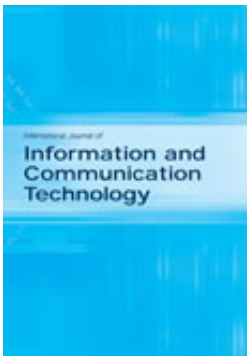
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research




**International Journal of Information and Communication Technology**

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

# Rotation-invariant face detection with guided deformable attention

Bin Deng, Guanghui Deng

**Article History:**

| | |
|---|---|
| Received: | 24 July 2024 |
| Last revised: | 30 August 2024 |
| Accepted: | 09 September 2024 |
| Published online: | 17 October 2024 |

# Rotation-invariant face detection with guided deformable attention

Bin Deng*

College of Computer Science,
Hunan University of Technology,
Zhuzhou, Hunan – 412007, China
Email: db1018@hut.edu.cn
*Corresponding author

Guanghui Deng

College of Science,
Hunan University of Technology,
Zhuzhou, Hunan – 412007, China
Email: 385829951@qq.com

**Abstract:** Detecting rotated faces has always been a challenging task. Fixed convolutional kernels struggle to effectively match features after rotation, while the sampling point offsets of deformable convolutions are limited by complex backgrounds. To address this issue, we propose a guided deformable attention (GDA) network. Guiding the offset direction of sampling points by adding constraints of facial structure to deformable convolutions. The GDA network adopts a dual-stream structure, with one branch detecting the inherent structural information for preliminary positioning of the face area; then, the second branch uses deformable convolution to perform pixel-level feature extraction on the face within the range. In addition, we introduce a novel loss, which, during the guidance process, aligns the activation areas in the feature maps extracted by the two branches through the KL divergence. Extensive experimental results validate that GDA network performs excellently on multiple face detection datasets, surpassing the current state-of-the-art face detection methods.

**Keywords:** rotated face detection; deformable convolution; attention; KL divergence; dual-stream; pixel-level.

**Reference** to this paper should be made as follows: Deng, B. and Deng, G. (2024) 'Rotation-invariant face detection with guided deformable attention', *Int. J. Information and Communication Technology*, Vol. 25, No. 8, pp.32–48.

**Biographical notes**: Bin Deng is an instructor and has a Master's degree. She graduated from Kunming University of Technology in 2008. She worked in Hunan University of Technology. Her research interests include digital image processing.

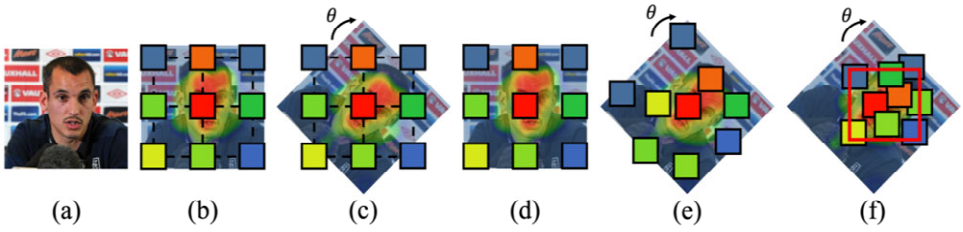
Guanghui Deng is an Associate Professor and graduated from Hunan Normal University in 1986. He worked at Hunan University of Technology. His research interests include mathematics and applied mathematics.

## 1 Introduction

Recently, there has been a growing demand for detecting rotated objects within scenes (Kumar et al., 2024; Pu et al., 2023), particularly within the realm of computer vision, where it has emerged as a prominently studied research area. Traditional object detection algorithms often assume that object instances are aligned with the image axes (Bah and Ming, 2020; Gao et al., 2024b); however, in real-world scenarios, objects often appear at various angles and orientations (Balasubaramanian et al., 2023). This situation is particularly evident in face detection (Zhang et al., 2024) because faces, as the core of human social interaction and cognition, exhibit a wide variety of poses and directions. In challenging application scenarios, such as face detection in surveillance videos, facial recognition, and face capture in virtual reality, accurately detecting and recognising rotated faces becomes particularly important.

Recently, significant progress has been made in the field of rotated face detection. Researchers have extensively analysed various representations of rotated faces and explored loss functions better suited to these representations (Zhou et al., 2022). Meanwhile, progressive networks and multi-task strategies for rotated face detection have also been thoroughly investigated (Xiong et al., 2023; Zhou et al., 2020a, 2020b). Additionally, some approaches have taken different routes by modelling the target as a Gaussian distribution and directly computing the distance between Gaussian distributions (Yang et al., 2022). However, current methods have limitations when handling rotated faces. Traditional fixed convolution kernels are ineffective at capturing local features of rotated objects, as the structural features of the face, such as the positions and shapes of key points, change with the rotation angle. Fixed convolution kernels are designed for features of fixed orientation and size, leading to inaccurate capture of local features after rotation. Furthermore, the overall structural features of the face also change with the rotation angle, with different face contours, proportions, and expressions presenting different forms at various angles. Traditional fixed convolution kernels struggle to adapt to these changes and fail to capture overall structural features effectively. In contrast, deformable convolutions can dynamically adjust the shape and position of the convolutional kernel to accommodate different rotation angles of the face, thereby better capturing local detail features of rotated objects and improving detection accuracy and robustness. It is worth noting that although deformable convolutions have been widely applied in the field of rotated object detection (Guo et al., 2022; Gao et al., 2024a), there is still a problem with the learned sampling point offsets being too random and dispersed. Despite performing well in areas like remote sensing images (Yin et al., 2023; Zha et al., 2023), deformable convolutions are more susceptible to background interference in rotated face detection due to the more complex background. As shown in Figure 1, Figure 1(a) is the original image, while Figures 1(b)–1(c) demonstrate the difficulty of traditional fixed convolutions in accurately matching features after image rotation. Figures 1(d)–1(e) indicate that although deformable convolutions can adaptively offset to enhance feature robustness to angles, the offsets of sampling points are too dispersed due to the complex background.

