac{h}{2} \right). \quad (15)$$

This refinement keeps detection boxes semantically aligned with the TCE-derived center and geometrically consistent with the underlying features, particularly under large pose variations.

6) *Additional Implementation Modules*: For completeness, we explicitly define the remaining components used in Fig. 3. These modules were previously only depicted but are now described in detail for clarity.

**BatchNorm2d (BN2d)**. Each convolutional block includes BN2d and a nonlinearity. Given channel  $k$  with batch-spatial mean  $\mu_k$  and variance  $\sigma_k^2$ , BN2d normalizes and rescales:

$$\text{BN2d}(x_k) = \gamma_k \frac{x_k - \mu_k}{\sqrt{\sigma_k^2 + \epsilon}} + \beta_k. \quad (16)$$

**Prediction Head**. From the LightTransformer output  $T_{\text{out}} \in \mathbb{R}^{N \times C}$  (with  $N = H'W'$ ), we predict a center heatmap and a size:

$$H = \text{reshape}(T_{\text{out}} W_h) \in \mathbb{R}^{H' \times W'}, \quad \hat{c} = \text{softargmax}(H), \quad (17)$$

$$(r_w, r_h) = \phi(T_{\text{out}} W_s), \quad (w, h) = s \cdot (e^{r_w}, e^{r_h}), \quad (18)$$

where  $s$  is the feature stride,  $W_h, W_s$  are learned projections, and  $\phi(\cdot)$  is a linear head producing the log-scales.

**RoIAlign.** Given feature map  $F_c$  and refined center  $c = (x_c, y_c)$ , define a square window  $b$  of side  $L$  centered at  $c$ . RoIAlign samples a fixed  $K \times K$  tensor by bilinear interpolation without quantization:

$$\text{RoIAlign}(F_c, b)[u, v] = \sum_{m,n} F_c[m, n] \omega_{uv}(m, n), \quad (19)$$

where  $[u, v] \in \{1, \dots, K\}^2$  are evenly spaced locations inside  $b$ , and  $\omega_{uv}(\cdot, \cdot)$  are bilinear weights. The pooled feature feeds the size head:

$$(r_w, r_h) = \psi(\text{RoIAlign}(F_c, b)), \quad (w, h) = s \cdot (e^{r_w}, e^{r_h}). \quad (20)$$

**PT-Module (Polar Transform).** Centered at  $c = (x_c, y_c)$ , we build a differentiable polar grid with radii  $\{\rho_m\}_{m=1}^M$  and angles  $\{\theta_t\}_{t=1}^T$ :

$$\rho = \sqrt{(x - x_c)^2 + (y - y_c)^2}, \quad \theta = \text{atan2}(y - y_c, x - x_c), \quad (21)$$

$$\mathcal{G}_{m,t} = (x_c + \rho_m \cos \theta_t, y_c + \rho_m \sin \theta_t). \quad (22)$$

$$\text{PT}(F_c, c)[m, t] = \sum_{i,j} F_c[i, j] w_{m,t}(i, j). \quad (23)$$

with bilinear weights  $w_{m,t}(i, j)$ . This resampling is fully differentiable and supports analysis in polar space.

### B. RAF: Rotation-Aware Face Dataset Construction

To effectively train rotation-robust face center detection models, we require datasets with precise center annotations and balanced angular pose distributions. WIDER FACE [21] provides a large number of images captured under diverse real-world conditions, which include variations in scale, occlusion, and background clutter. These characteristics make it a strong foundation for constructing a rotation-aware benchmark. Accordingly, we construct the RAF dataset by extending WIDER FACE with reliable face center annotations generated using the proposed TCE method. To ensure comprehensive rotational diversity, we further apply angular bin-based augmentation. Fig. 6 illustrates the creation process of the RAF dataset.

1) *TCE-Based Center Annotations and Pose Binning:* For each face, the ground-truth center  $c$  is computed *once offline* from five human-annotated landmarks using the deterministic, non-learned TCE rule. TCE builds a landmark connectivity graph and estimates the center from visible landmarks only. Missing or occluded nodes are removed, and shortest-path distances are computed on the largest remaining subgraph. The resulting centers are stored with the dataset and are not recomputed at train or test time, so labels do not depend on model predictions.

For pose analysis, we report performance as a function of RIP/roll. Results are aggregated over RIP angle ranges ( $0^\circ$ – $360^\circ$ ) to reveal behavior under extreme tilts.

To assess and balance angular pose distribution, we discretize the full orientation space  $\theta \in [0^\circ, 360^\circ]$  into  $K$  angular bins of width  $\beta = 36^\circ$ . Each bin  $b$  is defined as:

$$b = \left\lfloor \frac{\theta}{\beta} \right\rfloor, \quad b \in \{0, 1, \dots, K-1\}, \quad K = \frac{360}{\beta} \quad (24)$$

We construct a histogram  $H(b)$  over all faces to determine bin-wise sample distributions:

$$H(b) = \sum_{i=1}^N \Pi \left( b = \left\lfloor \frac{\theta_i}{\beta} \right\rfloor \right) \quad (25)$$

where  $N$  is the number of faces and  $\Pi$  is an indicator function. Bins with fewer than  $\tau \cdot \max(H)$  samples (with threshold  $\tau = 0.95$ ) are flagged for augmentation.

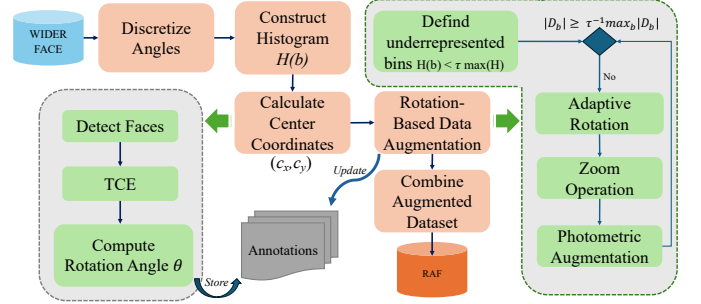


Fig. 6. RAF dataset creation pipeline, including TCE-based center annotation, angular bin analysis, and augmentation strategy

2) *Rotation and Photometric Augmentation:* For under-represented bins, we apply adaptive rotation and photometric transformations to synthetically balance angular diversity. Each face image is rotated about its TCE-derived center  $(c_x, c_y)$  by an angle  $\alpha$  within the bin range. We incorporate a scale factor  $z = 1.2$  to preserve facial details and avoid resolution loss. The transformation is computed using:

$$M_{\text{rot}} = \begin{pmatrix} z \cos \alpha & -z \sin \alpha & (1 - \cos \alpha)c_x + z \sin \alpha c_y \\ z \sin \alpha & z \cos \alpha & -z \sin \alpha c_x + (1 - \cos \alpha)c_y \end{pmatrix} \quad (26)$$

Photometric augmentations further diversify appearance and include:

- Brightness shift:  $\Delta \sim U(-51, 51)$
- Contrast scaling:  $\gamma \sim U(0.8, 1.2)$
- Gaussian blur: applied with  $3 \times 3$  or  $5 \times 5$  kernels

The augmented dataset  $D'$  is expressed as:

$$D' = D_{\text{orig}} \cup \bigcup_{b=0}^{K-1} A_b(D_b) \quad (27)$$

where  $A_b$  denotes the augmentation operator for bin  $b$ , applied until  $|D_b| \geq \tau^{-1} \max_b |D_b|$ .

