RepF-Net: Distortion-aware Re-projection Fusion Network for Object Detection in Panorama Image

Mengfan Li[1], Ming Meng[2] ⋆, and Zhong Zhou[1]

[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2] School of Data Science and Media Intelligence, Communication University of China, Beijing, China

Abstract. Panorama image has a large 360° field of view, providing rich contextual information for object detection, widely used in virtual reality, augmented reality, scene understanding, etc. However, existing methods for object detection on panorama image still have some problems. When 360° content is converted to the projection plane, the geometric distortion brought by the projection model makes the neural network can not extract features efficiently, the objects at the boundary of the projection image are also incomplete. To solve these problems, in this paper, we propose a novel two-stage detection network, RepF-Net, comprehensively utilizing multiple distortion-aware convolution modules to deal with geometric distortion while performing effective features extraction, and using the non-maximum fusion algorithm to fuse the content of the detected object in the post-processing stage. Our proposed unified distortion-aware convolution modules can be used to deal with distortions from geometric transforms and projection models, and be used to solve the geometric distortion caused by equirectangular projection and stereographic projection in our network. Our proposed non-maximum fusion algorithm fuses the content of detected objects to deal with incomplete object content separated by the projection boundary. Experimental results show that our RepF-Net outperforms previous state-of-the-art methods by 6% on mAP. Based on RepF-Net, we present an implementation of 3D object detection and scene layout reconstruction application.

1 Introduction

In recent years, panorama image has been widely used, and their 360° large field of view(FOV) provides rich contextual information for computer vision processing. As a fundamental task in computer vision, accurate object detection results enable subsequent applications such as virtual reality, augmented reality, and scene understanding to achieve better performance.

Before deep learning, in order to handle the object detection task, traditional object detection methods are usually subdivided into three steps: information region selection, feature extraction, and classification[32]. In the information region

⋆ Corresponding Author: mengming@buaa.edu.cn selection stage, a multi-scale sliding window is used to scan the entire image. And then feature extraction algorithms such as histogram of oriented gradients[7] and haar-like[17] are used to generate semantic and robust image representations. Finally, classification algorithms such as support vector machine[10] are chosen as the classifier.

With the development of the convolutional neural network in computer vision, there are two types of network architectures for object detection. In the first type, the network has two stages: a regional proposal generation network to replace both information region selection and feature extraction stage and a classification network. On the other hand, the second network has only one stage, which integrates feature extraction and classification, and uses anchors for informative region selection.

Although the convolutional neural networks have better performance than traditional object detection methods, object detection in panorama image remains challenging due to the sphere-to-plane projections. First, the geometric deformation brought by sphere-to-plane projections makes it difficult to extract features effectively. Second, sphere-to-plane projections also divide the original complete context information and make the object information incomplete on the projection boundary.

In this paper, we propose a re-projection fusion object detection network architecture RepF-Net, and perform better accuracy on panorama image than previous state-of-the-art methods. We propose a unified distortion-aware convolution module in the convolutional layer of our network architecture, which can both deal with equirectangular projection deformation in the information region selection stage and stereographic projection deformation in the feature extraction stage. Moreover, we propose a non-maximum fusion algorithm in the post-processing stage of our network architecture, which can fuse the incomplete information caused by the sphere-to-plane projection boundary.

Our contributions can be summarized as follows:

- We propose a re-projection fusion object detection network architecture RepF-Net for panorama image, utilizing multiple distortion-aware modules to perform effective feature extraction, while using re-projection and non-maximum fusion in the post-processing stage to obtain better performance.
- We propose a unified distortion-aware convolution module to handle various geometric distortions caused by geometric transforms and projection models. It makes our network focus on the information areas to extract features more efficiently, resulting in faster convergence and better performance. We propose a non-maximum fusion algorithm to handle the object incomplete problem caused by the projection boundary to obtain better detection.
- We conduct numerous ablation experiments and comparison experiments to verify the effectiveness of our proposed methods. Meanwhile, our proposed RepF-Net outperforms the state-of-the-art by 6% on mAP. Furthermore, we present an implementation of 3D object detection and scene layout reconstruction application based on our methods.

2 Related Work

CNN and Object Detection: With the application of convolutional networks, there are two main network architectures for object detection: one-stage detection and two-stage detection. The two-stage detector adopts the R-CNN architecture, and followed by its variants FastR-CNN[13], FasterR-CNN[24] and MaskR-CNN[15]. The two-stage detector first gets candidate proposals through a region proposal network(RPN), and then refines the proposals through a classification network to obtain the final detection results. On the other hand, the one-stage detector based on global regression and classification, uses pre-defined anchors instead of RPN-generated region proposals, allowing bounding boxes with relevant classes to be extracted directly from the input image. Mainstream object detection methods based on this architecture include You Only Look Once(YOLO)[21–23, 2] and Single Shot Detection(SSD)[18]. There are also some object detection network architectures that do not rely on proposals or anchors, such as CornerNet[16] which directly detects the corners of the object, while CenterNet[9] directly detects the center of the object. These detectors are less accurate due to the lack of prior information on proposals and anchors.