**Figure 1** The difference between fixed convolution kernels, deformable convolution kernels, and the proposed GDA (see online version for colours)




In this paper, we propose an innovative guided deformable attention (GDA) network to address the challenges in rotated face detection. We select deformable convolutions (Fu et al., 2023) as the main feature extraction method because they can dynamically adjust the position of convolution kernels to accommodate face rotation changes. However, deformable convolutions are prone to background interference when handling face rotation. To overcome this issue, we introduce a dual-stream structure in the GDA network to guide the offset of sampling points. Specifically, our network contains two key processing branches. The first branch performs coarse face localisation by learning an affine matrix. By reducing the dimensionality of the input feature map and using the affine matrix generated through fully connected layers, it performs initial rotation and scaling adjustments of the feature map, providing a precise face location reference for subsequent deformable convolutions. The second branch uses a KL divergence-guided mechanism to optimise the offsets of the sampling points, adjusting them in an orderly manner around the coarse localisation point. This KL divergence-guided mechanism ensures that the sampling points are more concentrated in the face region by calculating the differences between predicted and target features, thereby reducing the impact of background interference and enhancing the model's robustness in complex backgrounds. Through this dual-stream structure, the GDA network not only effectively combines the advantages of coarse localisation and fine adjustment but also significantly improves the precision of face feature extraction, demonstrating stronger adaptability and detection performance, as shown in Figure 1(f).

In summary, traditional fixed convolutional kernels cannot effectively capture the local features of rotated objects, while deformable convolution, although able to dynamically adjust the shape and position of the convolutional kernel to capture the features of rotated objects, the disordered sampling point offsets reduce the robustness of the features against complex backgrounds. Therefore, our research has the following three contributions:

1 We propose the GDA network for rotated face detection, which guides the offset of sampling points to enhance the model's robustness to rotation angles and complex backgrounds, and achieves excellent performance on multiple face detection datasets, reaching the state-of-the-art.

2 We introduce deformable convolution into the feature extraction of rotated faces to adapt to faces at different rotation angles, better capturing the local detail features of rotated faces.

3 We learn an affine matrix to roughly locate the detection target and guide the sampling points to offset orderly around the positioning points through KL divergence, improving the accuracy and efficiency of sampling point positioning.

## 2 Related work

### *2.1 Face detection*

Face detection stands as a cornerstone in research, with numerous methods having been proposed thus far. Among these techniques, the Haar cascade classifier (McCullagh, 2023) emerges as prominently utilised. Leveraging a sequence of classifiers and trained on diverse positive and negative image samples, it discerns areas within images likely to contain faces. Nonetheless, its computational efficiency and robustness exhibit certain constraints. The local binary pattern (LBP) (Shetty and Rebeiro, 2021) does not require the use of complex integral images like the Haar algorithm and is relatively less affected by changes in lighting and noise, as it is based on the differences between local pixels for feature extraction, rather than relying on the pixels of the entire area. However, LBP has less grasp of overall structure and shape information, which leads to poor performance in tasks that require global structural information, such as facial recognition where the overall contour and features need to be considered. To better address face detection under different scales, Jiang et al. (2022) designed a feature pyramid that captures information at various levels across different scales. By improving the method of feature fusion, the pyramid enables more comprehensive integration of feature information from different scales. Gao and Yang (2022) redesigned the YOLOv3 backbone network, introducing depthwise separable convolution, which not only reduces computational workload but also better captures complex face features in challenging environments by independently processing each channel, thus improving face detection performance. In addition, Khan et al. (2024) and others utilised a cascade of multiple networks to perform refined face detection at different stages, and by using feature fusion techniques, combined low-level detail information with high-level semantic information, achieving accurate face detection without the need for any facial alignment operations, enhancing its applicability and robustness in complex scenes.

### *2.2 Rotated face detection*

The rotation of the target brings about changes in the feature structure, and as the rotation angle changes, the distribution of key facial features in the image also changes, which poses higher requirements for feature extraction. To cope with this situation, Shi et al. (2018) and others proposed progressive calibration networks (PCN), which decompose the complex rotation angle prediction into multiple simple tasks, and achieve detection of faces at arbitrary angles through a three-stage cascaded network for progressive calibration of face rotation angles. Yang et al. (2019) and colleagues introduced angle-sensitivity cascaded networks (ASCN) as well. ASCN integrates face detection and alignment tasks in three cascaded networks. It leverages task correlations to enhance efficiency and handles intricate facial poses using a posture balance loss. Qi et al. (2022) introduced a five-point landmark regression head to predict five key facial landmarks, aiming to enhance the model's learning of facial geometry. They also used a wing loss function to constrain the keypoint prediction, thereby improving the prediction accuracy. Inspired by previous work, Zhou et al. (2022) made fuller use of facial structural information, enhancing the expressive power of features through a collaborative mechanism. Similar to the coarse-to-fine calibration process of PCN and the cascaded

network of ASCN, they predict faces with gradually decreasing rotation-invariant planar ranges at each stage, and integrate geometric angles into the penalty process, proposing a new training loss function. However, the utilisation of intricate cascaded networks in these methods adds to the model's complexity. In response, Xiong et al. (2023) and other researchers have opted to discard the multi-stage cascaded architecture. Instead, they employ a feature pyramid network to capture multi-scale features. Additionally, they introduce rotation-sensitive anchors, characterising the rotation state through an angle parameter. This approach aids the network in grasping directional information of the target and enhances the precision in predicting its actual shape. Consequently, the bounding box tightly encloses the target object, mitigating the risks of false positives and missed detections. Compared to other methods, Shao et al. (2021) consider facial alignment and facial action unit detection to be two highly related tasks. Therefore, they attempted to input high-level features from facial alignment into facial action unit detection, thereby achieving multi-scale feature alignment. Luo et al. (2021) explored the interrelationships between local regions through a graph convolutional network to restore geometric information. Additionally, they incorporated adversarial generative networks to mitigate the impact of imbalanced deformation attribute distribution on the results.

