Improved Face Detector on Fisheye Images via Spherical-Domain Attention

Jingbo Miao*†, Yanwei Liu*¶, Jinxia Liu‡, Antonios Argyriou§, Zhen Xu* and Yanni Han*
*Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
†School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
‡Zhejiang Wanli University, Ningbo, China
§University of Thessaly, Volos, Greece

***Abstract*—As one type of omnidirectional projection, fisheye images have been widely used in automatic driving and visual surveillance. However, they cannot be processed well by the traditional algorithms designed for the planar rectilinear images since they usually suffer from severe geometric distortion during image formation. In this paper, the conventional face detection algorithm is enhanced to fit the fisheye images via combining with the spherical convolution block by learning rotation-invariant features from the spherical domain. The learned features from both planar and spherical domains are subsequently mixed by the spatial attention mechanism. Consequently, the whole network can automatically learn the distorted features directly from different positions on the target image. Experimental results verify that our network can detect distorted faces on fisheye images effectively and maintain the performance on traditional planar images.**

***Index Terms*—Fisheye image, face detection, geometric distortion, spherical-domain attention**

I. Introduction

A 360-degree omnidirectional camera, that provides a super wide-angle view in space, uses a double fisheye lens for imaging to capture more information than ordinary lenses for the perception of the visual environment. For autonomous driving [1], intelligent 360-degree monitoring [2] is essential for rapidly detecting and recognizing the target in the omni-view scene. Furthermore, the fisheye image is one of the most widely used formats to represent an omni-view of the scene. Consequently, target detection on this view is of great significance for omni-directional visual applications.

In the past, several mature algorithms of object detection in traditional 2D rectilinear images have been developed. Early-developed detectors mainly used hand-crafted features such as the rectangle features on the integral image [3] and the Histograms of Oriented Gradient (HOG) features [4] to detect objects in images. As for the later studies, the work in [20, 26] used convolution to improve the performance of detection algorithms, while feature learning [16] based on Convolutional Neural Network (CNN) was adopted in the solution. In recent years, two mainstream CNN models have dominated due to their superior performance. The first is the two-stage model that adopts the strategy of predicting first and fine-tuning later [5, 6]. The second one is the single-stage model that predicts dense samples directly by regressing from anchor boxes to ground-truth object boxes.

The aforementioned schemes can attain superior performance relative to conventional rectilinear images. However, they are not suitable for fisheye images because the geometric projection leads to severe barrel distortion towards the perimeter of the circle. The geometric distortion in fisheye images poses a significant challenge for object detection. To alleviate this problem, one method that has been proposed is to rectify the fisheye image directly with explicit geometry constraints [9] or the line constraints [10]. However, this approach inevitably results in partial information loss of the original representation.

The approach in [11] proposed to re-train the traditional model that adopted the planar 2D convolution for fisheye face detection, something that makes challenging the capture of rotation-invariant features of objects. Recently, Cohen et al. [36] adopted the spherical convolution for capturing rotation-invariant features to counter the distortion. However, convolution on the sphere introduces a high computational cost which limits its practical applications. Inspired by this, we propose an attention mechanism with very few spherical convolution layers in this paper to improve the performance of a face detector in fisheye images.

The focus of this paper is to evolve existing algorithms into those conquering the geometric distortion. In this paper, spherical features, extracted with low computational complexity, are used to make up for the insufficiency of traditional planar convolution. The spherical feature is extracted only by a few convolutional layers and then combined with planar features via a spatial attention mechanism. The bi-domain features are regarded as the full features of the fisheye images, which are processed by the subsequent network for face detection later.

The main contributions are summarized as follows:

1). A novel face detector for fisheye images is proposed, which is based on fusing multi-scale and bi-domain features using feature pyramid and path enhancement.

2). We employ the attention mechanism to fuse the bi-domain features. To the best of our knowledge, it is the first work that proposes mixing signals from both the planar and spherical domains represented by the middle layers of a multi-scale feature network.

This work was supported in part by National Natural Science Foundation of China under Grants 61771469.

¶ Yanwei Liu is corresponding author.

3). Our approach integrates the two stages of rectification and detection for the face detection task of fisheye images into one network to alleviate the distortion problem, which omits the step of capturing prior knowledge of the camera. It also maintains the state-of-the-art performance in terms of both accuracy and speed for detecting faces in planar images.