CNN on Panorama Image: To make the convolution module extract features more efficiently on panorama image, deformable convolution(DeformConv) is proposed[6]. And Zhu et al.[34] further improved the deformable convolution to solve the problem of useless context regions interfering with feature extraction. While CNNs are able to learn invariance to common object transformations and intra-class variations, they require significantly more parameters, training samples, and training time to learn invariance to these distortions from the data. Meanwhile, Cohen et al.[5] proposed to use spherical CNN for classification and to encode rotational invariance into the network. However, overfitting combined with full rotation invariance reduces the discriminative power. In contrast, Benjamin et al.[4] encoded geometric distortions into convolutional neural networks, which are more compatible with existing CNN architectures and achieve better performance. And Clara et al.[12] directly improved deformable convolution and proposed equirectangular convolution(EquiConv), which is specially designed to eliminate geometric distortion under equirectangular projection. Similarly, orthographic convolution(OrthConv)[20] is designed to remove geometric distortions in orthographic projection.

Object Detection on Panorama Image: Deng et al.[8] first attempted to use existing object detection methods for object detection on panorama image. Due to the simplicity of converting a sphere into a Cartesian grid, equirectangular projection has been used as the primary sphere-to-plane projection method for projecting 360° content. However, the equirectangular projection applied to panorama image produces distortions leading to geometric deformation, which leads to different approaches to maintain performance. There are mainly two different approaches. The first approach proposes a multi-projection variant of the YOLO detector[28], which attempts to handle the geometric deformation problem with multiple stereographic projections. On this basis, Pengyu et al.[31] further optimized the parameters of multi-view projection and obtained better performance. On the other hand, the second approach optimizes the convolution layers by applying distortion-aware convolution modules, which handles the geometric deformation in the feature extraction stage[14].

Our method integrates these two main approaches, through a combination of multi-projection and distortion-aware convolution modules to deal with geometric distortions. In the stage of generating the candidate proposals of the projection area, we comprehensively use EquiConv and DeformConv for efficient feature extraction, and in the detection stage, we use a convolution module that can efficiently handle stereographic projection distortions. Moreover, in the post-processing stage, we fuse the re-projection detection results to handle the influence of the projection boundary to get the final results.

3 Method

Our goal is to design a network architecture for object detection from panorama image. Based on the trade-off of distortion reduction and efficiency improvement, we use the re-projection two-stage detector as our base network architecture. Before introducing our network, we first introduce our proposed unified distortion-aware convolution operator for general geometric distortions in Sec.3.1. Then in Sec.3.2, we introduce our proposed non-maximum fusion algorithm to fuse incomplete object content caused by the projection boundary. Subsequently, in Sec.3.3, we describe the architecture of our proposed network, which combines multiple distortion-aware convolution modules in the feature extraction stage and the non-maximum fusion algorithm in the post-processing stage.

3.1 Unified Distortion-Aware Convolution

Zhu and Dai et al.[6, 34] implement the convolution modules by adding additional parameters on the kernel offset, which can also be learned by the network. Therefore, the ability to learn object shape and deformation enables deformation convolution to extract features more efficiently. Although the offset parameters can be learned by training the network, they can also be calculated in advance for known geometric distortions[5, 12, 20]. Inspired by these works, we propose a unified distortion-aware convolution module, which can deal with all kinds of known geometric distortions.

The standard convolution sample a set of positions on the regular grid $R = \{(-1, -1), (-1, 0)..., (0, 1), (1, 1)\}$ as the convolution kernel, for each position p_0, the operation result of the regular grid structure is assigned to the corresponding element of the output feature map f_{l+1} of the l+1th layer, where p_0+p_n indicates that the sampling position p_n enumerates the relative position of the pixels in the convolution region R, while the deformable convolution improves the feature extraction capability by adding an offset $\triangle p_n$ in the convolution region:

$$f_{l+1} = \sum_{p_n \in R} w(p_n) \cdot f_l(p_0 + p_n + \triangle p_n). \tag{1}$$

Because (θ, ϕ) is the coordinate in the spherical domain without distortion, and (x, y) is the projection plane coordinate with geometric distortion, so once we have the conversion formula between the two coordinates, which is the projection formula, we can get a deformable convolution module for the certain geometric distortion.