To address the challenges of rotated target detection, deformable convolution has been widely applied in other datasets in this field (Chen and Wang, 2023; Wang et al., 2024; Su et al., 2022). Firstly, deformable convolution is combined with boundary-aware vectors to address the shortcomings of traditional convolution in dealing with the directionality and geometric changes of objects (Chen and Wang, 2023). Secondly, deformable convolution, in conjunction with the loss function, can adjust the shape of the convolutional kernel to adapt to densely distributed detection targets (Wang et al., 2024). Finally, due to its strong rotation detection capability and not relying on angle regression, deformable convolution has been widely used in arbitrary direction detection tasks in large-area remote sensing images (Su et al., 2022). However, we found that the promotion of deformable convolution in rotated face detection tasks is somewhat limited. Remote sensing image rotation target detection usually segments the image, making the background of most images tend to be consistent, allowing deformable convolution to ignore the impact of the background on the target. However, the images in face detection datasets come from daily life, resulting in a certain degree of complexity in the background, which brings certain challenges to the robustness of deformable convolution against complex backgrounds.

To address this, we propose the GDA network. First, the traditional fixed convolution roughly locates the face part of the image. Then, combined with deformable convolution, the offset of the sampling points is constrained to the positioning area to reduce the impact of the complex background on deformable convolution.

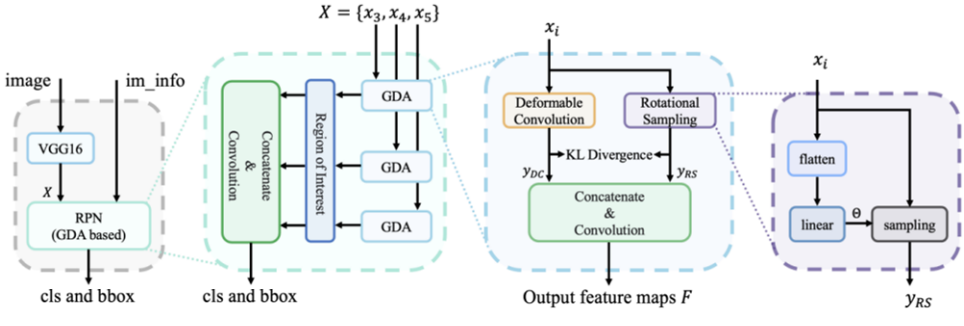
## 3 Proposed method

### *3.1 Overall framework*

Figure 2 illustrates the structure of the proposed GDA network, employing VGG16 (Simonyan and Zisserman, 2014) as its core architecture, and taking an image as input. The backbone network is segmented into three stages, labelled conv_3 (256), conv_4 (512), and conv_5 (512), denoting the dimensions of the feature maps generated at each

stage. The features $X = \{x_3, x_4, x_5\}$ output by VGG16 are used as the input for the region proposal network (RPN) (Ren et al., 2017). It should be noted that this is not the original RPN; we have embedded the proposed GDA into the RPN. Specifically, the GDA is directly processed on the feature maps of each stage, and then the region of interest (RoI) is extracted. Finally, based on the feature maps at different scales, a higher confidence class (cls) and bounding box (bbox) are obtained. The feature maps $x_i$ output by each stage are all processed by GDA to optimise the feature representation. The feature optimisation process can be divided into three steps. First, the process of GDA can be divided into two branches. The first branch, consisting of deformable convolutions, captures the possible presence of rotated targets in the feature map. The second branch samples and locates the area of the feature map that may contain the detection target through an affine matrix. Then, the KL divergence between the two intermediate feature maps $y_{DC}$ and $y_{RS}$, is calculated to align the activation areas, thereby constraining the range of the sampling point offsets of the deformable convolution. Finally, the two intermediate feature maps are integrated to obtain the optimised features $F$. It is worth mentioning that the rotation sampling step is not complex, but it can achieve good results. First, the feature map $x_i$ is flattened according to pixel values, then a series of fully connected networks are used, and finally, an affine matrix $\Theta$ is output to sample the original feature map $x_i$.

**Figure 2** Flowchart of the GDA network (see online version for colours)




## *3.2 Guided deformable attention*

GDA utilises deformable convolution to extract the parts of the input feature map $X = \{x_3, x_4, x_5\}$ that may contain rotated targets. The core of deformable convolution lies in its ability to dynamically adjust the sampling position of the convolutional kernel through the learned offset vector, adapting to the local changes of the input features. For any position $p_0$ in the feature map $x_i$, the output $y_{DC}(p_0)$ of deformable convolution is as shown in equation (1):

$$y_{DC}(p_0) = \sum_{p_n \in R} w(p_n) \cdot x_i(p_0 + p_n + \Delta p_n), \tag{1}$$

where $R$ represents the receptive field of the convolutional kernel, $w(p_n)$ is the weight in the convolutional kernel, $p_n$ is the position vector of the sampling point around the central point $p_0$ of the convolutional kernel, and $\Delta p_n$ is the corresponding position offset vector of the sampling point. In GDA, the receptive field size of the convolutional kernel is set

to 3 with padding of 1, meaning the convolutional kernel covers a local area of 3 × 3, and padding is used at the edges to maintain the size of the feature map.

The calculation of the offset vector $\Delta p_n$ is key to deformable convolution, as it determines the amount of displacement of the convolution kernel sampling points. In GDA, the offset vector is learned through the parameters of the convolutional kernel. For the input feature map $x_i \in R^{C\times H\times W}$ and the convolutional kernel $w \in R^{N\times N}$, the size of the output feature map $x_i^o$ is $(C + 2 \times N \times N) \times H \times W$, where $2 \times N \times N$ represents the offset of each element in the convolutional kernel.

Since the results of the offset vector $\Delta p_n$ are typically fractional and do not correspond to actual pixel points on the input feature map, bilinear interpolation is necessary to compute the shifted pixel values. The process of bilinear interpolation is depicted in Equation (2):

$$x_i(p) = \sum_q \max\left(0, 1-\left|q_x - p_x\right|\right) \cdot \max\left(0, 1-\left|q_y - p_y\right|\right) \cdot x(q), \tag{2}$$

where $p(p_x, p_y)$ represents the position of the shifted sampling point, and $q(q_x, q_y)$ represents the positions of the four nearest pixels in the feature map to $p$. Bilinear interpolation calculates the pixel value at the shifted position by weighting the values of the four nearest pixels based on the distance between $p$ and $q$.