3) *Pose-Balanced Angular Distribution:* To mitigate pose imbalance in existing benchmarks, we begin with the WIDER FACE training set (12,671 images, 80,143 annotations) and derive semantic face centers using TCE. We then quantize the full rotation range  $\theta \in [0^\circ, 360^\circ]$  into  $K = 10$  bins of width  $\beta = 36^\circ$  and compute the sample count  $H(b)$  in each bin. Bins containing fewer than  $\tau \max(H)$  samples (with  $\tau = 0.95$ ) are augmented via targeted rotations and photometric transformations. This process yields the RAF dataset with 24,532 images (109,647 faces), which achieves near-uniform coverage of all pose intervals, including extreme rotations ( $150^\circ$ – $210^\circ$ ) as illustrated in Fig. 7.

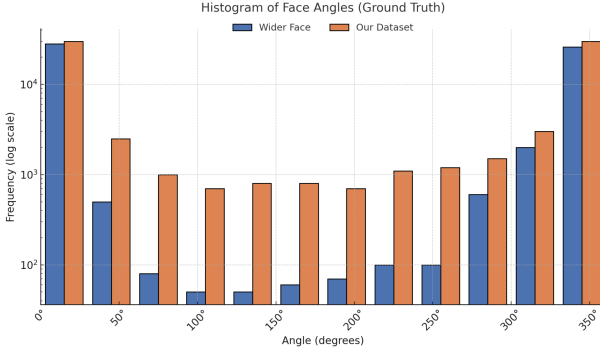


Fig. 7. Angle distribution comparison between RAF and the original WIDER FACE training subset [21].

### C. FCQA: Evaluating Face Center Quality

Our Hybrid-ViT model is trained on TCE-labeled centers within the RAF dataset. However, evaluating the quality of the predicted centers, especially under rotational variation, requires a reliable and rotation-aware metric. Traditional measures such as SSIM [31] and PSNR are not rotation-aware and often fail to capture structural misalignment in the polar domain. To address this, we propose a hybrid evaluation framework called FCQA, which combines feature-level and structural assessments tailored to polar-transformed face representations.

1) *Metric Design*: The FCQA metric is computed in two stages. First, we apply a polar transformation to the image centered on the predicted face center. Then, we extract feature embeddings through a pre-trained CNN and evaluate distributional and structural consistency.

a) *Polar Feature Consistency*: We adapt the Fréchet Inception Distance (FID) [32] to quantify feature alignment in the polar space. Given the feature embeddings of the polar-transformed predicted image ( $\mu_i^{\text{polar}}, \Sigma_i^{\text{polar}}$ ) and a reference distribution from real center-aligned faces ( $\mu_{\text{real}}^{\text{polar}}, \Sigma_{\text{real}}^{\text{polar}}$ ), the adapted FID score is:

$$FID_{\text{polar}} = \left\| \mu_{\text{real}}^{\text{polar}} - \mu_i^{\text{polar}} \right\|^2 + \text{Tr} \left( \Sigma_{\text{real}}^{\text{polar}} + \Sigma_i^{\text{polar}} - 2\sqrt{\Sigma_{\text{real}}^{\text{polar}} \Sigma_i^{\text{polar}}} \right) \quad (28)$$

b) *Structural Consistency*: Standard SSIM [31] does not account for circular shifts introduced by polar transformation. To address this, we define a circularly shifted SSIM metric  $SSIM_{\text{shift}}(\Delta\theta)$ , where the angular offset  $\Delta\theta$  maximizes alignment between the predicted and reference polar images:

$$SSIM_{\text{shift}}(\Delta\theta) = SSIM(I_{\text{polar}}, CShift(I_{\text{polar}_{\text{rt}}}, \Delta\theta)) \quad (29)$$

c) *Final Hybrid Metric*: The FCQA score integrates both components using a weighted sum. The final metric is defined as:

$$FCQA = \alpha \cdot SSIM(x, y) + \beta \cdot FID_{\text{polar}} \quad (30)$$

TABLE I  
SENSITIVITY TO CENTER OFFSETS AT 90° ROTATION

| Offset  | FCQA ↓ | SSIM <sub>Diff</sub> ↓ | SSIM <sub>Shift_Diff</sub> ↓ |
|---------|--------|------------------------|------------------------------|
| (0,0)   | 0.017  | 0.796                  | 0.036                        |
| (2,2)   | 0.194  | 0.794                  | 0.076                        |
| (5,5)   | 0.590  | 0.795                  | 0.185                        |
| (10,10) | 1.202  | 0.800                  | 0.540                        |
| (15,15) | 1.850  | 0.803                  | 0.600                        |
| (20,20) | 2.046  | 0.805                  | 0.640                        |

TABLE II  
SENSITIVITY TO ROTATION ANGLES AT OFFSET = (10,10)

| Rotation (°) | FCQA ↓ | SSIM <sub>Diff</sub> ↓ | CartFID <sub>Scaled</sub> ↓ |
|--------------|--------|------------------------|-----------------------------|
| 0            | 1.150  | 0.005                  | 0.000                       |
| 30           | 1.201  | 0.050                  | 0.120                       |
| 60           | 1.205  | 0.300                  | 0.350                       |
| 90           | 1.202  | 0.800                  | 0.820                       |
| 120          | 1.210  | 0.900                  | 0.600                       |
| 150          | 1.235  | 0.950                  | 0.880                       |
| 180          | 1.195  | 0.010                  | 0.000                       |

In our experiments, we use  $\alpha = 0.4$  and  $\beta = 0.6$ , which yield the lowest standard deviation across rotation and offset scenarios.

