Sphere-CenterNet: A Geometry-Aware Center-based Detection on ERP Images

Mengyi Lyu Yafeng Zhao Wanqi Cheng Gang Shi*
Xinjiang University, Urumqi, China
{107552304104, zhaoyafeng, 107552304041}@stu.xju.edu.cn, shigang@xju.edu.cn

Abstract

Object detection in omnidirectional (360°) images is a critical task for robotics and immersive applications. However, its development is severely hampered by the significant geometric distortion and longitudinal circular topology inherent in the Equirectangular Projection (ERP). Directly applying standard 2D detectors often leads to a substantial performance drop as they ignore the underlying spherical geometry. To address these challenges, we propose Sphere-CenterNet, a framework that systematically embeds spherical awareness into a single-stage, anchor-free detector. Our core contribution lies in a novel, geometry-aware network architecture featuring two key innovations. First, we introduce the Latitude-Aware Deformable Convolutional Network (LADCN), which integrates our original Spherical Latitude Attention (LAA) mechanism with Deformable Convolution (DCN). This allows the network to adaptively adjust feature sampling locations and response intensities according to the varying distortion patterns across different latitudinal regions. Second, we construct a Seamless Multi-Scale Fusion Neck, based on a Feature Pyramid Network and built entirely with Circular Convolutions, ensuring the strict preservation of 360° topological continuity during cross-scale feature fusion. This architecture is embedded within an end-to-end spherical detection pipeline that operates entirely in the spherical coordinate system. Extensive experiments on the public 360-Indoor and PANDORA datasets demonstrate the state-of-the-art performance of our method. On 360-Indoor, our model achieves 16.8% mAP, 32.8% AP50, and 15.2% AP75. On the more challenging PANDORA oriented detection task, it also achieves excellent results of 15.3% mAP and 31.5% AP50, fully validating the effectiveness and generalization capability of our proposed architecture.

Keywords: Equirectangular Projection Image, Sphere Object Detection, Spherical Awareness, Spherical CenterNet.

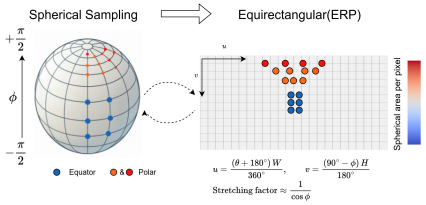


Figure 1. Non-uniform sampling on the sphere. The formulas show the conversion between spherical element coordinates and ERP image coordinates, as well as an approximate distortion factor formula. Interpolation algorithms are required when converting a panoramic spherical image to an ERP image.

1. Introduction

With the rapid development of fields such as virtual reality (VR), augmented reality (AR), autonomous driving, and robotics, the demand for intelligent agents with 360° environmental perception capabilities is increasingly urgent [15, 30]. Omnidirectional Images (ODIs), particularly panoramas presented in the Equirectangular Projection (ERP) format, have become a key data carrier for achieving this goal due to their ability to provide a complete, unobstructed field of view. However, as introduced by Khasanova et al. [13], while the ERP format is convenient for data storage, its unique projection method poses fundamental challenges to high-precision object detection, severely hindering the direct application of existing advanced detection algorithms. Unlike traditional planar perspective images, the pixel coordinates (u, v) in an ERP image do not correspond to a uniform Euclidean grid [27]. Instead, they can be understood as a non-linear sampling of points on a unit sphere, as shown in Figure 1. Each pixel in an ERP image uniquely corresponds to a spherical coordinate (θ, ϕ), where θ represents the polar angle (latitude) and ϕ represents the azimuthal angle (longitude). This "unfolding" mapping from the sphere to a 2D plane leads to a series of fundamental challenges, and methods for object detection based on ERP images and spherical coordinates are known as panoramic image object detection.

In general, panoramic images present two key problems: (1) Severe geometric distortion: The shape, size, and orien-

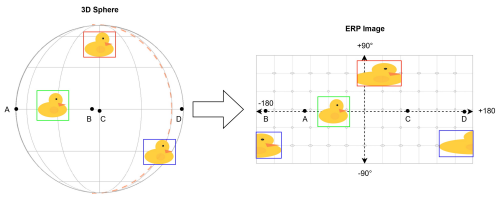


Figure 2. Geometric distortion caused by ERP projection. A regularly shaped object on the sphere is severely stretched near the top of the image (high latitude). A standard 2D bounding box (red) includes a large amount of background to enclose the object, while an ideal bounding box (green) should conform to the spherical geometry. When an object crosses the seam (blue), it is split into two parts in the ERP image.

tation of an object in the image change significantly with its latitudinal position on the sphere. As shown in Figure 2, a regularly shaped object in the real world is nonlinearly stretched into an extremely wide, flat shape when it is near the top and bottom edges of the image (i.e., the north and south poles of the sphere). This geometric distortion not only makes it difficult for deep learning models to extract and understand object features but also complicates the calculation of Intersection over Union (IoU) for objects in panoramic images. (2) Topological discontinuity: The left and right boundaries of an ERP image are seamlessly connected on the sphere. This causes an object that straddles the $\pm 180^\circ$ meridian to be split into two separate parts, appearing at opposite ends of the 2D image. These characteristics severely violate the fundamental assumptions of "translation invariance" [27] and "spatial feature stationarity" [7] that standard Convolutional Neural Networks (CNNs) rely on. This makes it difficult for the model to understand the unity of objects crossing the seam, leading to a significant drop in performance when directly applying advanced 2D detectors (e.g., YOLO, R-CNN), as their receptive fields cannot effectively match the deformed objects.