To address the randomness issue of the initial offset vectors during training, GDA introduces a guiding mechanism. By learning affine transformation matrices, GDA can guide the sampling points of deformable convolutional kernels to shift in directions more conducive to feature extraction. The affine transformation matrices consider not only the local information of the feature map but also the global structural information, thereby enhancing the network's robustness to complex backgrounds.

In the second branch of the GDA network, the feature map $x_i \in R^{C\times H\times W}$ is first flattened into a one-dimensional vector for processing through fully connected layers. Subsequently, the flattened feature map is dimensionally reduced and transformed by two fully connected layers, $\alpha_l$ and $\beta_l$, to ultimately output the affine transformation matrix Θ. The specific process is illustrated in Equation (3):

$$y_{RS} = ReLu\left(\beta_l\left(\alpha_l\left(flatten\left(x_i\right)\right)\right)\right), \tag{3))}$$

where $\alpha_l \in R^{(C\times H\times W)\times\left(\frac{C}{2}\times H\times W\right)}$ and $\beta_l \in R^{\left(\frac{C}{2}\times H\times W\right)\times 6}$ represent the parameters of the two fully connected layers. ReLU is an activation function utilised to introduce nonlinearity.

During the initialisation stage, to ensure that the affine transformation matrix Θ does not affect the sampling range of the images, we adopt a specific strategy to initialise the weights in $\beta_l$. Specifically, all weights in $\beta_l$ are initialised to zero, while the bias is set to [1, 0, 0, 0, 1, 0]. This initialisation strategy ensures that at the beginning of training, the affine transformation approximates the identity transformation, thus avoiding unstable sampling caused by random initialisation at the early stages of training.

This process involves affine transformation of the feature map to achieve preliminary positioning of rotated faces. The sampled feature map $y_{RS}$ can more accurately reflect the global structural information of the face, providing positional guidance for subsequent deformable convolutions.

Finally, we concatenate the local feature map $y_{DC}$ and the global feature map $y_{RS}$. The concatenation operation is performed along the channel dimension, stacking the channels of the two feature maps to form a new feature map with double the number of channels. However, the new feature map has higher dimensionality, leading to increased computational complexity and feature redundancy. To reduce dimensionality and integrate information, we utilise a 1 × 1 convolution $\gamma$ to effectively reduce the dimensionality of the concatenated feature map and simultaneously learn complex relationships between different features, achieving feature fusion. The specific process is illustrated in equation (4):

$$F = \gamma[y_{RS}]\|[y_{DC}], \tag{4}$$

where $F$ is the final output feature map, and $\gamma$ is a 1 × 1 convolution with a scale of $R^{(2\times C)\times C}$.

### *3.3 Guidance mechanism*

During the initial stages of training, because the offset vectors of deformable convolutions are randomly initialised, their activation areas may not align with the activation areas of the global feature map $y_{RS}$. This misalignment can lead to a decrease in detection performance because local features may not accurately reflect the global structural information. Therefore, the introduction of the guiding mechanism is crucial for enhancing model performance.

To achieve guidance between feature maps, we compute the two-dimensional KL divergence between $y_{DC}$ and $y_{RS}$ as the loss function. KL divergence is a measure of the similarity between two probability distributions. When applied to feature maps, it encourages the predicted distribution $y_{DC}$ to approximate the ground truth distribution $y_{RS}$. The process of computing the two-dimensional KL divergence is illustrated in equation (5):

$$KL\left(y_{RS}\middle|y_{DC}\right) = \sum_{i=1}^{W}\sum_{j=1}^{H} y_{RS}\left(x_i, y_j\right)\log\frac{y_{RS}\left(x_i, y_j\right)}{y_{DC}\left(x_i, y_j\right)}, \tag{5}$$

where $W$ and $H$ represent the width and height of the feature map, respectively, and $x_i$ and $y_j$ represent the positional coordinates of the feature map in the horizontal and vertical directions. $y_{RS}(x_i, y_j)$ and $y_{DC}(x_i, y_j)$ represent the pixel values at the position $(x_i, y_j)$.

By using the two-dimensional KL divergence as the loss function, we can update the parameters in the network using the backpropagation algorithm. In each iteration, the gradient of the loss function is propagated through the network to adjust the offset vectors of deformable convolutions and the parameters of the affine transformation matrix. As training progresses, the activation areas of $y_{DC}$ will gradually align with those of $y_{RS}$, achieving effective guidance between feature maps.

However, in practical applications, to improve training stability and efficiency, we may need to optimise the weights of the loss function. Since the problem of random pixel offset is more severe in the initial stages of training, and after multiple iterations, the direction of pixel offset tends to stabilise, further constraining the pixel offset direction would lead to learning redundant feature representations. Therefore, we introduce a

hyperparameter ϵ to the KL divergence, reducing its impact on the model after a certain number of iterations.

### *3.4 Loss function*

In our endeavour, we utilised both the cross-entropy loss function (Wang et al., 2022) and the L2 loss function (Xun et al., 2022) to refine the classification and bounding box regression aspects within rotated face detection. The cross-entropy loss function evaluates the disparity between the model's predicted facial presence and the actual annotations, enhancing the model's ability to differentiate between facial and non-facial features more accurately. Conversely, the L2 loss function aims to minimise the Euclidean gap between the predicted bounding box and the ground truth bounding box, thereby facilitating precise facial localisation. In addition to these, we introduced new loss functions to guide the pixel offset direction of deformable convolutions during the initial stages of training, enhancing adaptability to rotational pose variations. The cross-entropy loss function and the L2 loss function are represented by equations (6) and (7) respectively:

$$L_{cls} = y \log \hat{y} + (1-y) \log (1-\hat{y}), \tag{6}$$

$$L_{l2} = \|t - \hat{t}\|_2^2, \tag{7}$$

where $y$ and $t$ are the ground truth labels, $\hat{y}$ and $\hat{t}$ are the predicted face confidence scores and bounding box predictions by the RPN. In summary, our joint loss function is represented by equation (8):

$$Loss = L_{cls} + L_{l2} + \in L_{kl}. \tag{8}$$

## 4 Experiments

### *4.1 Datasets*

To verify the effectiveness of our proposed method, we conducted tests on two datasets: the WIDER FACE (Yang et al., 2016) dataset and the FDDB (Jain and Learned-Miller, 2010) dataset.