2) *Evaluating Metric Robustness*: We validate FCQA's robustness to center displacement and rotational changes by comparing it to traditional metrics in two stress-test settings: center offset and pose rotation.

a) *Sensitivity to Center Offsets*: Table I demonstrates that FCQA effectively captures center misalignment, with scores rising significantly as the offset increases. In contrast, traditional metrics such as  $SSIM_{\text{Diff}}$ ,  $SSIM_{\text{Shift_Diff}}$ , and  $PSNR_{\text{Inverse}}$  show limited sensitivity to such changes.

b) *Robustness to Rotation Variations*: Table II shows that FCQA provides stable and reliable evaluations across a wide range of face rotation angles. Even under extreme poses, it consistently reflects the structural similarity and feature alignment of the predictions. In comparison,  $SSIM_{\text{Diff}}$  and  $CartFID_{\text{Scaled}}$  vary significantly or are affected by symmetry, which makes their evaluations less dependable.

## IV. EXPERIMENTS

### A. Datasets

We employ widely adopted benchmarks to evaluate face detection and facial landmark localization under unconstrained and rotation-varying conditions.

1) *Face Detection Datasets*: These benchmarks assess detector performance across a spectrum of real-world conditions and include specialized rotations to test orientation robustness.

*WIDER FACE* [34] comprises 32,203 images with 393,703 annotated faces spanning 61 event categories. It exhibits extreme variability in scale, pose, illumination, expression, occlusion, and background clutter. We use its standard train-val-test split and additionally construct a rotated test subset to measure robustness under arbitrary in-plane and out-of-plane rotations.

*Multi-Oriented FDDB* [5] extends the FDDB benchmark [33] by including faces at more diverse orientations

and higher rotation angles. The original Fddb contains 2,845 images (5,171 faces) with elliptical annotations and provides a complementary evaluation of detector performance under unconstrained acquisition.

2) *Facial Landmark Datasets*: These datasets vary in landmark annotation density and complexity, which enable evaluation of localization accuracy under diverse facial appearances and distortions.

*MTFL* [35] provides approximately 12,000 images annotated with five key points (two eyes, nose tip, two mouth corners) alongside binary attributes (e.g., gender, smile, eye-glasses). Its multi-task setup facilitates concurrent study of landmark localization and auxiliary attribute prediction.

*CelebA* [36] contains over 200,000 celebrity images with 40 attribute labels and five landmark annotations per face. Its scale and appearance diversity make it well-suited for large-scale landmark training.

*WFLW* [43] includes 10,000 images densely annotated with 98 landmarks and segmented into six attribute-based subsets (pose, expression, illumination, makeup, occlusion, blur). This partitioning enables fine-grained analysis of algorithmic robustness across different real-world distortions and extreme poses. This dataset provides a comprehensive benchmark for evaluating state-of-the-art facial landmark detection (FLD) methods [44], [45].

### B. Implementation Details

Our network was trained on input images resized to 512x512 pixels, with output feature maps of 128x128 pixels. Training was performed using batch size of 4 and an initial learning rate of 0.0005, with scheduled learning rate decay at epochs 60, 70, and 100. To enhance model robustness and prevent overfitting, we employed data augmentation techniques including random horizontal flipping with a probability of 0.5, scale augmentation with a factor of 1.6, and translation shift of 0.1. Notably, we use low factor of 0.1 augmentation due to RAF dataset contributions. The model checkpoints were preserved throughout the training process, with the final model weights. Training was parallelized across 4 workers to optimize data loading and preprocessing effectively. All experiments were conducted using PyTorch 2.1.1 with CUDA 12.1 and cuDNN 8902 support on a single RTX 3090Ti GPU.

### C. Evaluation Metrics

Common evaluation metrics for face detection include Recall rate, which measures a model's capability to detect true positives out of all actual positive cases [5], and Normalized Mean Error (NME), which is commonly used for evaluating facial landmark localization accuracy. Additionally, FCQA specifically measures the precision of predicted face centers relative to ground truth and emphasizes spatial accuracy.

NME quantifies facial landmark localization accuracy by calculating the average Euclidean distance between predicted and ground truth landmark positions. Normalizing by the Inter-Ocular Distance (IOD) [46] ensures robustness to variations in face scale, pose, and image resolution, which enables consistent comparisons across diverse datasets and conditions.

Formally, NME is computed as:

$$\text{NME} = \frac{1}{M} \sum_{j=1}^M \frac{\|L_{p,j} - L_{gt,j}\|_2}{D_{\text{IOD}}}. \quad (31)$$

where:

- $M$  is the total number of landmark points.
- $L_{p,j}$  and  $L_{gt,j}$  represent the predicted and ground truth coordinates of the  $j^{\text{th}}$  landmark, respectively.
- $\|\cdot\|_2$  denotes the Euclidean distance.

The IOD normalization factor is defined as:

$$D_{\text{IOD}} = \|L_{gt,eye\_right} - L_{gt,eye\_left}\|_2, \quad (32)$$

where  $L_{gt,eye\_left}$  and  $L_{gt,eye\_right}$  are the ground truth coordinates of the left and right eye landmarks, respectively.

### D. Results and Analysis

1) *Face Detection Under Rotation*: We evaluate our method on the Multi-Oriented Fddb benchmark, which reports recall at 100 false positives for four extreme head poses (Up, Down, Left, Right). As Table III shows, our Hybrid-ViT+CGM model achieves an average recall of 91.5%, a 0.4 percentage-point improvement over the previous best (RP-Net). Moreover, it maintains consistent performance across all rotations: 91.8% (Up), 92.1% (Down), 91.0% (Left), and 90.9% (Right), whereas classical CNN-based RIFD detectors suffer larger drops under extreme angles. These results confirm that our design preserves strong detection accuracy even under severe rotations.

The next section will show that landmark localization benefits even more dramatically from our topology-guided and rotation-aware approach.

2) *Landmark Localization Across Datasets*: To demonstrate the effectiveness of our approach, we compare it to seven state-of-the-art detectors MTCNN [8], RetinaFace [1], MediaPipe [40], SCRFD [3], YuNet [4], YOLOv8-Face [39], and RP-Net [7] using NME measured at six head-rotation angles.

a) *Comparisons on MTFL and CelebA*: Table IV reports per-angle NME on MTFL and CelebA. We group test images by rotation into subsets at 90°, 120°, 150°, 180°, 210°, and 240°.

On MTFL, our method achieves the lowest average NME across angles ( $0.84 \pm 0.01$ ) with notably low variance. While RetinaFace and SCRFD are best at 90° and 120° respectively (Table IV), our curve remains competitive across the remaining angles, which yields a lower overall mean. For reference, RP-Net records  $0.87 \pm 0.14$ , RetinaFace  $0.89 \pm 0.14$ , and SCRFD  $0.97 \pm 0.19$ .

On CelebA, our method achieves the lowest average NME across angles ( $0.25 \pm 0.01$ ) with low variance and leads at each angle (e.g., 0.23 at 180° and 0.24 at 240°). For reference, RP-Net records  $0.83 \pm 0.25$ , and RetinaFace  $0.85 \pm 0.27$ .