First, we need to calculate $(\triangle\theta, \triangle\phi)$ according to the size of the current convolutional feature layer s, and the conversion formula are represented as $xy2\theta\phi$ and $\theta\phi 2xy$:

$$\triangle\theta, \triangle\phi = \frac{xy2\theta\phi(s, s) - xy2\theta\phi(0, 0)}{s}. \quad (2)$$

After that, we can calculate the offset of the current convolution kernel according to the position (x, y) of the convolution kernel in the feature layer:

$$\theta, \phi = xy2\theta\phi(x, y), \triangle x, \triangle y = \theta\phi 2xy(\theta + \triangle\theta, \phi + \triangle\phi). \quad (3)$$

Finally, we can calculate the offset applied to deformable convolution, which is an offset relative to the original convolution kernel, not relative to the image domain itself, while s_f represents the size of the feature map:

$$R_{offset} = (\{(\triangle x, \triangle y)...\} - \{(x, y)...\}) * (s_f, s_f) - \{(-1, -1), ..., (1, 1)\}. \quad (4)$$

As shown in Fig. 1 (a), we show the projection of a panorama image of the sphere onto the tangent plane. We assume that the radius of the sphere is $r = 1$, the viewpoint V is at $(1, 0, 0)$, the projection direction is towards the negative X-axis, and the center of the tangent plane is at $x = (-1, 0, 0)$. Now, the values of the point $PP(X, Y)$ on the projection tangent plane are projected from $P(\theta, \phi)$ on the sphere as:

$$\frac{d+1}{d+\cos\phi} = \frac{-X + s/2}{\sin\phi}, \frac{d+1}{d+\cos\theta} = \frac{-Y + s/2}{\sin\theta}. \quad (5)$$

While s represents the size of the projection plane, the original coordinates of the point $PP(X, Y)$ are $(-1, y, z)$. And Equ. 5 is the stereographic conversion formula.

We adopt the stereographic projection model which means d is constantly equal to 1. By substituting into the stereographic projection model, we can obtain stereographic convolution(SteConv), which removes the stereographic projection distortion. The comparison between the kernel sample region of SteConv and standard convolution(StdConv) is shown in Fig. 1, (b) shows the sampling effect of the two convolution kernels, and (c) shows the comparison of the effects of the two kernels with different dilation and kernel size. Moreover, we substituted the equirectangular projection formula and implement a unified equirectangular convolution.

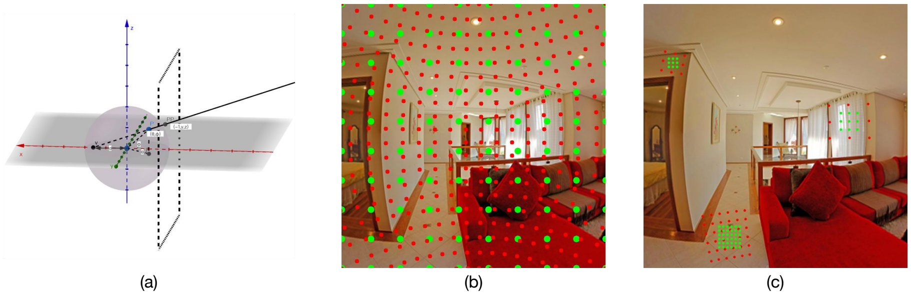

Fig. 1. Visualization of the stereographic projection model and stereographic convolution. Green - StdConv, red - SteConv.

3.2 Non-maximum Fusion

In the state-of-the-art object detection pipelines, region proposals generated by convolutional neural networks replace traditional sliding windows, but multiple proposals often regress to the same region of interest. Hence, it is necessary to use non-maximum suppression(NMS) as a post-processing step to obtain the final detection as it significantly reduces the number of false positives. As an important part of the object detection pipeline, NMS first sorts all detection boxes according to their scores, and selects the detection box with the maximum score, while suppressing all other detection boxes whose overlapping score exceeds the predefined threshold. This process is recursively applied to all detection boxes.

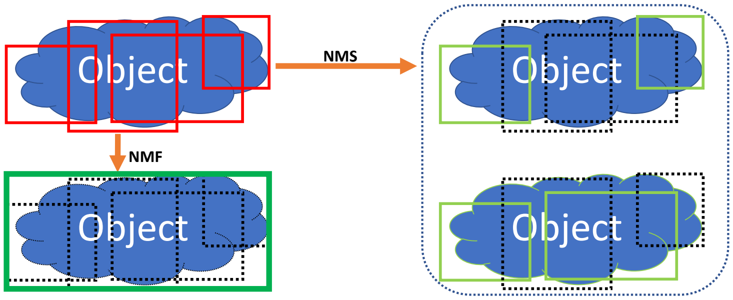

Fig. 2. Schematic illustration of NMS vs. NMF. Red - origin detect boxes, light green - NMS result, dark green - NMF result, black dotted line - boxes which be suppressed.