To address these challenges, existing research has largely followed two main technical routes: Projection-Driven and Distortion-Aware [15]. Projection-driven methods aim to circumvent distortion by re-projecting the ERP image into one or more formats with less distortion (e.g., Cubemap [18], projections based on heterogeneous polyhedra by Cohen et al. [6], or perspective sub-images by Yang et al. [25]), thereby directly utilizing mature 2D detectors designed for perspective images. However, while these methods reduce distortion locally, they introduce new problems: they disrupt the integrity of the image, leading to boundary stitching issues between views, and often result in a significant increase in computational overhead due to the need for multiple model forward passes and image re-projection and restoration. In contrast, distortion-aware methods are a more mainstream approach. They choose to process the ERP image directly, adapting the network architecture itself to handle the distortion. For instance, some works design deformable convolutional kernels or define convolution operations directly on the sphere (e.g., SphereNet, SPHCONV [7]) to allow the model to autonomously learn and adapt to geometric deformations. Cho et al. [4] used a Transformer-based architecture with an attention mechanism to enhance the perception of distortion. Although these methods have made significant progress in feature extraction, they are often "piecemeal" solutions that only address an isolated problem in the entire pipeline or lead to difficulties in model convergence and high computational costs due to their special feature extraction methods.

Therefore, we believe that a systematic, end-to-end solution is necessary. We propose Sphere-CenterNet, a novel framework that systematically injects spherical awareness into every stage of a single-stage, anchor-free detector. Unlike previous work that focused on improving a single module, our work deeply optimizes both the feature extraction and feature fusion stages:

First, in the feature extraction stage, we propose a novel hybrid spherical feature extraction architecture. This architecture follows the design philosophy of using strict geometric priors at the low level and data-driven adaptation at the high level. In the shallow layers of the network, we use strict spherical convolution (SphereConv, proposed by Coors et al. [7]) to establish geometrically correct initial features. In the deeper layers, we use our original Latitude-Aware Deformable Convolutional module (LADCN). By integrating the Spherical Latitude Attention (LAA) mechanism with Deformable Convolution (DCN [23]), this module can adaptively adjust sampling locations and feature responses according to the varying distortion patterns at different latitudes, thus flexibly modeling complex semantic deformations. Second, in the feature fusion stage, we construct a seamless multi-scale feature fusion neck. This neck is based on the efficient Bi-directional Feature Pyramid Network (EBiFPN [3]), uses SphericalWConv convolution for weighted computation of the horizontal spatial dimension, and is built entirely with Circular Convolutions, strictly ensuring the preservation of 360° topological continuity during cross-scale fusion. We believe that this comprehensive geometry-aware design, from feature extraction to feature fusion, is the optimal path that balances theoretical completeness and model flexibility. The main contributions of this paper can be summarized as follows:

- **A systematic spherical detection framework:** We propose Sphere-CenterNet, an end-to-end panoramic object detection framework. Unlike the piecemeal improvements in existing work, this framework natively supports spherical geometric coordinates at multiple

levels, including supervision, feature extraction, feature fusion, and post-processing, forming a complete and self-consistent solution.

- **A novel hybrid feature extractor:** We designed a hybrid spherical feature extractor following the philosophy of using strict geometric priors at the low level, data-driven adaptation at the high level. It efficiently handles the complex distortions of ERP images in a hierarchical manner by using spherical convolution (SphereConv) in the shallow layers, integrating our original Latitude-Aware Deformable Convolution (LADCN) in the deep layers, and designing a more lightweight and adaptable EBiFPN Lite module for the neck based on the BiFPN feature fusion idea.

2. Related Work

This section will review the three areas most relevant to our work: general 2D object detection, object detection in omnidirectional images, and technologies related to spherical bounding boxes and IoU calculation.