These results demonstrate that combining topology-guided semantic center estimation with RAF dataset augmentation via binning enriches the training set. This increased diversity enables our model to learn more robust features, which results in high accuracy and consistent performance under extreme head rotations.

TABLE III  
ROTATION-INVARIANT FACE-DETECTION RECALL ON Fddb (RECALL@100 FP), MODEL SIZE AND PARAMETER COUNT.

| Method                  | Venue (Year)    | Recall (%)  |             |             |             |             | Size       |            |
|-------------------------|-----------------|-------------|-------------|-------------|-------------|-------------|------------|------------|
|                         |                 | Up          | Down        | Left        | Right       | Avg.        | Model (MB) | Params (M) |
| Prior Methods           |                 |             |             |             |             |             |            |            |
| MTCNN-aug [8]           | CVPR (2016)     | 86.2        | 84.1        | 83.6        | 83.5        | 84.4        | 8.8        | 2.2        |
| SSD300 [42]             | ECCV (2016)     | 86.7        | 86.5        | 85.5        | 85.4        | 86.2        | 105        | 26.3       |
| Faster R-CNN [41]       | TPAMI (2016)    | 87.6        | 87.3        | 86.9        | 86.7        | 87.1        | 167        | 41.7       |
| PCN [5]                 | CVPR (2018)     | 88.5        | 88.1        | 87.7        | 87.5        | 87.9        | 16.8       | 4.2        |
| ASCN [9]                | CVPR (2019)     | 89.3        | 88.2        | 88.3        | 88.2        | 88.5        | 10.0       | 2.5        |
| MTCNet [10]             | PR (2022)       | 89.7        | 88.9        | 88.7        | 88.4        | 88.9        | 21.6       | 5.4        |
| Xiong et al. [37]       | J. Phys. (2023) | <b>92.2</b> | 71.6        | 75.5        | 75.3        | 78.7        | 151        | 37.7       |
| RP-Net [7]              | PR (2025)       | 91.5        | 91.2        | <b>91.8</b> | 90.0        | 91.1        | 29.6       | 7.4        |
| Our Proposed            |                 |             |             |             |             |             |            |            |
| Ours (MobileNetV4)      | —               | 89.8        | 88.0        | 90.5        | 89.0        | 89.3        | 7.6        | 1.9        |
| Ours (Hybrid-ViT)       | —               | 91.2        | 90.8        | 90.8        | 90.2        | 90.8        | 62.5       | 16.4       |
| Ours (Hybrid-ViT + CGM) | —               | 91.8        | <b>92.1</b> | 91.0        | <b>90.9</b> | <b>91.5</b> | 62.5       | 16.4       |

TABLE IV  
LANDMARK LOCALIZATION ACROSS DATASETS (MTFL, CELEBA). NME AT VARIOUS HEAD ROTATION ANGLES. LOWER IS BETTER. BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

| Dataset | Model            | Venue (Year)  | @90° | @120° | @150° | @180° | @210° | @240° | Avg  | Std Dev |
|---------|------------------|---------------|------|-------|-------|-------|-------|-------|------|---------|
| MTFL    | MTCNN [8]        | CVPR (2016)   | 0.73 | 0.93  | 1.03  | 1.14  | 1.00  | 0.84  | 0.95 | 0.12    |
|         | RetinaFace [1]   | CVPR (2020)   | 0.61 | 0.87  | 0.96  | 1.10  | 0.94  | 0.86  | 0.89 | 0.14    |
|         | MediaPipe [40]   | Google (2020) | 0.73 | 1.06  | 1.04  | 1.18  | 1.14  | 1.02  | 1.03 | 0.13    |
|         | SCRFD [3]        | ICLR (2022)   | 0.82 | 0.81  | 1.02  | 1.39  | 0.97  | 0.82  | 0.97 | 0.19    |
|         | YuNet [4]        | MIR (2023)    | 1.06 | 1.09  | 1.18  | 1.13  | 1.05  | 0.87  | 1.06 | 0.09    |
|         | YOLOv8-Face [39] | ADICS (2024)  | 0.96 | 1.10  | 1.19  | 1.33  | 1.14  | 0.99  | 1.12 | 0.11    |
|         | RP-Net [7]       | PR (2025)     | 0.89 | 0.90  | 0.94  | 1.05  | 0.92  | 0.84  | 0.87 | 0.14    |
|         | Ours             | —             | 0.85 | 0.86  | 0.82  | 0.82  | 0.84  | 0.85  | 0.84 | 0.01    |
| CelebA  | MTCNN [8]        | CVPR (2016)   | 0.53 | 0.91  | 1.22  | 1.32  | 1.22  | 0.98  | 1.03 | 0.27    |
|         | RetinaFace [1]   | CVPR (2020)   | 0.29 | 0.76  | 0.95  | 1.06  | 1.00  | 1.01  | 0.85 | 0.27    |
|         | MediaPipe [40]   | Google (2020) | 0.14 | 1.35  | 1.43  | 1.51  | 1.43  | 1.37  | 1.21 | 0.48    |
|         | SCRFD [3]        | ICLR (2022)   | 0.26 | 0.86  | 1.20  | 1.33  | 1.30  | 1.01  | 0.99 | 0.37    |
|         | YuNet [4]        | MIR (2023)    | 0.87 | 1.20  | 1.34  | 1.36  | 1.32  | 1.17  | 1.21 | 0.17    |
|         | YOLOv8-Face [39] | ADICS (2024)  | 0.30 | 1.06  | 1.37  | 1.47  | 1.39  | 1.26  | 1.14 | 0.40    |
|         | RP-Net [7]       | PR (2025)     | 0.28 | 0.70  | 0.90  | 1.00  | 0.95  | 0.96  | 0.83 | 0.25    |
|         | Ours             | —             | 0.24 | 0.25  | 0.26  | 0.23  | 0.26  | 0.24  | 0.25 | 0.01    |

b) *Evaluation on WFLW*: We evaluated our method on the WFLW dataset, which is the most challenging face-alignment benchmark derived from WIDER FACE across six distortion categories (blur, expression, illumination, makeup, occlusion, pose) [47]. To assess rotation robustness, we additionally rotated each face in 30° increments from 0° to 240°.

As shown in Table V, at 0° state-of-the-art detectors such as RetinaFace and YOLOv8-Face achieve the lowest NME (0.10–0.11), with our method close behind at 0.17. However, their errors climb above 0.40 at 90° and exceed 1.00 beyond 120°. In contrast, our approach maintains an NME below 0.20 at every angle and across all distortion subsets. For example, on the Blur subset, the NME rises only from 0.17 at 0° to 0.20 at 240°, whereas RetinaFace’s error grows from 0.11 to 1.07.