The rest of the paper is organized as follows. Section II introduces the related work, and Section III presents the specific structure of the proposed face detector for fisheye images, the mechanism of the attention block, and the implementation details of the network. Then Section IV provides the experimental results, and Section V concludes the paper.

II. RELATED WORK

A. Face Detection

Face detection on planar images has undergone a significant evolution from traditional methods [17, 18, 19] towards the deep-learning-based techniques [20, 21, 22], which have shown compelling accuracy and speed in recent studies. Rowley et al. [17] used multi-layer perception to detect frontal faces initially. AdaBoost framework based on the ensemble learning [3, 23] derived the hierarchically cascaded face classifiers with different types of face features [25]. The two-stage target detection methods [5, 6] predicted class and coordinates with higher accuracy, but its speed was on the opposite side. The one-stage methods [29, 30, 31, 32] benefit from obtaining not the separate region suggestions but the dense anchors in multi-scale feature maps. Furthermore, as multi-task learning [27] was introduced, face recognition, face alignment, and key point alignment were jointly used as the training target of the network [26, 28]. In addition, networks with excellent performance recently [30, 31] used context modules to take advantage of different ranges of information perception.

The approaches mentioned above were all targeted for the rectilinear images and obtained superior performance. However, they are not suitable for fisheye images as they cannot efficiently deal with geometric distortion. Following the well-known structure for multi-scale feature extraction [15, 48], we propose a face detector that not only integrates the low and high-level information but also combines two-dimension features from both the planar and spherical domains of fisheye images.

B. Countermeasures for Distortion in 360-degree Image

The field of view (FOV) of the fisheye lens is greater than 180 degrees, which is beneficial to generating images with more information than the traditional rectilinear ones, but this causes the images to suffer geometric distortion. The distortion becomes more severe from the center to the periphery of the images. Research that aims to solve the distortion problem is currently underway. The most direct way is to correct distorted images to represent characteristics consistent with flat images. The framework in [12] classifies faces according to the degrees of distortion through CNN and then sends them to three networks with different weights for rectification and detection. Though this paper obtained the improved face rectification performance, it ignored the face detection process in the actual recognition application. Moreover, it only divided the image into three discrete distortion-level regions, which did not perfectly reflect the continuity of the degree of distortion with position changing.

Another way to deal with fisheye image distortion is by improving the adaptability of the algorithms directly. In kernel transformer networks [34], based on position information of ERP (equirectangular projection) image, the convolution kernel is modified by an adaptive branch. However, according to the specific network structure, adjustments must be made, which means poor portability that the original network cannot be applied to omnidirectional images of other projection formats (i.e., fisheye images). Su et al. [13] designed the kernel with different shapes according to varying latitudes of pixels, while other works [33, 35] adjusted the kernel on the sphere and performed re-sampling or projected the feature to the tangent plane. Nevertheless, interpolating the planar 2D ERP image while defining the feature on the sphere is irrational [34]. Besides, all of these methods cannot learn the actual geometric-invariant characteristics of fisheye images only by optimizing linear convolution. Cohen et al. [36] extended convolution to the spectral domain, guided by rotation-invariant characteristics extracted by spherical CNNs. Although the convolution speedup benefits from the fast Fourier transform, it is still considered computationally expensive since point multiplication itself is a complicated operation in the frequency domain.

C. Attention Mechanism

Since its introduction, the attention mechanism was gradually developed so as to deal with image-related tasks, which leads networks to focus on the specific position in an image while ignoring the unimportant information. Jaderberg et al. [41] used the affine transformation learned by the convolution module to simulate the rotation characteristics of the image, leading to a network that possesses transformation invariance. This method essentially exploited the spatial attention mechanism. The study in [42] focused on using the channel attention information in the image, and CBAM[43] combined both channel and space attention in a cascade structure to extract features, which encoded the space and channel information as the weights for position or channels, respectively. Diverse weights of convolution kernels in [44] were used for different images, while different types of attention were concentrated through multi-layer attention in [45]. It is worth noting that the long-range dependence in space was captured through the self-attention mechanism and non-local operation operators in [46]. The attention mechanism in these pioneering works has achieved significant advancements, but it has not been applied to mitigate the geometric distortion of fisheye images. Our network, which pulls together feature distributions from both spherical and planar domains employing the attention mechanism, is able to focus on the distorted region.