The main problem of NMS is that it directly suppresses adjacent detection boxes. For the re-projection two-stage detection algorithm, there is no detection box located at the maximum value in re-projection detection boxes, while each detection box is a part of the detected object. Therefore, according to the design of the algorithm, after applying NMS to the re-projection detection box, the original complete object is detected as multiple continuous detection boxes or multiple incomplete components, which will lead to a decrease in average precision. This is because the NMS algorithm is designed to process the output value of the neural network, and only takes the local maximum value as the final detection output, and when processing the re-projection detection boxes, what we need is to associate a detection box cluster with an object. The multiple detection boxes in one cluster are fused, so as to fuse multiple incomplete parts of the object into the final detection box. While this problem can not be solved by NMS, even with some improvements to it[3, 25]. We show an illustration of the problem in Fig. 2.

To this end, we propose a non-maximum fusion(NMF) algorithm, which improves the original NMS algorithm and fuses all detection boxes that have overlapping relationships with the maxima instead of direct suppression. The steps of the NNF algorithm are described as follows:

```
program non_max_fusion (B={b1, ... b_n}, S={S1, ... Sn}, Nt)
  {
    B is the list of initial detection boxes.
    S contains corresponding detection scores.
    Nt is the NMF threshold.
  };
begin:
  F ← {};
  while B is not empty do:
    m ← argmax S;
    C ← {bm};
    B ← B - C;
    for bi in B do:
      if iou(C, bi) > Nt then:
        B ← B - {bi}; S ← S - {si};
        C ← C U {bi};
      end
    end
    F ←  F U fusion(C);
  end
  return F, S;
end.
```

The NNF algorithm leads to improvements in average precision measured over multiple overlap thresholds for re-projection two-stage object detectors. Since the NMF algorithm does not require any additional training and is simple to implement, it can be easily integrated into the object detection pipeline.

3.3 Re-projection Fusion Network Architecture

In the detection step, it is a common consensus that two-stage detectors can achieve higher accuracy. The first stage is the multi-view projection region proposal network(MVP-RPN), which can efficiently generate proposals on equirectangular projection images, and the second stage is the stereographic convolutional detector(SteNet), which can accurately refine proposals based on stereographic projection images. Projection region of interest align(PRoI-Align) is additionally introduced to bridge the multi-view projection region proposal network and the stereographic convolutional detector, by transforming proposals into projection field of view to obtain fixed-size stereographic projection images as input to SteNet. In the post-processing step. The detection boxes are first preprocessed using re-projection, and then non-maximum fusion is used to obtain the final detection. The overall architecture of RepF-Net is shown in Fig. 3.

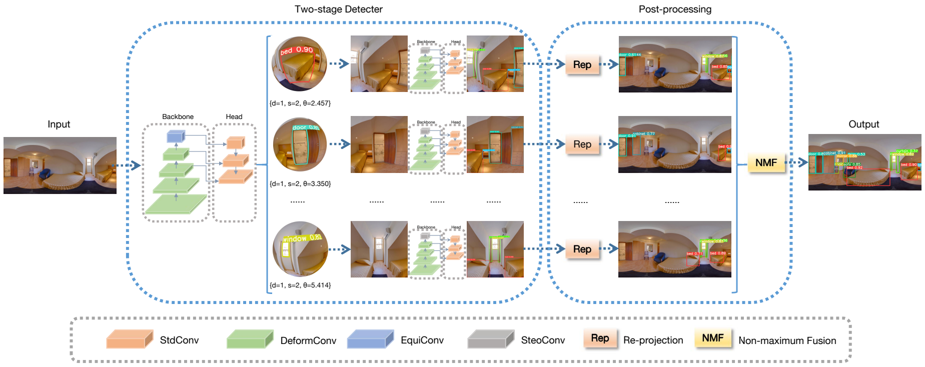

Fig. 3. This figure visualizes the two-stage network and post-processing architecture of RepF-Net.

MVP-RPN: Given a panorama image, MVP-RPN generates the objectness score for each candidate region proposal from its equirectangular projection representation. Different from ordinary RPN[24], in order to handle the geometric distortion brought by equirectangular projection, MVP-RPN comprehensively applies deformable convolution and equirectangular convolution in the backbone network to efficiently extract a distortion-aware feature map. Finally, MVP-RPN generates the position of the selected region proposal as the input for the next stage.