The Pose subset is the most challenging: severe rotations cause many landmarks to move outside their usual spatial neighborhoods or become self-occluded. Here, conventional

detectors fail (NME above 2.00 at 150°–210°), but our method remains robust, with NME increasing modestly from 0.23 at 0° to 0.26 at 150°, and staying below 0.25 through 240°.

This robustness comes from two key steps. First, our topology-guided estimator combines fine-grained CNN features with global attention to predict landmark positions even when they shift or become occluded. Second, the Center-Guided Module aligns initial center estimates with semantic heatmaps and corrects large positional errors caused by extreme rotations.

### 3) Face Center Estimation Evaluation:

a) *Protocol*: Given an image  $I_i$ , a method outputs a predicted center  $\hat{c}_i = g(I_i)$ . We compare  $\hat{c}_i$  to the fixed RAF label  $c_i$  using

$$\|\hat{c}_i - c_i\|_2, \quad \text{NME}(\hat{c}_i, c_i), \quad \text{FCQA}(\hat{c}_i, c_i),$$

and report averages aggregated across RIP angles (0°–360°; Table VI). Labels are produced offline from human landmarks using the deterministic TCE rule; evaluation depends only on

TABLE V  
NORMALIZED MEAN ERROR (IN %) ON WFLW ACROSS SIX DISTORTION SUBSETS. BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY.

| Subset | Model       | 0°   | 90°  | 120° | 150° | 180° | 210° | 240° |
|--------|-------------|------|------|------|------|------|------|------|
| Blur   | RetinaFace  | 0.11 | 0.42 | 0.84 | 1.05 | 1.14 | 1.04 | 1.07 |
|        | MediaPipe   | 0.33 | 0.24 | 1.43 | 1.72 | 1.57 | 1.56 | 1.61 |
|        | SCRFD       | 0.18 | 0.39 | 1.13 | 1.33 | 1.35 | 1.32 | 1.12 |
|        | YuNet       | 0.21 | 0.97 | 1.36 | 1.48 | 1.41 | 1.42 | 1.41 |
|        | YOLOv8-Face | 0.10 | 0.32 | 0.55 | 0.68 | 0.59 | 0.64 | 0.54 |
|        | RP-Net      | 0.20 | 0.30 | 0.65 | 0.81 | 0.73 | 0.79 | 0.68 |
|        | Ours        | 0.17 | 0.17 | 0.19 | 0.18 | 0.17 | 0.18 | 0.20 |
| Exp.   | RetinaFace  | 0.10 | 0.52 | 1.04 | 1.40 | 1.28 | 1.47 | 1.16 |
|        | MediaPipe   | 0.25 | 0.40 | 1.56 | 1.38 | 1.37 | 1.55 | 1.29 |
|        | SCRFD       | 0.16 | 0.48 | 1.08 | 1.30 | 1.32 | 1.29 | 1.19 |
|        | YuNet       | 0.21 | 0.92 | 1.20 | 1.37 | 1.34 | 1.35 | 1.21 |
|        | YOLOv8-Face | 0.10 | 0.34 | 0.59 | 0.68 | 0.63 | 0.71 | 0.63 |
|        | RP-Net      | 0.19 | 0.33 | 0.70 | 0.86 | 0.78 | 0.81 | 0.73 |
|        | Ours        | 0.15 | 0.15 | 0.14 | 0.15 | 0.13 | 0.16 | 0.15 |
| Ill.   | RetinaFace  | 0.10 | 0.43 | 1.08 | 1.33 | 1.30 | 1.36 | 1.03 |
|        | MediaPipe   | 0.26 | 0.43 | 1.14 | 1.36 | 1.42 | 1.37 | 1.37 |
|        | SCRFD       | 0.15 | 0.36 | 0.64 | 0.75 | 0.69 | 0.73 | 0.63 |
|        | YuNet       | 0.19 | 1.23 | 1.60 | 1.52 | 1.41 | 1.57 | 1.62 |
|        | YOLOv8-Face | 0.10 | 0.33 | 0.64 | 0.91 | 0.66 | 0.76 | 0.67 |
|        | RP-Net      | 0.18 | 0.30 | 0.58 | 0.72 | 0.65 | 0.71 | 0.62 |
|        | Ours        | 0.15 | 0.15 | 0.13 | 0.14 | 0.13 | 0.15 | 0.16 |

| Subset | Model       | 0°   | 90°  | 120° | 150° | 180° | 210° | 240° |
|--------|-------------|------|------|------|------|------|------|------|
| Mu.    | RetinaFace  | 0.11 | 0.50 | 1.98 | 1.28 | 1.31 | 1.26 | 0.99 |
|        | MediaPipe   | 0.32 | 0.42 | 1.27 | 1.44 | 1.44 | 1.43 | 1.38 |
|        | SCRFD       | 0.16 | 0.34 | 0.64 | 0.83 | 0.79 | 0.92 | 0.66 |
|        | YuNet       | 0.20 | 1.32 | 1.32 | 1.55 | 1.32 | 1.53 | 1.41 |
|        | YOLOv8-Face | 0.09 | 0.44 | 0.75 | 0.98 | 0.83 | 0.90 | 0.71 |
|        | RP-Net      | 0.19 | 0.32 | 0.68 | 0.80 | 0.73 | 0.78 | 0.66 |
|        | Ours        | 0.17 | 0.19 | 0.16 | 0.15 | 0.16 | 0.14 | 0.15 |
| Occ.   | RetinaFace  | 0.11 | 0.69 | 1.21 | 1.49 | 1.45 | 1.51 | 1.25 |
|        | MediaPipe   | 0.32 | 0.52 | 1.31 | 1.45 | 1.48 | 1.53 | 1.35 |
|        | SCRFD       | 0.10 | 0.54 | 0.79 | 0.92 | 0.86 | 0.97 | 0.79 |
|        | YuNet       | 0.21 | 1.31 | 1.73 | 1.70 | 1.44 | 1.73 | 1.55 |
|        | YOLOv8-Face | 0.11 | 0.50 | 0.83 | 1.03 | 0.94 | 0.99 | 0.89 |
|        | RP-Net      | 0.20 | 0.48 | 0.78 | 0.95 | 0.85 | 0.88 | 0.82 |
|        | Ours        | 0.18 | 0.18 | 0.20 | 0.19 | 0.17 | 0.22 | 0.19 |
| Pose   | RetinaFace  | 0.17 | 0.70 | 1.30 | 2.06 | 1.70 | 2.61 | 1.57 |
|        | MediaPipe   | 0.62 | 0.60 | 1.64 | 1.83 | 1.90 | 2.33 | 1.45 |
|        | SCRFD       | 0.25 | 0.62 | 1.02 | 1.16 | 1.02 | 1.22 | 0.99 |
|        | YuNet       | 0.32 | 2.05 | 2.41 | 2.02 | 1.79 | 2.14 | 2.32 |
|        | YOLOv8-Face | 0.16 | 0.52 | 0.82 | 1.25 | 0.87 | 1.29 | 0.97 |
|        | RP-Net      | 0.25 | 0.55 | 0.88 | 1.18 | 1.03 | 1.24 | 1.06 |
|        | Ours        | 0.23 | 0.25 | 0.26 | 0.24 | 0.24 | 0.23 | 0.22 |

(a) Blur, Expression, and Illumination subsets.