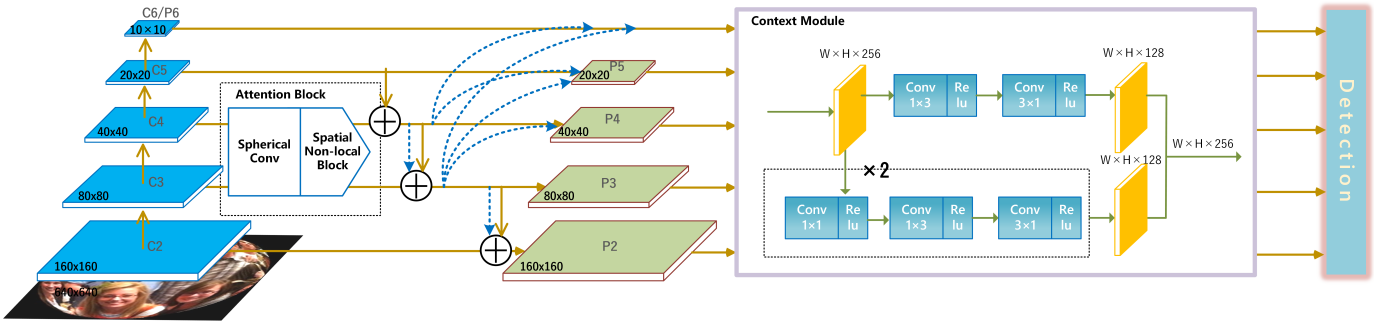

Fig. 1. An overview of the proposed face detector for fisheye images. The flow direction of the features from two domains of fisheye images is shown. The solid orange line denotes the planar feature from the backbone network, and the dotted blue line indicates the spherical domain feature after the attention block. Planar feature maps of middle layers fuse with the spherical-domain features from middle layers and then are added to other layers through path enhancement and up-sampling. Enhanced by the optimized context module, the detector finally carries out the multi-task learning.

III. FACE DETECTOR FOR FISHEYE IMAGE

This section presents the overall architecture of the proposed face detector for fisheye images.

A. Network Structure

Our goal is to design the face detection algorithm for fisheye images, aiming to capture truly spherical representations of fisheye images at a low cost. Fig. 1 shows the architecture of the proposed detector.

Feature Extraction Structure. To extract more accurate features, referring to [27], we use Resnet [7] as the backbone of the detector and extract {c5, c4, c3, c2} layers as the multi-scale features of images. We set the scaling step to 2 (that is, the stride of the last layer of each block is 2) and obtain the feature maps of different sizes via successive convolution layers. The output of the middle layers is led straight into an attention block to extract the features from the spherical domain. We finally simulate the multi-scale Feature Pyramid Network (FPN) [48] network by up-sampling the feature map to complete the feature extraction of the fisheye image.

During the experiment, we found that the medium-sized feature map has a more vital perception of distorted faces in fisheye images, which is more conducive to focus on the position-based distortion: More advanced feature map, with richer semantic information, has a more robust capability to encode semantic information; On the contrary, feature maps of a larger size learn more content information, such as contours and edges, and the ability of which to learn distortion is worse than those of a smaller size. Based on the above observations, considering path enhancement has been used as a feature enhancement method [15] [47] for detecting objects, we add it to the feature pyramid network to get features among middle levels (including the planar and spherical domain information) propagated to features of upper layers. It obviously shortens the message path among layers. More importantly, path enhancement avoids the repetitive spherical convolution for high-level feature maps, reducing the amount of calculation and forcing the training pace. At the same time, up-sampling also transfers the middle-layer information to the bottom-layer feature maps to achieve information sharing. Then the multi-layer features are sent to the context module to be enhanced.