PRoI-Align: Given the region proposals generated by MVP-RPN, PRoI-Align converts the location information of region proposals into three-dimensional FOV parameters (d, s, θ), where d represents the distance of the projection plane from the sphere center, which is inversely proportional to region proposal size, s represents the size of the projection plane, which is directly proportional to region proposal size, θ represents the rotation angle of the projection plane relative to the sphere plane, which constitutes a one-to-one mapping relationship from the horizontal position of the region proposals. The three-dimensional FOV parameters can be substituted into the stereographic projection formula to obtain fixed-size stereographic projection images as the input of the next stage.

SteNet: Given the fixed-size stereographic projection images generated by PRoI-Align, SteNet applies another detection network to further localize region proposals. The same as MVP-RPN, SteNet comprehensively applies deformable convolution and stereographic convolution in the backbone network to efficiently extract a distortion-aware feature map, and offset the geometric distortion brought by stereographic projection. In the end, SteNet refines the detection box of the selected region proposal as the input for the next stage.

Post-processing: In the first step, the post-processing stage re-projects the detection box onto the equirectangular image as the input for the next stage. Non-maximum fusion is then applied to reduce the number of false positives. Since the incomplete object content has been fused in the post-processing stage, the final detection results of the network architecture are obtained.

4 Experiments

In this section, we conduct numerous of experiments aimed at evaluating the effectiveness of our proposed method for object detection in panorama image. We first describe our collection and extension of the datasets. Then explain the implementation details of the experiment, including training and development strategy. Next, our proposed unified distortion-aware convolution module achieves better performance through qualitative and quantitative comparative evaluation. After that, we verified our proposed non-maximum fusion algorithm through ablation experiments, which can achieve better performance in the post-processing stage. Finally, we compare our method with other state-of-the-art methods of object detection in panorama image and find that our method can outperform them.

4.1 Dataset

Collecting high-quality datasets with a sufficient number of images and the corresponding object detection groundtruth is critical for training complex models. However existing equirectangular projection image datasets, including Sun360[27], PanoContext[30], SunCG[26], Stanford2D3D[1], and Structured3D[33], all lack standard object detection annotations. We define a dataset annotation protocol for object detection through protobuf, and according to the protocol converting equirectangular projection image annotation from the above datasets[27, 30, 26, 1, 33]. Simultaneously, we use the projection parameters $\{d = 1; s = 2, 3; \triangle\theta = 0, \frac{\pi}{24}, \frac{\pi}{12}, \ldots, 2\pi\}$ in the stereographic conversion formula to convert stereographic projection image annotation. We also made corrections for low-quality images in the original dataset, as well as wrong object annotations. Finally, the dataset we constructed contains 3423 equirectangular category annotations and 69760 stereographic category annotations. With the definition of the protocol, our dataset can be conveniently applied to various experiments and tasks. The split strategy of train/validation/test for the dataset is similar to [14], the dataset is divided into 85% for train and validation and 15% for test.

4.2 Implementation Details

Training Strategy: We implement our method using PyTorch and CUDA 11.6, and test it on two NVIDIA Titan X GPUs. All input RGB images are 640×640. Based on Yolov5 pre-training, we employ AdamW optimizer[19] to train the network for 500 epochs with a batch size of 8. Moreover, we use clustering of dataset annotation to generate anchors, while using Mosaic as data augmentation strategy[2].

Development Strategy: In order to reduce the time and memory requirement for the calculation of projection matrix and convolution offset matrix, we use the serialized MD5 value of the parameter as the cache key, and store the serialized calculated value in memory and file system. Moreover, we define protobuf for major APIs such as dataset processing, projection and detection, and communication between network architecture via gRPC.

4.3 Results of Unified Distortion-Aware Convolution

Performance Analysis of SteConv: A quantitative comparison of the object detection effect between our proposed SteConv, which is implemented through our proposed unified distortion-aware convolution module, and other convolutions modules is summarized in Table 1. DeformConv achieves better performance than StdConv because the added offset parameter can extract features more efficiently. On the other hand, the pre-defined offset parameters for geometric deformation in SteConv are more efficient than the parameters learned by the network, thus obtaining better performance than DeformConv.