WIDER FACE, a comprehensive face detection dataset, comprises 32,203 images, featuring a total of 393,703 annotated facial instances. Each subset, categorised as easy, medium, and hard, presents varying levels of difficulty, with the hard subset posing the greatest challenge. Consequently, evaluating the face detector's efficacy is most accurately achieved through its performance on the hard subset.

FDDB, a renowned face detection benchmark, encompasses 2,845 images and 5,171 annotated faces. To conduct a more thorough assessment of rotated face detection algorithms, we employed multi-oriented FDDB. This variant includes facial images in diverse orientations, facilitating a more precise evaluation of the algorithm's ability to detect rotated faces.

### *4.2 Experimental setup*

We implemented the GDA network using the PyTorch 1.6 framework and trained it using the Adam optimiser for a total of 120 epochs. The weight decay parameter of the network was set to $6 \times 10^{-5}$, and the momentum parameter was set to 0.7. The batch size was 128. The initial learning rate was set to 0.3, and the learning rate was reduced to one-tenth of its original value after every 30,000 iterations. The hyperparameter $\epsilon$ for the KL divergence loss function was initially set to 0.7 and was adjusted to 0.1 at the beginning of the 60th epoch. The entire network training process was completed on an Nvidia RTX 3090 GPU.

### *4.3 Ablation Study*

To evaluate the impact of different components on the GDA network, we conducted an ablation study on the WIDER FACE dataset. The results are shown in Table 1.

**Table 1** Performance (%) comparison of the GDA network with the baseline

| | *Deformable convolution* | *Rotational sampling* | *KL divergence* | *mAP* | | |
|---|---|---|---|---|---|---|
| | | | | *Easy* | *Medium* | *Hard* |
| Baseline | × | × | × | 91.3 | 89.2 | 81.5 |
| | ✓ | × | × | 94.5 | 91.7 | 80.9 |
| | × | ✓ | × | 93.7 | 92.1 | 82.3 |
| | ✓ | ✓ | × | 95.3 | 93.4 | 83.6 |
| GDA | ✓ | ✓ | ✓ | 96.7 | 95.3 | 87.7 |

It can be observed that the introduction of deformable convolution improved the model's performance on the easy and medium subsets of the WIDER FACE dataset, with increases of 3.2% and 2.5%, respectively. This indicates that deformable convolution allows the model to better adapt to the local feature changes of rotated faces, especially in capturing the changes of key feature points at different rotation angles. However, there was a decrease of 0.6% on the hard subset, which includes more extreme or uncommon situations that are more challenging for deformable convolution. The introduction of rotational sampling led to a stable improvement in performance on all three subsets, with increases of 2.4%, 2.9%, and 0.8%, respectively. This suggests that rotational sampling helps the model to recognise and locate rotated face areas, especially for faces that are not front-facing or horizontally oriented. Finally, the introduction of KL divergence significantly improved the model's performance on all three subsets, with increases of 5.4%, 6.1%, and 6.2%, respectively. KL divergence guides the sampling point offsets of deformable convolution by aligning the activation areas, enabling the model to learn more accurate feature representations during training, thereby enhancing the detection performance for rotated faces.

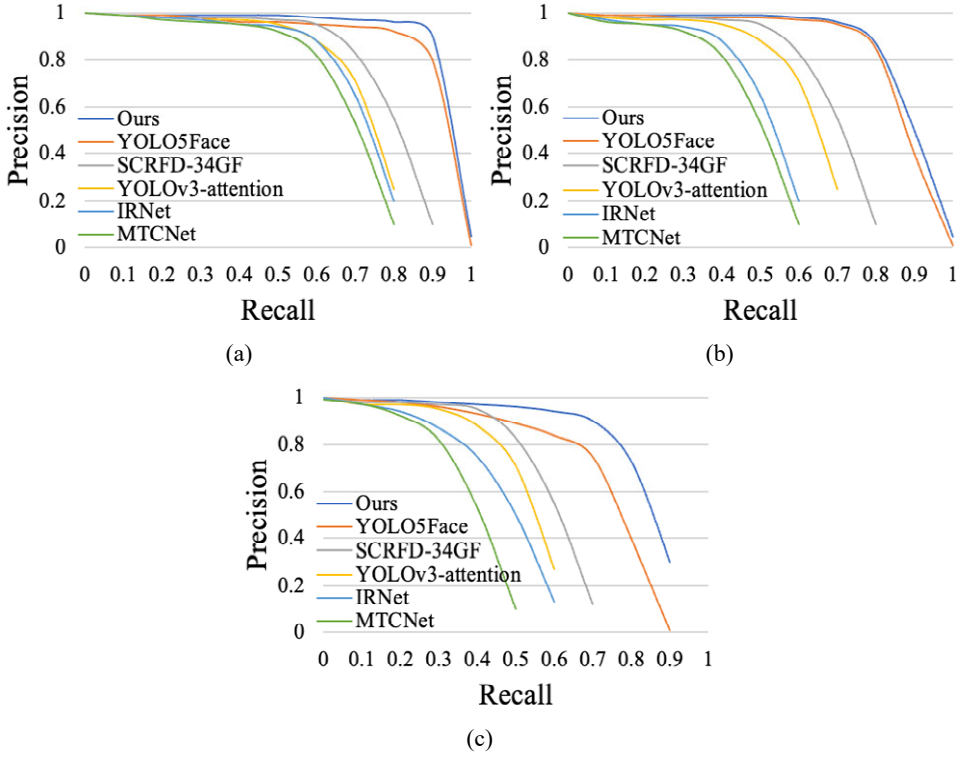
### *4.4 Comparison with state-of-the-art methods*

Currently, the methods for rotated face detection mainly include YOLO5Face (Qi et al., 2022), SCRFD-34GF (Guo et al., 2021), YOLOv3-attention (Liu et al., 2021), IRNet

(Jiang et al., 2022), MTCNet (Zhou et al., 2022), MTCNN-v2 (Gu et al., 2022), and FaceSSD (Ye et al., 2021).

We plotted the precision-recall (PR) curves for some of the models on the three subsets of the WIDER FACE dataset, and the results are shown in Figure 3.