Context Network Module. We use feature extraction branches with kernels in 3x3 and 5x5 scales, as shown in Fig. 1, to improve the ability to extract contextual information through receptive fields of different sizes. One of the main concerns of face detection is to improve the detection accuracy of smaller faces. Consequently, we do not adopt 7x7 or even larger convolution kernels in consideration of reducing the complexity of the network since large-scale convolution contributes little to detecting tiny faces. To reduce the number of parameters, we also adopt the same idea as [30], converting the 5x5 convolution into two concatenated 3x3 convolutions layers. In addition, we perform convolution operations in the frequency domain and spatial domain separately as MobileNetV2 [49]. First, feature maps of each channel are convolved separately, then 1x1 feature extraction is performed on the feature maps across these channels. Besides, the 3x3 convolution kernel is further divided into two, 1x3 and 3x1, maintaining the same receiving field while minimizing the number of parameters and computational complexity. The context module processes the feature map of each layer and finally sends them to the detection module, predicting the coordinates and categories after bounding box head and regression head (1x1 convolution layer).

B. Attention Block based on Spherical Convolution

In the attention block, shown in Fig. 2, the planar feature map is sent to the spherical convolution network, then the extracted feature map and the input are fused to derive the bi-level feature map computed by the pairwise function f.

The spatial attention operation in the spherical convolution block is defined as

$$\mathrm{y}_i = softmax \sum\nolimits_{\forall j} f(x_i, x_j) g(x_j), \tag{1}$$

where x_i, x_j represent the planar image feature map for the point i and the spherical feature map for point j, and y_i denotes the planar feature map based on spherical attention, with the same size as x_i. We adopt a single-layer convolution: $g(\mathrm{x}_j) = \mathrm{W}_g^T x_j$ instead of linear embedding, where W_g^T is the

weight vector that encodes a planar image as the representation of input signal. Furthermore, we use the concatenation version of function f described in [46], which captures the long-distance dependence of the specific point x_i in the planar feature map with all other points. Note that the other points belong to the feature maps extracted from spherical convolution, which contains distortion information. Hence:

$$f(x_i, x_j) = \text{LeakyReLU}(W_f^T[\theta(x_i), \varphi(x_j)]) \tag{2}$$

As the spherical convolution manages the rotation-invariant signals on the spherical domain, we denote spherical CNNs as $\varphi(x_j)$ to catch the attention of the spherical signal, which employs two consecutive layers of spherical convolution as shown in Fig. 2. By focusing on the trade-off between the amount of calculation and the performance, this block only employs two layers of the spherical convolution, transforming the feature map into S2 and SO(3) domains for convolution operation. The first convolution layer is carried out on spherical domain, with the H×W feature map in the plane coordinate system converted to that in size of $\alpha \times \beta$ in the spherical coordinate system. In the second layer, the conversion to the SO(3) domain with an $\alpha \times \beta \times \gamma$ output is performed on the spherical domain. $\theta(x_i)$= $\mathrm{W}_\theta^T x_i$ represents a layer of traditional planar

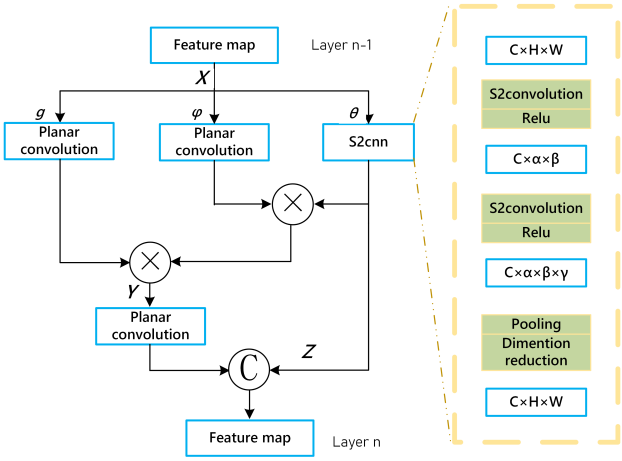

Fig. 2. Attention block based on spherical convolution. X, Y, and Z denotes input, planar feature map enhanced by spherical information, and spherical feature map. $\otimes$ denotes matrix multiplication, and © denotes concatenation in the channel direction. C, H, and W characterize the feature maps on the planar domain while α, β, γ denotes those on the spherical domain.

convolution; LeakyReLU serves as the activation function to achieve improved smoothness.