Table 1. Comparison experiments of different kinds of convolution modules. The boldface denotes the best performance in this experiment.

model	mAP	bed	painting	tv	sofa	curtain	table	bedside
StdConv	78.7	88.1	87.5	86.8	79.3	77.6	71.0	74.7
DeformConv	78.9	86.0	87.0	**89.0**	**81.2**	**78.4**	70.4	74.7
SteConv	**79.8**	**90.0**	**88.8**	87.2	80.6	77.5	**72.0**	**76.8**

Ablation Study of SteConv: We experiment with the effect of the position and number of applying SteConv and DeformConv on detection accuracy. As shown in Fig. 4, layer1∼5 represents the position of SteConv in backbone, and layer1∼5+ represents the number of SteConv layers. Through the comparison experiment in (a), we can get that the main factors affecting the performance of SteConv are the size of the convolution layer and the richness of features. With the movement of the position, applying SteConv or DeformConv to the backbone layer closest to the detection head, accuracy reaches the best performance. We conclude that this is because more abstract convolution features can handle distortion better than larger feature layers, thus the final optimal effect is located in the last layer of the backbone. From the comparison experiments in (b), we can see that the accuracy continues to improve as we continue to replace layers with DeformConv. However, replacing more SteConv layers did not improve accuracy. We conclude this is because the SteConv pre-calculated from the geometric deformation formula is theoretically the upper limit of deformation that DeformConv can handle, while more SteConv stacking will bring anti-stereographic distortion.

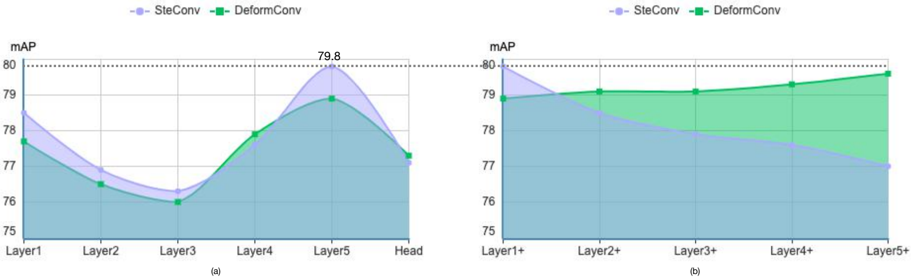

Fig. 4. The figure shows line charts of the position and number of SteConv and DeformConv layers and mAPs. In (a) layer1-5 means the layer position, in (b) layer1-5+ means the number of layers.

4.4 Results of Non-maximum Fusion

Quantitative results: We conduct comparison experiments for the application of the non-maximum fusion algorithm in the post-processing stage of our proposed RepF-Net. From the analysis of our experimental results in Table 2, the non-maximum fusion algorithm fuse the detection boxes from multiple projection images, thus the addition of the non-maximum fusion algorithm further improves the detection accuracy especially when detecting large objects. Based on its effectiveness in handling detection box for both small and large objects, the non-maximum fusion algorithm achieves better performance than the non-maximum suppression algorithm.

Table 2. Quantitative results of RepF-Net, with or without NMF in post-processing stage. The boldface denotes the best performance in this experiment.

model	mAP	tv	painting	bed	curtain	window	bedside	mirror
w/o. NMF	80.5	93.3	**93.1**	80.6	62.4	80.9	**80.4**	73.3
w. NMF	**84.2**	**95.5**	87.6	**87.0**	**79.2**	**81.0**	79.9	**79.2**

Qualitative Results: We show the qualitative results of the comparison between our proposed NMF and NMS in Fig. 5. NMS can achieve good results when detecting objects that can be completely detected in a single projection image, such as (a) in Fig. 5. However, when detecting objects that require the fusion of multiple projection images, it is inevitable that objects will be detected as multiple continuous detection boxes, such as (b) in Fig. 5, or only detect certain components of the object, such as (c) in Fig. 5. In contrast, our proposed NMF outperforms NMS in the above cases.

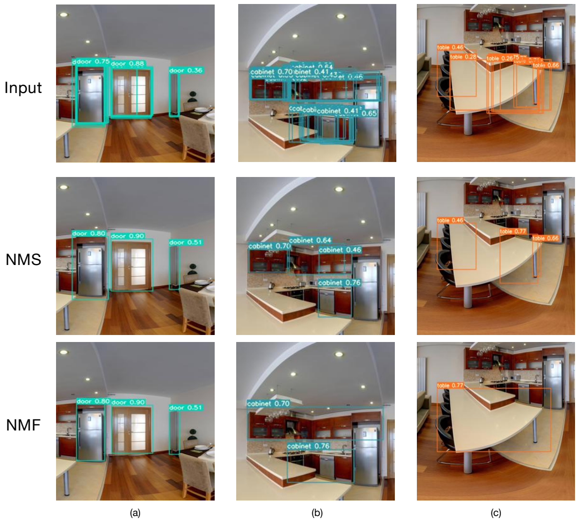

Fig. 5. Qualitative results of the comparison between the non-maximum suppression algorithm and the non-maximum fusion algorithm.