**Figure 3** PR curves on the three subsets of the WIDER FACE dataset, (a) easy (b) medium (c) hard (see online version for colours)




(a) (b)

(c)

It can be observed that on the easy and medium subsets, the proposed GDA network performs similarly to YOLO5Face, but on the hard subset, YOLO5Face's performance begins to decline due to the inclusion of more extreme angles or partially occluded faces, while the GDA network remains stable. This is because in the GDA network, deformable convolution and rotational sampling better adapt to faces at different angles, and the guidance mechanism and KL divergence optimise the adaptability to complex backgrounds. In contrast, YOLO5Face may rely more on its underlying YOLOv5 architecture to process features, which works well on the easy and medium subsets but is not as effective on the hard subset.

Additionally, for a more intuitive comparison, we also numerically compared the proposed GDA network with existing work on the WIDER FACE dataset and the FDDB dataset, and the results are shown in Tables 2 and 3.

**Table 2** The mAP (%) comparison with state-of-the-art methods on WIDER FACE

| | *WIDER FACE* | | |
|---|---|---|---|
| | *Easy* | *Medium* | *Hard* |
| YOLO5Face (Qi et al., 2022) | 96.6 | 95.0 | 86.5 |
| SCRFD-34GF (Guo et al., 2021) | 96.1 | 94.9 | 85.3 |
| YOLOv3-attention (Liu et al., 2021) | 94.2 | 91.9 | 82.1 |
| IRNet (Jiang et al., 2022) | 91.8 | 89.3 | 76.6 |
| MTCNet (Zhou et al., 2022) | 84.8 | 82.5 | 59.8 |
| Baseline | 91.3 | 89.2 | 81.5 |
| Ours | 96.7 | 95.3 | 87.7 |

It can be observed that the proposed GDA network achieved excellent performance on all three subsets. Similar to the rotational sampling step, SCRFD-34GF adjusts the distribution of training samples to enable the model to focus on faces at different scales. However, to further optimise the network structure, SCRFD-34GF employs a two-step search strategy, focusing on the computational redistribution of the backbone and detector parts, which further increases the network's complexity. In contrast, the GDA network, through a simple and effective rotational sampling module, learns the affine matrix that can adaptively sample faces at different scales while maintaining low complexity. This design allows the GDA network to achieve a good balance between performance and computational complexity, making it more widely applicable and operable in practical applications.

On the FDDB dataset, the proposed GDA network still achieved the best performance. Although the two YOLO-based methods – YOLO5Face and YOLOv3-attention – perform similarly to the GDA network in terms of performance by introducing attention mechanisms and the advanced YOLO5 framework, they still face challenges in dealing with rotated face detection tasks. This limitation mainly stems from two aspects: first, these methods rely on traditional fixed convolutional kernels, which struggle to effectively match facial features when facing extreme posture changes, thus limiting the success rate of feature recognition; second, the traditional loss functions they use are not fully optimised for the specific needs of rotated face detection tasks, lacking sensitivity to the task's particular requirements. In contrast, the GDA network can achieve excellent performance because it employs flexible deformable convolution technology, which can dynamically adjust the convolutional kernel to adapt to the feature changes of rotated faces, making feature matching possible even under extreme postures. Compared with traditional fixed convolutional kernels, deformable convolution can better capture the local detail features of rotated faces, improving detection accuracy and robustness. In addition, the GDA network also introduces an innovative loss function design, guiding the sampling point offsets of deformable convolution through the calculation of KL divergence, further optimising the network's ability to extract features from rotated faces. This targeted design enables the GDA network to achieve more accurate feature positioning and higher detection accuracy in rotated face detection tasks, especially on the FDDB dataset.

**Table 3** The mAP (%) comparison with state-of-the-art methods on FDDB

| | *FDDB* |
|---|---|
| | *mAP* |
| YOLO5Face (Qi et al., 2022) | 98.8 |
| YOLOv3-attention (Liu et al., 2021) | 98.6 |
| MTCNet (Zhou et al., 2022) | 89.7 |
| MTCNN-v2 (Gu et al., 2022) | 96.0 |
| FaceSSD (Ye et al., 2021) | 92.4 |
| Baseline | 95.3 |
| Ours | 99.1 |

## *4.5 Visualisation experiment*

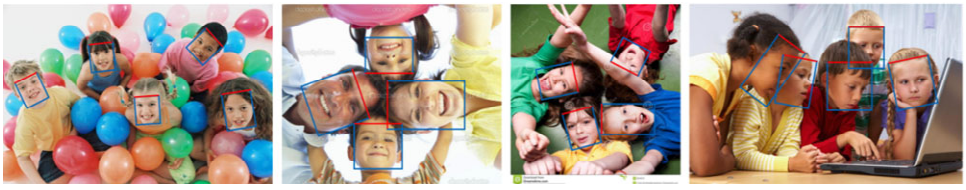
We selected some representative images from the WIDER FACE dataset and visualised the detection of rotated faces, with the results shown in Figure 4. Figure 4(a) shows the detection visualisation for the Baseline, and to more clearly demonstrate the performance of GDA, Figure 4(b) shows the detection visualisation for YOLO5Face, and Figure 4(c) is the detection visualisation for the GDA network, where red represents the upward direction. It can be observed that the three methods perform similarly on some simple images, but the Baseline and YOLO5Face have omissions in detecting faces with extreme rotations and occlusions. In contrast, the GDA network has further improved performance in detecting extremely rotated faces through deformable convolution and rotational sampling modules. In Figure 4(c), it can be clearly seen that the GDA network successfully detected faces with large rotation angles and was also able to accurately locate the position of faces even in the presence of occlusions. This indicates that the GDA network has stronger robustness and accuracy in tasks of detecting faces with extreme rotations and occlusions.