In terms of self-attention, $\varphi(x_j)$ here is considered to be the query, which is extracted by spherical convolution, and $\theta(x_i)$ here is regarded as the key for distortion degrees. By calculating the similarity of the query and the key, we can make the point i perceive the area with similar distortion characteristics at the pixel level. Furthermore, the direction and degree of distortion at different positions are determined. The object will show a particular law of deformation at a specific position. Spherical convolution is used to capture the distorted area related to position i and express the relevant information in the form of weights, which is used to generate y_i through multiplication with $g(x_i)$. Then we concatenate y_i and $\theta(x_i)$ along the channel direction to get the final bi-domain output.

In the implementation, as presented in Subsection A, the distortion is mainly concentrated in the middle feature layers while the higher and lower layers contribute less. Our attention block only extracts feature maps from the intermediate layers as input, as shown in Fig. 1, and then merges them with that from other layers through path enhancement and up-sampling.

C. Training

Implementation Details. In the feature extraction network, we use Resnet50 pre-trained in Imagenet-11k as the backbone. Similar to Single-stage headless (SSH) face detector [30], we use step sizes of {4, 8, 16, 32, 64} to detect faces with different scales on the feature maps with the sizes set to {160, 80, 40, 20, 10}. For all the anchors generated by the network, 1:1 is used as the aspect ratio, and 0.5 is set as the positive threshold; that is, an anchor will be judged as a positive sample when its intersection over union (IOU) with the certain ground truth is greater than this value. Correspondingly, the threshold of negative samples is set to 0.3. In loss head (1x1 convolution layer), the weights of the network are shared among the feature layers of multiple scales {p5, p4, p3, p2}.

Loss Function. The loss function adopts the usual face detection loss function, multi-task loss, including classification and regression.

$$L = \sum_{\forall i} L_{cls}(p_i, p_i^*) + \lambda \sum_{\forall i} \mathrm{I}(p_i^* = 1) L_{box}(t_i, t_i^*) \tag{3}$$

where i represents each possible anchor predicted under multiple scales. Regarding the classification loss $L_{cls}(p_i, p_i^*)$, p_i denotes the predicted probability of the anchor i, and p_i^* has two values $\{0, 1\}$, corresponding to anchor i whether being positive or negative. Since face detection belongs to the class of binary classification problems, we use SoftMax as the loss function. Correspondingly, coordinate prediction is a numerical regression problem, so we adopt SmoothL1 Loss as the loss function $L_{box}(t_i, t_i^*)$. $\mathrm{I}(\cdot)$ is a conditional judgment function, which indicates the predicted coordinate value, that will only be regressed when the anchor is positive. We parameterize the coordinates of the left-top point as well as both length and width of the anchor, then convert them to log-space. $t_i = \{t_x, t_y, t_w, t_h\}$ and $t_i^* = \{t_x^*, t_y^*, t_w^*, t_h^*\}$ respectively denote the predicted value and the ground truth of the coordinate towards the specific positive anchors. λ is a loss parameter used to weaken the imbalance between the number of positive and negative anchors.

IV. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed face detector on fisheye images.

A. Experimental Settings

The network is trained on four NVIDIA Tesla P40 (24G) GPUs with the batch size of 4x8. We use stochastic gradient

descent with momentum set to 0.9 as the optimization method for the entire network. In different training stages, we use a gradually decayed (set to $5e^{-4}$) learning rate to update the weights. The learning rate is initialized as $1e^{-3}$ and descends by a factor of 10 after $\{190, 220\}$ epochs for Mobilenet and $\{70, 90\}$ epochs for Resnet. The training process terminates at 250 epochs for Mobilenet and 100 epochs for Resnet.

B. Datasets

We train the network with the training set of Wider-360 and use the full set of FDDB-360 and the test set of Wider-360 as the basis for evaluating network performance. The test method is the same as that of the planar image dataset. We also use the original Wider Face [8] to assess the performance of our network on planar images.

FDDB-360 Dataset. This dataset includes a collection of images from the face detection dataset and benchmark (FDDB) that have been processed to look like fisheye images from a typical 360-degree camera [11]. It consists of 17052 images with 26640 faces.

Wider-360 Dataset. This dataset is created by post-processing the planar rectilinear images collected from the well-known dataset Wider Face, using a projection model [37] that maps planar rectilinear images to fisheye ones. It contains 63897 fisheye images, of which the training set occupies 50982, and the test set occupies 12915. Unlike the traditional image dataset version, all of its data are not divided into three subsets (easy, medium, and hard), but merged into one dataset. Hence, its detection difficulty is approximately equivalent to that of the medium set of a Wider Face. This dataset can not only show the detector's ability to capture distorted faces but also bring out the network's effectiveness for small faces.