(b) Makeup, Occlusion, and Pose subsets.

TABLE VI  
FACE CENTER ESTIMATION RESULTS EVALUATED ON THE MTFD DATASET  
ACROSS RIP ANGLES (0°–360°). LOWER IS BETTER. TOP BLOCK:  
PRETRAINED DETECTORS EVALUATED WITHOUT RAF SUPERVISION.  
BOTTOM BLOCK: DETECTORS TRAINED ON RAF.

| Method                      | Euclid. Dist ↓ | NME ↓       | FCQA ↓       |
|-----------------------------|----------------|-------------|--------------|
| <i>Without RAF training</i> |                |             |              |
| MTCNN [8]                   | 32.65          | 1.33        | 0.188        |
| CenterFace [21]             | 29.46          | 1.23        | 0.539        |
| MediaPipe [40]              | 17.30          | 0.55        | 0.161        |
| YOLOv8-Face [39]            | 33.51          | 1.31        | 0.195        |
| SCRFD [3]                   | 44.46          | 1.06        | 0.220        |
| RetinaFace [1]              | 48.57          | 1.11        | 0.240        |
| RP-Net [7]                  | 10.50          | 0.35        | 0.120        |
| <i>With RAF training</i>    |                |             |              |
| CenterFace [21]             | 36.19          | 0.82        | 0.415        |
| YOLOv8-Face [39]            | 27.72          | 0.62        | 0.135        |
| SCRFD [3]                   | 51.80          | 1.21        | 0.300        |
| RetinaFace [1]              | 19.82          | 0.45        | 0.145        |
| RP-Net [7]                  | 8.50           | 0.28        | 0.110        |
| <b>TCE (Ours)</b>           | <b>7.99</b>    | <b>0.27</b> | <b>0.108</b> |

( $\hat{c}_i, c_i$ ), so there is no feedback from predictions into labeling or scoring.

*b) Results:* Table VI shows that conventional detectors exhibit sizable center drift under large RIP. Pretrained models reach FCQA values between 0.161 and 0.539, which indicates substantial deviation from the semantic center. MediaPipe [40] performs better than most anchor-based and anchor-free detectors because its center is derived directly from dense landmarks. After training on RAF, several detectors reduce their errors, but the magnitude of improvement differs across models. RetinaFace [1] shows one of the largest gains, with Euclidean distance decreasing from 48.57 px to 19.82 px and NME decreasing from 1.11 to 0.45, a result that reflects sub-

stantially improved center stability under rotation. YOLOv8-Face [39] also benefits from RAF supervision, whereas CenterFace [21] improves mainly in FCQA because the coarse heatmap representation limits the precision of center updates. RP-Net [7] exhibits only a small improvement because the landmark-driven design already produces near-semantic centers. SCRFD [3] shows no improvement, as dense anchor priors constrain how far predicted centers can shift toward the TCE-defined target. The TCE estimator attains the strongest performance across all metrics and reaches an FCQA of 0.108, a result that reflects highly consistent, semantically aligned, and rotation-robust center predictions.

*c) Cross-model validation of RAF:* We validate generalization using detectors from three major families: anchor-based methods (SCRFD [3], RetinaFace [1]), anchor-free methods (CenterFace [21], YOLOv8-Face [39]), and keypoint or hybrid regression methods (MTCNN [8], MediaPipe [40], RP-Net [7]). SCRFD and RetinaFace estimate centers as the midpoints of regressed anchor boxes, while CenterFace uses heatmap peaks and YOLOv8-Face uses grid-cell regression. Keypoint and hybrid methods compute centers from landmarks or pose-normalized boxes. RAF operates independently of architecture because its TCE-based labels are generated offline and evaluation depends only on ( $\hat{c}_i, c_i$ ), which allows a uniform protocol across detector types. In summary, RAF training yields improvements across multiple detectors. The results support the view that RAF offers generalizable benefits beyond our Hybrid-ViT design.

*4) Qualitative Analysis:* We next present visual results to illustrate how our method behaves in challenging scenarios.

*a) Comparison under ROP Rotations:* Fig. 8 compares face-center predictions and bounding-box outputs from several leading detectors against our TCE-guided approach as the head rotates out of plane. In each row, baseline centers drift from the true semantic midpoint and their boxes warp at steep angles.

By contrast, our method keeps the red center tightly on the face’s true center and the blue box aligned, even at extreme tilts.

*Note on occasional nose truncation.* We observe a rare failure mode at large yaw: when CGM blends the heatmap center with the TCE-derived center from predicted landmarks, a slight lateral bias toward the far cheek can appear. This conservative width bias may cause mild nose-side truncation, while the vertical extent remains accurate.



Fig. 8. A visual comparison of face-center predictions under ROP conditions shows that our TCE-based method aligns more closely with the true semantic centers than other state-of-the-art detection approaches.

*b) Results on WIDER FACE:* Fig. 10 shows sample detections on the WIDER FACE validation set. After rotation-aware training and semantic-center estimation, our model detects faces with higher recall and more accurate landmarks across a wide variety of poses, scales, and occlusions. The bounding boxes remain well-fitted and the landmarks stay consistently aligned, which demonstrates robust performance on real, in-the-wild images.

### E. Ablation Studies

Table VII summarizes the accuracy–efficiency trade-off across existing face detectors (CenterFace [21], SCRFD [3], YOLOv8-Face [39], RetinaFace [1], RP-Net [7]) and our variants (Baseline, Baseline+RAF, Baseline+RAF+Hybrid-ViT, Full Model). Under a shared protocol, the table reports Params, FLOPs, model size, FDDb recall, and FCQA. The corresponding accuracy–compute trend appears in Fig. 9.