**Figure 4** Visualisation of rotated face detection on the WIDER FACE dataset for the baseline, YOLO5Face and GDA networks, (a) baseline (b) YOLO5Face (c) GDA network (see online version for colours)

(a)

(b)

**Figure 4** Visualisation of rotated face detection on the WIDER FACE dataset for the baseline, YOLO5Face and GDA networks, (a) baseline (b) YOLO5Face (c) GDA network (continued) (see online version for colours)



(c)

Finally, to further investigate the impact of the rotational sampling module and KL divergence on the sampling point offsets in deformable convolution, we visualised the sampling points to observe the different offset directions before and after the calculation of KL divergence. The visualisation results are shown in Figure 5, where Figure 5(a) shows the sampling point offsets without the constraint of KL divergence, and Figure 5(b) shows the sampling point offsets with the addition of KL divergence. In the figure, the green box represents the facial positioning obtained through the affine matrix sampling.

**Figure 5** Visualisation of sampling points in deformable convolution (see online version for colours)

(a) (b)

Observing Figure 5(a), it can be found that although the sampling points can cover the facial area, they are still affected by the surrounding background, which may have an adverse effect on the model's robustness. In contrast, the sampling points in Figure 5(b) are more concentrated around the target area of the face. This is because the rotational sampling has positioned the general direction of the face and, through KL divergence, has

brought the activation areas of the feature map output by deformable convolution closer to the positioning area. This more precise positioning of sampling points helps to improve the model's robustness against complex backgrounds, making the model more reliable in locating and recognising rotated faces.

## 5 Conclusions

In practical scenarios, due to variations in rotation angles and the influence of complex environments, traditional fixed and rotational convolutional kernels exhibit limitations in handling rotated faces. To address this issue, we propose a new network architecture called GDA. This approach combines deformable convolutions with a rotational sampling mechanism, enabling flexible adaptation to the local feature changes of rotated faces. Additionally, by introducing the KL divergence loss function, we optimise the sampling point offsets in deformable convolutions, thereby enhancing the detection performance and robustness for rotated faces. Experimental results demonstrate that GDA significantly outperforms existing rotated face detection algorithms on the WIDER FACE and FDDB datasets. However, during the process of guiding sampling points, the model tends to focus on global control, which may lead to reduced detection performance for small targets, particularly in dense scenes. Consequently, future research will explore the integration of the proposed guidance mechanism within a multi-scale network to enhance the model's ability to extract features at different levels of granularity.

## References

Bah, S.M. and Ming, F. (2020) 'An improved face recognition algorithm and its application in attendance management system', *Array*, Vol. 5, p.100014, DOI: https://doi.org/10.1016/j.array.2019.100014.

Balasubaramanian, S., Cyriac, R., Roshan, S., Paramasivam, K.M. and Jose, B.C. (2023) 'An effective stacked autoencoder based depth separable convolutional neural network model for face mask detection', *Array*, Vol. 19, p.100294, https://doi.org/10.1016/j.array.2023.100294.

Chen, H. and Wang, K. (2023) 'Fusing DCN and BBAV for remote sensing image object detection', *International Journal of Cognitive Informatics and Natural Intelligence*, Vol. 17, No. 1, pp.1–16, DOI: 10.4018/IJCINI.335496

Fu, X., Yuan, Z., Yu, T. and Ge, Y. (2023) 'DA-FPN: deformable convolution and feature alignment for object detection', *Electronics*, Vol. 12, No. 6, p.1354, https://doi.org/10.3390/electronics12061354.

Gao, F., Cai, C., Tang, W., Tian, Y. and Huang, K. (2024a) 'RA2DC-Net: a residual augment-convolutions and adaptive deformable convolution for points-based anchor-free orientation detection network in remote sensing images', *Expert Systems with Applications*, Vol. 238, p.122299, https://doi.org/10.1016/j.eswa.2023.122299.

Gao, S., Chen, Y., Cui, N. and Qin, W. (2024b) 'Enhancing object detection in low-resolution images via frequency domain learning', *Array*, Vol. 22, p.100342, DOI: https://doi.org/10.1016/j.array.2024.100342.

Gao, J. and Yang, T. (2022) 'Face detection algorithm based on improved TinyYOLOv3 and attention mechanism', *Computer Communications*, Vol. 181, pp.329–337, https://doi.org/10.1016/j.comcom.2021.10.023.

Gu, M., Liu, X. and Feng, J. (2022) 'Classroom face detection algorithm based on improved MTCNN', *Signal, Image and Video Processing*, Vol. 16, No. 5, pp.1355–1362, https://doi.org/10.1007/s11760-021-02087-x.

Guo, H., Bai, H., Yuan, Y. and Qin, W. (2022) 'Fully deformable convolutional network for ship detection in remote sensing imagery', *Remote Sensing*, Vol. 14, No. 8, p.1850, https://doi.org/10.3390/rs14081850.

Guo, J., Deng, J., Lattas, A. and Zafeiriou, S. (2021) *Sample and Computation Redistribution for Efficient Face Detection*, arxiv preprint arxiv:2105.04714.

Jain, V. and Learned-Miller, E. (2010) *FDDB: A Benchmark for Face Detection in Unconstrained Settings*, UMass Amherst Technical Report, Vol. 2, No. 6, pp.1–11.

Jiang, C., Ma, H. and Li, L. (2022) 'IRNet: an improved retinanet model for face detection', in *2022 7th International Conference on Image, Vision and Computing (ICIVC)*, pp.129–134, DOI: 10.1109/ICIVC55077.2022.9886975.

Khan, S.S., Sengupta, D., Ghosh, A. and Chaudhuri, A. (2024) 'MTCNN++: a CNN-based face detection algorithm inspired by MTCNN', *The Visual Computer*, Vol. 40, No. 2, pp.899–917, https://doi.org/10.1007/s00371-023-02822-0.

Kumar, S., Mishra, R., Jain, T. and Shankar, A. (2024) 'Advances image-based automated security system', *International Journal of Distributed Systems and Technologies*, Vol. 15, No. 1, pp.1–12.

Liu, Q., Lu, S. and Lan, L. (2021) 'YOLOv3 attention face detector with high accuracy and efficiency', *Computer Systems Science & Engineering*, Vol. 37, No. 2, https://doi.org/10.32604/csse.2021.014086.

Luo, M., Cao, J., Ma, X., Zhang, X. and He, R. (2021) 'FA-GAN: face augmentation GAN for deformation-invariant face recognition', *IEEE Transactions on Information Forensics and Security*, Vol. 16, pp.2341–2355, DOI: 10.1109/TIFS.2021.3053460.