C. Ablation Study

Effectiveness of Attention Block and Context Module. As shown in Table I, we evaluate the performance of the proposed network under different settings on the Wider-360 validation set and the FDDB-360 dataset. It can be seen that the accuracy of the network with attention block on the Wider-360 dataset significantly improves the face box average precision (AP) 3.85% by the use of Resnet and 1.93% by the use of Mobile-net, suggesting that attention block is indispensable for detecting distorted faces. The contribution to the accuracy of the FDDB dataset is numerically smaller due to the dataset is relatively easy to detect. However, the attention block still improves the performance in the case of the two backbones, with the AP value of 99.1% for Resnet and 98.1% for Mobile-net. We should emphasize that our test on the FDDB-360 dataset uses the model trained on Wider-360, so the results indicate the excellent generalization ability of the model. Nevertheless, adding a context module can further improve the AP on Wider-360 by 0.44% for Resnet and 0.41% for Mobile-net compared to adding the attention block only.

Fig. 3 shows several detection examples. The red box marks the faces detected by the network. Larger and more regular faces closer to the middle of the image are detected by both detectors. In areas with a large degree of distortion, our network (Fig. 3 right) always has a better detection effect for more severely distorted faces and can maintain the same effect when the size of the face is small. It should be noted that the spatial distribution of the network we designed (Fig. 4 right) is more uniform, which further shows that our network pays more attention to the distortion characteristics of the sphere, while the traditional network has a higher error rate in severely distorted positions.

TABLE I
ABLATION STUDY FOR ATTENTION BLOCK (AB) AND CONTEXT MODULE (CM).

Method	Backbone	FDDB-360	Wider-360
FPN	Mobile-net	97.8%	55.43%
FPN+AB	Mobile-net	98.1%	57.36%
FPN+CM	Mobile-net	97.8%	55.92%
FPN+AB+CM	Mobile-net	98.2%	57.77%
FPN	Resnet	98.5%	64.34%
FPN+AB	Resnet	99.1%	68.19%
FPN+CM	Resnet	98.7%	64.87%
FPN+AB+CM	Resnet	99.1%	68.63%

Fig. 3. Fisheye image face detector results by FPN (left) and FPN+Attention Block (right).

D. Effectiveness of the Proposed Network

Since Wider-360 does not provide the standard evaluation protocol, we reproduced several current state-of-the-art face detection networks, namely SSH [30], Retinaface [27], S^3FD [31], to test their performance on the planar dataset Wider Face and fisheye dataset Wider-360 in the same experimental environment. Like the original dataset, Wider-360 contains plenty of tiny faces, many of which are concentrated in severe-distorted areas, so the detection task on Wider-360 is considerably more difficult.

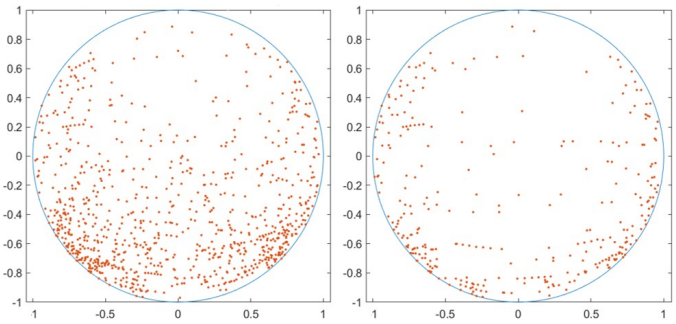

Fig. 4. False negative distribution for using attention block (right) or not (left)

TABLE II
COMPARISON OF THE DETECTORS

method	Wider-360	Wider Face		
		Hard	Medium	Easy
SSH [30]	62.77%	81.45%	88.97%	91.88%
Retinaface [27]	64.34%	83.57%	90.44%	93.52%
S^3FD [31]	65.72%	84.16%	90.53%	93.49%
Ours	70.19%	85.72%	90.56%	93.67%

Table II shows that our approach outperforms other methods on fisheye images with all networks trained on Wider-360. More specifically, the proposed network produces the best AP value of 70.19% for Wider-360, 4.47% higher than S^3FD with the best results in other planar networks. At the same time, the performance of the proposed approach for the planar dataset is slightly better, with the AP of 93.67% on the easy subset and 85.72% on the hard subset, which indicates the attention mechanism in our network can also capture some of the rotation features in the planar image. As shown in Fig. 5, on the simpler FDDB-360 dataset, our method (blue curve) also shows better Precision-Recall performance than other networks.