Fig. 9 provides a complementary overview by jointly plotting FDDb Recall against FCQA, with bubble size indicating model complexity. A lower FCQA corresponds to more accurate and semantically aligned center predictions, while higher recall reflects stronger detection capability. The chart reveals a clear performance trade-off across detectors: models with higher recall often exhibit poorer center-quality accuracy,

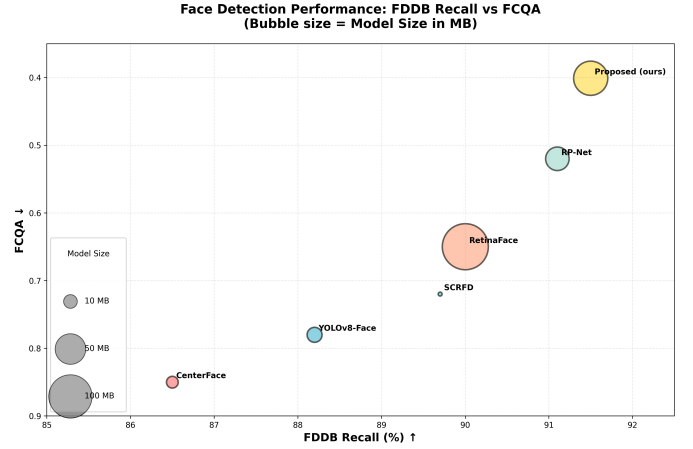


Fig. 9. Comparison of face detection performance across FDDb Recall and FCQA. Each point represents a detector, with bubble size indicating model size (MB). The upper-right region corresponds to the optimal operating zone, where detectors achieve both high recall and low FCQA.

especially those relying heavily on anchor-based regression or geometric priors. For instance, RetinaFace achieves competitive recall but shows high FCQA scores, which indicates a drift of the bounding-box midpoint from the semantic center under rotation. SCRFD and YOLOv8-Face follow the same pattern and demonstrate the difficulty CNN-only detectors face when handling large head-pose variations.

In contrast, RP-Net and our proposed method cluster toward the desirable region of high recall and low FCQA. RP-Net benefits from a rotation-aware design, while the FCQA remains around 0.52, a level consistent with reliance on polar alignment. Our method achieves the best combination of both metrics and reaches the highest FDDb recall (91.5%) together with the lowest FCQA (0.401). These results show that the Hybrid-ViT and CGM components improve semantic-center precision without incurring excessive model size or computational overhead.

Table VIII further isolates the contribution of TCE supervision and CGM by using identical training and evaluation settings, which allows a direct comparison of the incremental benefit of RAF via TCE, Hybrid-ViT, and CGM.

*1) Impact of RAF (with TCE):* Removing RAF supervision (Baseline) reduces recall from 91.5% to 87.2% and increases FCQA from 0.401 to 0.561. This decline shows that RAF provides essential rotation-aware samples and TCE-based semantic center labels, both of which are necessary for maintaining stable predictions under large out-of-plane rotations.

*2) Impact of Hybrid-ViT:* Replacing the Hybrid-ViT backbone with MobileNetV4 (Baseline+RAF) decreases recall from 91.5% to 89.3% and raises FCQA from 0.401 to 0.492. Hybrid-ViT therefore contributes meaningful gains by supplying global-context modeling through lightweight attention layers, which improves robustness to extreme head poses compared to a purely convolutional backbone.

*3) Impact of CGM:* Removing CGM while keeping RAF and Hybrid-ViT (Baseline+RAF+Hybrid-ViT) lowers recall from 91.5% to 90.8% and increases FCQA from 0.401 to



Fig. 10. Illustration of detection performance under the challenging conditions of the WIDER FACE dataset.

TABLE VII

ACCURACY–EFFICIENCY COMPARISON ACROSS SPECIALIZED DETECTORS AND OUR MODELS UNDER A COMMON PROTOCOL: LOWER FCQA IS BETTER.

| Method                                | Params (M)  | FLOPs (G)   | Model Size (MB) | FDDb Recall (%) ↑ | FCQA ↓       |
|---------------------------------------|-------------|-------------|-----------------|-------------------|--------------|
| CenterFace [21]                       | 2.30        | 2.28        | 7.42            | 86.5              | 0.850        |
| SCRFD [3]                             | <b>0.82</b> | <b>2.20</b> | <b>0.82</b>     | 89.7              | 0.720        |
| YOLOv8-Face [39]                      | 16.80       | 2.70        | 12.05           | 88.2              | 0.780        |
| RetinaFace [1]                        | 41.70       | 38.84       | 113.51          | 90.0              | 0.650        |
| RP-Net [7]                            | 7.40        | 3.83        | 29.60           | 91.1              | 0.520        |
| <i>Proposed (ours)</i>                |             |             |                 |                   |              |
| Baseline + RAF + Hybrid-ViT (w/o CGM) | 16.39       | 6.56        | 62.53           | 90.8              | 0.408        |
| Full Model (RAF + Hybrid-ViT + CGM)   | 16.39       | 6.57        | 62.53           | <b>91.5</b>       | <b>0.401</b> |

TABLE VIII

ABLATION ON TCE SUPERVISION AND CGM REFINEMENT. TCE DEFINES THE SEMANTIC CENTER DURING TRAINING AND GUIDES CGM REFINEMENT AT INFERENCE. LOWER FCQA INDICATES BETTER QUALITY.

| Method                                | TCE (Sup.) | TCE (CGM) | FDDb Recall (%) ↑ | FCQA ↓       |
|---------------------------------------|------------|-----------|-------------------|--------------|
| Baseline (w/o RAF)                    | ×          | ×         | 87.2              | 0.561        |
| Baseline + RAF (w/o Hybrid-ViT)       | ✓          | ×         | 89.3              | 0.492        |
| Baseline + RAF + Hybrid-ViT (w/o CGM) | ✓          | ×         | 90.8              | 0.408        |
| Full Model (RAF + Hybrid-ViT + CGM)   | ✓          | ✓         | <b>91.5</b>       | <b>0.401</b> |

0.408. This indicates that CGM provides valuable refinement by reducing residual misalignment between coarse predictions and the TCE-derived semantic center.

Overall, these ablations show that effective face-center estimation relies on both high-quality supervision and suitable architectural components. RAF provides rotation-aware and topology-guided supervision through TCE, which establishes a stable semantic center during training. Hybrid-ViT offers

global spatial cues that strengthen robustness to large pose variations. CGM refines the predicted center by using the TCE estimate again at inference and this step corrects residual localization errors. The progressive improvements confirm the modules' complementary contributions, and the full model attains the highest recall and best FCQA.

## V. CONCLUSION

In this work, we propose a rotation-invariant framework for face detection and landmark localization that estimates semantic face centers without relying on rigid geometric priors. The framework is further supported by a rotation-aware dataset that reduces the need for extensive online augmentation and improves robustness across arbitrary poses. Our experiments demonstrate that the proposed method outperforms conventional center estimation approaches in FCQA scores and other key metrics. RAF provides broad rotational coverage, but its landmark annotations remain noisy under heavy occlusion or low resolution, and its bounding-box labels are imprecise. Future work will focus on refining these annotations to enable more reliable RIFD evaluation.