McCullagh, P. (2023) 'Face detection by using Haar cascade classifier', *Wasit Journal of Computer and Mathematics Science*, Vol. 2, No. 1, pp.1–5, https://doi.org/10.31185/wjcm.109.

Pu, Y., Wang, Y., Xia, Z., Han, Y., Wang, Y., Gan, W., Wang, Z., Song, S. and Huang, G. (2023) 'Adaptive rotated convolution for rotated object detection', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.6589–6600.

Qi, D., Tan, W., Yao, Q. and Liu, J. (2022) 'YOLO5Face: why reinventing a face detector', in *European Conference on Computer Vision*, pp.228–244, https://doi.org/10.1007/978-3-031-25072-9_15.

Ren, S., He, K., Girshick, R. and Sun, J. (2017) 'Faster R-CNN: towards real-time object detection with region proposal networks', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp.1137–1149, DOI: 10.1109/TPAMI.2016.2577031.

Shao, Z., Liu, Z., Cai, J. and Ma, L. (2021) 'JAA-Net: joint facial action unit detection and face alignment via adaptive attention', *International Journal of Computer Vision*, Vol. 129, pp.321–340, DOI: https://doi.org/10.1007/s11263-020-01378-z.

Shetty, A.B. and Rebeiro, J. (2021) 'Facial recognition using Haar cascade and LBP classifiers', *Global Transitions Proceedings*, Vol. 2, No. 2, pp.330–335, https://doi.org/10.1016/j.gltp.2021.08.044.

Shi, X., Shan, S., Kan, M., Wu, S. and Chen, X. (2018) 'Real-time rotation-invariant face detection with progressive calibration networks', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2295–2303.

Simonyan, K. and Zisserman, A. (2014) *Very Deep Convolutional Networks for Large-Scale Image Recognition*, arxiv preprint arxiv:1409.1556.

Su, N., Huang, Z., Yan, Y., Zhao, C. and Zhou, S. (2022) 'Detect larger at once: large-area remote-sensing image arbitrary-oriented ship detection', *IEEE Geoscience and Remote Sensing Letters*, Vol. 19, pp.1–5, DOI: 10.1109/LGRS.2022.3144485.

Wang, L., Shen, Y., Yang, J., Zeng, H. and Gao, H. (2024) 'Rotated points for object detection in remote sensing images', *IET Image Processing*, Vol. 18, No. 6, pp.1655–1665, https://doi.org/10.1049/ipr2.13011.

Wang, X., Wang, S., Liang, Y., Gu, L. and Lei, Z. (2022) 'Rvface: reliable vector guided softmax loss for face recognition', *IEEE Transactions on Image Processing*, Vol. 31, pp.2337–2351, DOI: 10.1109/TIP.2022.3154293.

Xiong, Y., Meng, W., Yan, J. and Yang, J. (2023) 'A rotation-invariance face detector based on RetinaNet', in *Journal of Physics: Conference Series*, Vol. 2562, No. 1, p.012066, IOP Publishing, DOI: 10.1088/1742-6596/2562/1/012066.

Xun, Z., Wang, L. and Liu, Y. (2022) 'Improved face detection algorithm based on multitask convolutional neural network for unmanned aerial vehicles view', *Journal of Electronic Imaging*, Vol. 31, No. 6, pp.61804–61804, https://doi.org/10.1117/1.JEI.31.6.061804.

Yang, B., Yang, C., Liu, Q. and Yin, X.C. (2019) 'Joint rotation-invariance face detection and alignment with angle-sensitivity cascaded networks', in *Proceedings of the 27th ACM International Conference on Multimedia*, pp.1473–1480, https://doi.org/10.1145/3343031.3350877.

Yang, S., Luo, P., Loy, C.C. and Tang, X. (2016) 'Wider face: a face detection benchmark', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5525–5533.

Yang, X., Zhang, G., Yang, X., Zhou, Y., Wang, W., Tang, J., He, T. and Yan, J. (2022) 'Detecting rotated objects as Gaussian distributions and its 3-D generalization', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 4, pp.4335–4354, DOI: 10.1109/TPAMI.2022.3197152.

Ye, B., Shi, Y., Li, H., Li, L. and Tong, S. (2021) 'Face SSD: a real-time face detector based on SSD', in *2021 40th Chinese Control Conference (CCC)*, pp.8445–8450, DOI: 10.23919/CCC52363.2021.9550294.

Yin, S., Wang, L., Wang, Q., Yang, J. and Jiang, M. (2023) 'Remote sensing image segmentation based on a novel Gaussian mixture model and SURF algorithm', *International Journal of Swarm Intelligence Research*, Vol. 14, No. 2, pp.1–15, DOI: 10.4018/IJSIR.322301.

Zha, W., Hu, L., Sun, Y. and Li, Y. (2023) 'ENGD-BiFPN: a remote sensing object detection model based on grouped deformable convolution for power transmission towers', *Multimedia Tools and Applications*, Vol. 82, No. 29, pp.45585–45604, https://doi.org/10.1007/s11042-023-15584-7.

Zhang, J., Hou, C., Yang, X., Yang, X., Yang, W. and Cui, H. (2024) 'Advancing face detection efficiency: utilizing classification networks for lowering false positive incidences', *Array*, Vol. 22, p.100347, https://doi.org/10.1016/j.array.2024.100347.

Zhou, L., Zhao, H. and Leng, J. (2022) 'MTCNet: multi-task collaboration network for rotation-invariance face detection', *Pattern Recognition*, Vol. 124, p.108425, https://doi.org/10.1016/j.patcog.2021.108425.

Zhou, L.F., Gu, Y., Wang, P.S., Liu, F.Y., Liu, J. and Xu, T.Y. (2020a) 'Rotation-invariant face detection with multi-task progressive calibration networks', in *International Conference on Pattern Recognition and Artificial Intelligence*, pp.513–524, https://doi.org/10.1007/978-3-030-59830-3_44.

Zhou, L.F., Gu, Y., Liang, S., Lei, B.J. and Liu, J. (2020b) 'Direction-sensitivity features ensemble network for rotation-invariant face detection', in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp.581–590, https://doi.org/10.1007/978-3-030-60639-8_48.