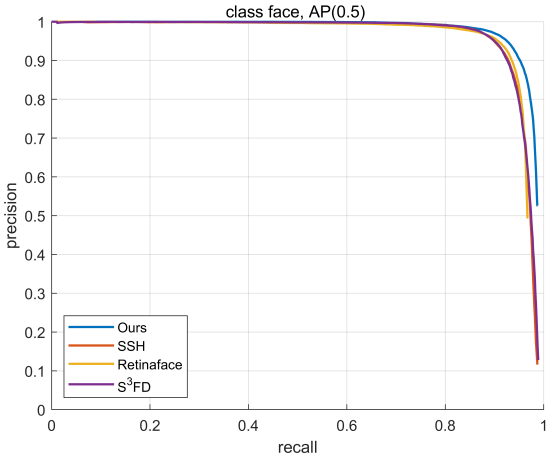

Fig. 5. Comparison of our network on the FDDB-360 with other networks.

Time. In terms of operating efficiency, spherical convolution suffers high complexity, so we need to evaluate the inference speed of the network. The test environment uses Tesla P40 and cuDNN v7.6.3 as well as Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz. Our detector runs at 30 FPS and achieves real-time detection performance.

V. CONCLUSION

In this work, we proposed a face detector for fisheye images, which alleviates fisheye image distortion and improves detection accuracy. We propose to utilize spherical convolution to extract the rotation-invariance features of the fisheye image on the spherical domain and fuse it with the planar features through the attention mechanism. The path enhancement and up-sampling strategy can fuse the spherical and planar information between feature maps of different scales. In addition, our context awareness module also enhances the feature extraction effect of the detector. Experimental results show that, while maintaining the accuracy and inference speed of traditional algorithms, the proposed face detector achieves better performance than traditional networks in face detection tasks for both fisheye and planar images.

REFERENCES

[1] S. Yogamani, C. Hughes, J. Horgan, G. Sistu, P. Varley, D. O'Dea, et al., "WoodScape: A multi-task multi-camera fisheye dataset for autonomous driving," IEEE International Conference on Computer Vision (ICCV), 2019.

[2] H. Kim, E. Chae, G. Jo and J. Paik, "Fisheye lens-based surveillance camera for wide field-of-view monitoring," IEEE International Conference on Consumer Electronics (ICCE), 2015.

[3] P. Viola and M. Jones. "Rapid Object Detection using a Boosted Cascade of Simple Features," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2001.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[5] R. Girshick, "Fast R-CNN," in IEEE International Conference on Computer Vision (ICCV), 2015.

[6] S. e. a. Ren, "Faster R-CNN: Towards real-time object detection with region proposal networks," Advances in neural information processing systems (NIPS), 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[8] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[9] Z. Xue, N. Xue, G.-S. Xia, and W. Shen, "Learning to calibrate straight lines for fisheye image rectification," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[10] M. Zhang, J. Yao, M. Xia, K. Li, Y. Zhang, and Y. Liu, "Line-based multi-label energy optimization for fisheye image rectification and calibration," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[11] J. Fu, S. R. Alvar, I. V. Bajic, and R. G. Vaughan, "FDDB-360: Face detection in 360-degree fisheye images," IEEE Conference on Multimedia Information Processing and Retrieval (CMIPR), 2019.

[12] Y. -H. Li, I. -C. Lo and H. H. Chen, "Deep Face Rectification for 360° Dual-Fisheye Cameras," in IEEE Transactions on Image Processing, vol. 30, pp. 264-276, 2021.

[13] Y.-C. Su and K. Grauman, "Flat2sphere: Learning spherical convolution for fast features from 360° imagery," Advances in neural information processing systems (NIPS), 2017.

[14] Q. Zhao, C. Zhu, F. Dai, Y. Ma, G. Jin, Y. Zhang. "Distortion-aware CNNs for Spherical Images," In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2018.