## REFERENCES

- [1] 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.*, 2020, pp. 5202–5211.
- [2] Z. P. Yu, H. B. Huang, W. J. Chen, Y. X. Su, Y. H. Liu, and X. Y. Wang, "YOLO-FaceV2: A scale and occlusion aware face detector," *Pattern Recognit.*, vol. 155, Art. no. 110714, 2024.
- [3] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, "Sample and computation redistribution for efficient face detection," *arXiv preprint arXiv:2105.04714*, 2021. [Online]. Available: <https://arxiv.org/abs/2105.04714>
- [4] W. Wu, H. Peng, and S. Yu, "YuNet: A tiny millisecond-level face detector," *Mach. Intell. Res.*, vol. 20, no. 5, pp. 656–665, 2023.
- [5] X. Shi, S. Shan, M. Kan, S. Wu, and X. Chen, "Real-time rotation-invariant face detection with progressive calibration networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2295–2303.
- [6] X. Li, D. Zhang, M. Li, and D.-J. Lee, "Accurate head pose estimation using image rectification and a lightweight convolutional neural network," *IEEE Trans. Multimedia*, vol. 25, pp. 2239–2251, 2023.
- [7] H. Kaewkorn, L. Zhou, and W. Li, "RP-Net: A robust polar transformation network for rotation-invariant face detection," *Pattern Recognit.*, vol. 158, Art. no. 111044, 2025.
- [8] K. Zhang, Z. Zhao, Z. Chen, Z. Lei, and S. Li, "Joint face detection and alignment using multitask cascaded convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3697–3706.
- [9] B. Yang, C. Yang, Q. Liu, and X.-C. Yin, "Joint rotation-invariance face detection and alignment with angle-sensitivity cascaded networks," in *Proc. ACM Multimedia*, 2019, pp. 1473–1480.
- [10] L. Zhou, H. Zhao, and J. Leng, "MTCNet: Multi-task collaboration network for rotation-invariance face detection," *Pattern Recognit.*, vol. 124, Art. no. 108425, 2022.
- [11] A. Asperti and D. Filippini, "Deep learning for head pose estimation: A survey," *SN Comput. Sci.*, vol. 4, no. 4, Art. no. 349, 2023.
- [12] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, 2016.
- [13] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 2144–2151.
- [14] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 146–155.
- [15] Z. He, K. Zhao, B. Li, and Y. Li, "Regulating confidence by corner discrepancy and center score," *J. Intell. Fuzzy Syst.*, vol. 44, no. 6, pp. 10709–10720, 2023.
- [16] Y. T. Feng, S. Q. Yu, H. Y. Peng, Y. R. Li, and J. G. Zhang, "Detect faces efficiently: A survey and evaluations," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 4, no. 1, pp. 1–18, 2021.
- [17] J. Li et al., "DSFD: Dual shot face detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5055–5064.
- [18] S. Zhang et al., "S<sup>3</sup>FD: Single shot scale-invariant face detector," *arXiv preprint arXiv:1708.05237*, 2017.
- [19] B. Zhang et al., "ASFD: Automatic and scalable face detector," *arXiv preprint arXiv:2003.11228*, 2020.
- [20] C. Chi et al., "Selective refinement network for high performance face detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 8231–8238.
- [21] Y. Xu et al., "CenterFace: Joint face detection and alignment using face as point," *Sci. Program.*, vol. 2020, Art. no. 7845384, 2020.
- [22] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.
- [23] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [24] Y. Tian et al., "FPNFormer: Rethink the method of processing the rotation-invariance and rotation-equivariance," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–10, 2024.
- [25] Z. Yuke, R. Yumeng, Y. Lei, and S. Guo, "RotaTR: Detection transformer for dense and rotated object," *arXiv preprint arXiv:2312.02821*, 2023.
- [26] S. Chen, M. Ye, and B. Du, "Rotation invariant transformer for recognizing object in UAVs," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 2565–2574.
- [27] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021.
- [28] R. W. Floyd, "Algorithm 97: Shortest path," *Commun. ACM*, vol. 5, no. 6, p. 345, 1962.
- [29] D. Qin et al., "MobileNetV4—Universal models for the mobile ecosystem," *arXiv preprint arXiv:2404.10518*, 2024. Available: <https://arxiv.org/abs/2404.10518>
- [30] Q. Wang et al., "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11534–11542.
- [31] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
- [32] D. C. Dowson and B. V. Landau, "The Fréchet distance between multivariate normal distributions," *J. Multivariate Anal.*, vol. 12, no. 3, pp. 450–455, 1982.
- [33] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," Univ. Massachusetts Amherst, Tech. Rep. UM-CS-2010-009, 2010.
- [34] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5525–5533.
- [35] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Fine-grained evaluation on face detection in the wild," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recognit.*, 2015, pp. 1–7.
- [36] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3730–3738.
- [37] Y. Xiong, W. Meng, J. Yan, and J. Yang, "A rotation-invariance face detector based on RetinaNet," *J. Phys. Conf. Ser.*, vol. 2562, 2023.
- [38] R. Varghese and S. M., "YOLOv8: A novel object detection algorithm with enhanced performance and robustness," in *Proc. Int. Conf. Adv. Data Eng. Intell. Comput. Syst. (ADICS)*, 2024, pp. 1–6.
- [39] Lindevs, "YOLOv8-Face," [Online]. Available: <https://github.com/lindevs/yolov8-face>, Apr. 2024.
- [40] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019. [Online]. Available: <https://arxiv.org/abs/1906.08172>
- [41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2016.
- [42] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. ECCV*, 2016, pp. 21–37.
- [43] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Cao, "Look at boundary: A boundary-aware face alignment algorithm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 2129–2138.
- [44] J. Wan, Z. Lai, J. Liu, J. Zhou, and C. Gao, "Robust Face Alignment by Multi-Order High-Precision Hourglass Network," *IEEE Trans. Image Process.*, vol. 30, pp. 121–133, Jan. 2021, doi:10.1109/TIP.2020.3032029.
- [45] 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, Jan. 2023, doi:10.1109/TIP.2023.3261749.
- [46] Z. Tang, X. Peng, K. Li, and D. N. Metaxas, "Towards efficient U-Nets: A coupled and quantized approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2038–2050, Aug. 2020.
- [47] X. Lan, Q. Hu, and J. Cheng, "ATF: An alternating training framework for weakly supervised face alignment," *IEEE Trans. Multimedia*, vol. 25, pp. 1798–1809, 2023.