4.5 Comparison with the State-of-the-Art Methods

Comparison Experiment: We compare RepF-Net with baseline methods of object detection in panorama image and the results are given in Table 3. Rep-CNN as a two-stage detector achieves the best performance among the baseline methods due to the projection greatly reduced geometric distortion, and PanoBlitzNet as a one-stage detector also achieves good detection accuracy due to the introduction of the EquiConv. Finally, our RepF-Net combines the advantages of both methods and achieves better performance than the SoTA. Moreover, in RepF-Net+ we apply DeformConv in the backbone of our proposed RepF-Net to replace the convolution layout except for the SteConv layer. And it can be concluded that, on the basis of handling the geometric distortion by SteConv, DeformConv can more efficiently extract features, therefore better generalize the geometric shape of the object and finally achieve the best performance.

Table 3. Performance comparison between baseline methods and RepF-Net. RepF-Net+ represents applying DeformConv on the basis of RepF-Net. The boldface denotes the best performance in this experiment.

model	mAP	tv	painting	bed	curtain	window	bedside
DPM[11]	29.4	31.0	56.0	35.2	29.5	21.8	-
Deng et al.[8]	68.7	70.0	68.0	76.3	69.5	62.6	-
Multi-Project Yolo[28]	69.4	87.5	80.7	17.2	73.4	76.9	78.2
PanoBlitzNet[14]	77.8	93.3	83.9	**95.3**	75.9	70.9	**91.4**
Rep-RCNN[31]	79.6	92.4	**92.2**	70.3	69.5	75.0	83.2
RepF-Net(w. NMF)	84.2	95.5	87.6	87.0	79.2	**81.0**	79.9
RepF-Net+	**86.0**	**95.5**	89.4	90.5	**79.4**	**81.0**	79.9

Qualitative Comparison: We show the qualitative results of the comparison between our proposed RepF-Net and other state-of-the-art methods. As shown in Fig. 6, the one-stage detector PanoBlitzNet can detect large objects and handle incomplete objects well, while has difficult detecting all small objects and has lower accuracy. And the two-stage detector Rep-CNN can improve the detection accuracy of small objects, while it is difficult to detect large objects, and there is a problem of incomplete detection of objects. In contrast, our proposed RepF-Net solves the problem of incomplete detection objects, and achieves better performance in both small and large object detection accuracy.

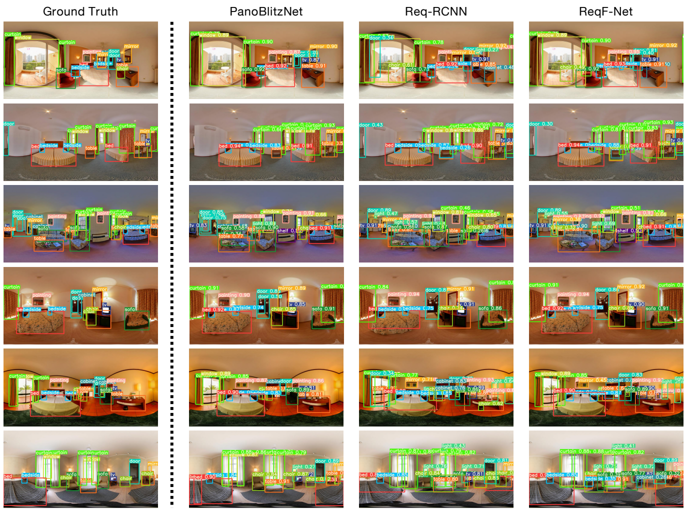

Fig. 6. Qualitative comparison of different object detection methods on panorama image.

5 Applications

On indoor scene understanding tasks, because of the richer contextual information encoded by the larger field of view, using panorama image can achieve better performance than using perspective images. In the task of indoor scene understanding, there are two main steps, object detection and layout recovery[29]. While many methods achieve good performance in layout recovery[35], there is still space for improvement in object detection. Based on our proposed RepF-Net, we present an implementation of indoor scene understanding. And the qualitative results of 3D object detection and scene layout reconstruction on three datasets are shown in Fig. 7.