[15] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. "Path Aggregation Network for Instance Segmentation," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[16] H. A. Rowley, S. Baluja and T. Kanade, "Neural network-based face detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23-38, Jan. 1998.

[17] H.A. Rowley, S. Baluja, T. Kanade, "Rotation invariant neural network-based face detection," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 1998.

[18] S. Z. Li, L. Zhu, Z.Q. Zhang, A. Blake, H. J. Zhang, and Harry Shum. "Statistical Learning of Multi-view Face Detection," In Proceedings of the European Conference on Computer Vision (ECCV), 2002.

[19] H. Li, Z. Lin, X. Shen, J. Brandt and G. Hua, "A convolutional neural network cascade for face detection," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[20] P. Hu and D. Ramanan, "Finding tiny faces," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[21] S. Yang, P. Luo, C. C. Loy and X. Tang, "Faceness-Net: Face Detection through Deep Facial Part Responses," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 8, pp. 1845-1859, 1 Aug. 2018.

[22] P. Viola and M. Jones, "Robust real-time face detection," IEEE The International Conference on Computer Vision (ICCV), 2001.

[23] B. Yang, J. Yan, Z. Lei and S. Z. Li, "Aggregate channel features for multi-view face detection," IEEE International Joint Conference on Biometrics (IJCB), 2014.

[24] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," IEEE International Conference on Image Processing (ICIP), 2002.

[25] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," in IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, Oct. 2016.

[26] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "DenseBox: Unifying Landmark Localization with End to End Object Detection," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[27] J. Deng, J. Guo, E. Ververas, I. Kotsia and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[28] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified real-time object detection," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[29] W Liu, D Anguelov, D Erhan, C. Szegedy, S. Reed et al., "SSD: single shot multibox detector," In Proceedings of the European Conference on Computer Vision (ECCV), 2016 .

[30] M. Najibi, P. Samangouei, R. Chellappa and L. S Davis, "Ssh: Single stage headless face detector," IEEE The International Conference on Computer Vision (ICCV), 2017

[31] S. Zhang, X. Zhu, Zh. Lei, H. Shi, X. Wang and S. Z. Li, "Single Shot Scale-Invariant Face Detector," IEEE The International Conference on Computer Vision (ICCV), 2017.

[32] X. Tang, D. K. Du, Z. Q. He, and J. T. Liu, "Pyramidbox: A context-assisted single shot face detector," In Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[33] B. Coors, A. Paul Condurache, and A. Geiger. "Spherenet: Learning spherical representations for detection and classification in omnidirectional images," In Proceedings of the European Conference on Computer Vision (ECCV), 2018

[34] Y.-C. Su and K. Grauman, "Kernel transformer networks for compact spherical convolution," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[35] Z. Zhang, Y. Xu, J. Yu and S. Gao, "Saliency Detection in 360° Videos," In Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[36] T. Cohen, M. Geiger, J. Kohler, M. Welling, "Spherical cnns," International Conference on Learning Representations (ICLR), 2018.

[37] J. Fu, I. V. Bajić, and R. G. Vaughan, "Datasets for face and object detection in fisheye images," Data in Brief, 2019.

[38] D. Britz, A. Goldie, M. -T. Luong, and Q. V. Le, "Massive Exploration of Neural Machine Translation Architectures," CoRR abs/1703.03906, 2017.

[39] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," Advances in neural information processing systems (NIPS), 2015.

[40] A. Agrawal, D. Batra, and D. Parikh, "Analyzing the behavior of visual question answering models," arXiv:1606.07356, 2016.

[41] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, "Spatial transformer networks," Advances in neural information processing systems (NIPS), 2015.

[42] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[43] S. Woo, J. Park, J. Lee et al., "CBAM: Convolutional Block Attention Module," In Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[44] X. Li, W. Wang, X. Hu and J. Yang, "Selective Kernel Networks," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[45] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, et al., "Residual attention network for image classification," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[46] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[47] G. J et al., ultralytics/yolov5: v4.0, Jan. 2021, [online] Available: https://doi.org/10.5281/zenodo.4418161.

[48] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan and S. J. Belongie, "Feature pyramid networks for object detection," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[49] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2018.