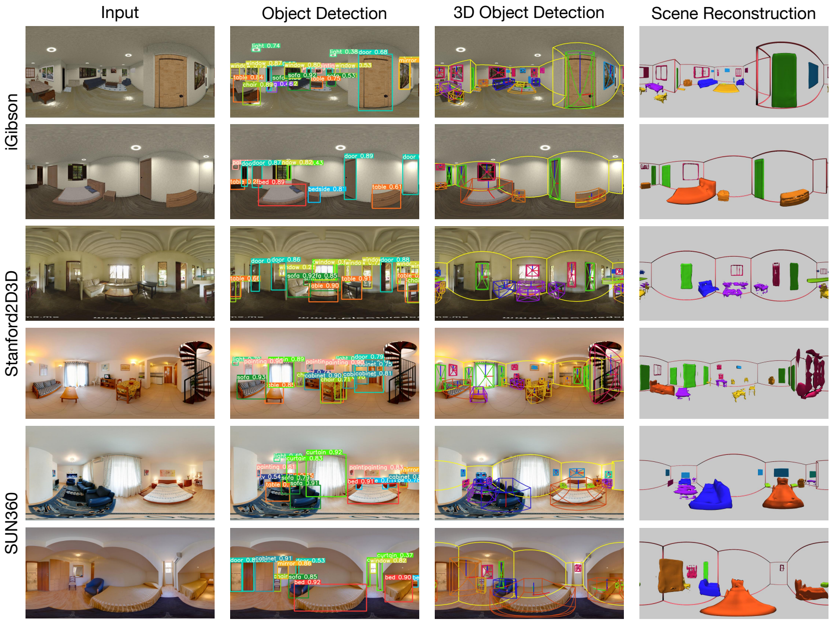

Fig. 7. Qualitative results of 3D object detection and scene layout reconstruction.

6 Conclusion

This paper presents a novel two-stage detection network, RepF-Net for object detection in panorama image, while including a unified distortion-aware convolution module for geometric distortions, and a non-maximum fusion algorithm for post-processing. Experiments validate the effectiveness of each module in our method, and show that our network performs better performance than other state-of-the-art object detectors. In addition, our network model has also been applied to tasks of 3D object detection and scene reconstruction.

References

1. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017)
2. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
3. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms–improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision. pp. 5561–5569 (2017)
4. Cohen, T.S., Geiger, M., Köhler, J., Welling, M.: Spherical cnns. arXiv preprint arXiv:1801.10130 (2018)
5. Coors, B., Condurache, A.P., Geiger, A.: Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In: Proceedings of the European conference on computer vision (ECCV). pp. 518–533 (2018)
6. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 1, pp. 886–893. Ieee (2005)
8. Deng, F., Zhu, X., Ren, J.: Object detection on panoramic images based on deep learning. In: 2017 3rd International Conference on Control, Automation and Robotics (ICCAR). pp. 375–380. IEEE (2017)
9. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6569–6578 (2019)
10. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. the Journal of machine Learning research **9**, 1871–1874 (2008)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence **32**(9), 1627–1645 (2010)
12. Fernandez-Labrador, C., Facil, J.M., Perez-Yus, A., Demonceaux, C., Civera, J., Guerrero, J.J.: Corners for layout: End-to-end layout recovery from 360 images. IEEE Robotics and Automation Letters **5**(2), 1255–1262 (2020)
13. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
14. Guerrero-Viu, J., Fernandez-Labrador, C., Demonceaux, C., Guerrero, J.J.: What's in my room? object recognition on indoor panoramic images. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 567–573. IEEE (2020)
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
16. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision (ECCV). pp. 734–750 (2018)
17. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. In: Proceedings. international conference on image processing. vol. 1, pp. I–I. IEEE (2002)
18. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)

19. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)
20. Meng, M., Xiao, L., Zhou, Y., Li, Z., Zhou, Z.: Distortion-aware room layout estimation from a single fisheye image. In: 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 441–449. IEEE (2021)
21. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
22. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263–7271 (2017)
23. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
25. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing **107**, 104117 (2021)
26. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1746–1754 (2017)
27. Xiao, J., Ehinger, K.A., Oliva, A., Torralba, A.: Recognizing scene viewpoint using panoramic place representation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2695–2702. IEEE (2012)
28. Yang, W., Qian, Y., Kämäräinen, J.K., Cricri, F., Fan, L.: Object detection in equirectangular panorama. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 2190–2195. IEEE (2018)
29. Zhang, C., Cui, Z., Chen, C., Liu, S., Zeng, B., Bao, H., Zhang, Y.: Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12632–12641 (2021)
30. Zhang, Y., Song, S., Tan, P., Xiao, J.: Panocontext: A whole-room 3d context model for panoramic scene understanding. In: European conference on computer vision. pp. 668–686. Springer (2014)
31. Zhao, P., You, A., Zhang, Y., Liu, J., Bian, K., Tong, Y.: Spherical criteria for fast and accurate 360 object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12959–12966 (2020)
32. Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X.: Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems **30**(11), 3212–3232 (2019)
33. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: European Conference on Computer Vision. pp. 519–535. Springer (2020)
34. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9308–9316 (2019)
35. Zou, C., Su, J.W., Peng, C.H., Colburn, A., Shan, Q., Wonka, P., Chu, H.K., Hoiem, D.: Manhattan room layout reconstruction from a single 360 image: A comparative study of state-of-the-art methods. International Journal of Computer Vision **129**(5), 1410–1431 (2021)