2.1. General 2D Object Detection

In recent years, 2D object detection based on deep convolutional neural networks (CNNs) has made significant progress, achieving extremely high accuracy on large public datasets such as COCO and ImageNet. Existing object detection methods can be broadly divided into two-stage and single-stage detectors. Two-stage methods, represented by Faster R-CNN [20], first generate candidate regions through a Region Proposal Network (RPN) and then perform classification and regression, typically achieving higher accuracy but at a slower speed. Another CenterNet-based [33] two-stage detector, CenterNet2 [32], essentially uses CenterNet as an RPN to propose better candidate regions. Although it improves the accuracy of candidate regions, it still leads to certain errors in panoramic object detection tasks. In contrast, single-stage methods, represented by YOLO [19] and SSD [16], perform dense prediction directly on feature maps, eliminating the region proposal step and achieving faster detection speeds. However, most of these mainstream methods rely on a set of predefined anchor boxes with fixed sizes and aspect ratios for prediction. This prior design faces fundamental challenges when dealing with panoramic images with severe geometric deformations. The apparent shape of objects in an ERP image changes dramatically with latitude; a fixed object may appear square near the equator but be stretched into an extremely wide, flat shape at high latitudes. Therefore, designing a set of anchor boxes that can effectively match all possible deformations becomes almost impossible, leading to a large number of low-quality proposals, often with localization deviations and boundary tearing issues, which seriously affect detection performance. To overcome the reliance on anchor boxes, anchor-free detectors have emerged, with FCOS [22] and CenterNet [33] being outstanding representatives. These methods no longer rely on fixed anchor boxes but instead treat object detection as a problem of locating keypoints (such as center points or corner points) and regressing related attributes (such as size). This paradigm is naturally more suitable for panoramic images because it focuses on the relatively more stable geometric features of objects (like the center point) rather than their easily variable apparent shapes on a 2D plane. In particular, CenterNet simplifies the object detection task to keypoint estimation of object centers and directly predicts the object's size through regression. Its simple, efficient, and powerful paradigm makes it an ideal baseline model that is more robust geometrically. Furthermore, its design of center points and sizes naturally aligns with the annotation format of panoramic object detection (see Section 2.3), making it a better choice for this task. In summary, our Sphere-CenterNet is built upon the powerful anchor-free idea of CenterNet, trained and evaluated entirely based on spherical coordinates, and has been systematically adapted for spherical geometry in every aspect from supervision to network architecture to fully unleash its potential in omnidirectional image detection tasks, truly achieving an end-to-end panoramic image detection pipeline.

2.2. Object Detection in 360° Images

The core challenge of object detection in omnidirectional images lies in handling the three structural problems caused by Equirectangular Projection (ERP): geometric distortion, non-uniform sampling, and boundary continuity [15]. These characteristics severely violate the fundamental assumptions, such as "translation invariance," that CNNs designed for standard perspective images rely on. To address these challenges, researchers have conducted extensive explorations. Early work focused on modifying the core operators of the network, which later evolved into two mainstream technical routes: Distortion-Aware and Projection-Driven.

Projection-based Methods. To directly circumvent the severe distortion of the ERP format, one class of methods first re-projects the ERP image into one or more formats with less distortion. For example, as discussed in Ref. [1], Cubemap (CMP/CP) projects the spherical image onto the six 90° orthogonal faces of a cube; Tangent Projection (TP) projects the spherical content onto multiple local tangent planes, generating approximately distortion-free image patches; Stereographic Projection is another common projection method, for example, the work of Yang et al. (2018) [25] converts one ERP image into four stereographic sub-images. The workflow of these methods typically involves running mature 2D detectors (like YOLO)

independently on each projected view and then merging the detection results from each view back into the original ERP coordinate system through a post-processing step. A representative work is Reprojection R-CNN [28]. However, while this strategy reduces distortion locally, it introduces a series of new problems: 1) Object splitting: Objects that cross the boundaries of different views are segmented, requiring complex post-processing heuristics for stitching and merging results. 2) High computational cost: It usually requires performing model forward propagation on multiple views separately, leading to a significant increase in computational overhead. 3) Loss of global context: Processing each view independently destroys the integrity of the scene, preventing the model from utilizing the rich global context information provided by the panoramic image.

Distortion-Aware Network Design. Unlike reprojection methods, this approach aims to natively handle the geometric properties of ERP images by directly modifying the architecture of the deep learning model. The research focuses on how to adapt the convolution operation to the non-Euclidean manifold structure of ERP. 1) Data-driven adaptive convolution: Some works use Deformable Convolution (DCN) [23]. DCN allows the sampling points of the convolution kernel to be dynamically offset on the feature map, enabling the network to autonomously learn the local geometric deformation of objects from data. For the non-linear stretching of objects in ERP images that varies with latitude, DCN provides a flexible, data-driven solution, such as in PV-YOLO [12]. 2) Spherical convolution based on strict geometric priors: Other works pursue stricter geometric correctness. Yu et al. [26] used a network for feature extraction from panoramic images. The most authoritative is SphereNet [7], which proposed a true spherical convolution and pooling method. Since the correspondence between spherical coordinates and pixel coordinates is only related to the image width and height, this method uses `grid_sample` to make the convolution sample fixed spherical grid pixels, ensuring that the receptive field of the convolution is uniform and distortion-free on the sphere, achieving rotation equivariance. These methods have made significant progress in feature extraction, enabling the network to "understand" distorted objects.

2.3. Evolution of Spherical Bounding Box and IoU

As research deepened, it has been widely recognized that, in addition to modifying the network structure, a more fundamental challenge lies in how to geometrically represent and evaluate objects in panoramic images correctly. As shown in Figure 2, directly using axis-aligned 2D rectangular boxes to enclose objects, especially in high-latitude regions, inevitably introduces a large amount of irrelevant background area or fails to tightly fit the object's contour, thereby severely affecting the model's localization accuracy. This has driven the development of spherical bounding boxes and their Intersection over Union (IoU) calculation methods.

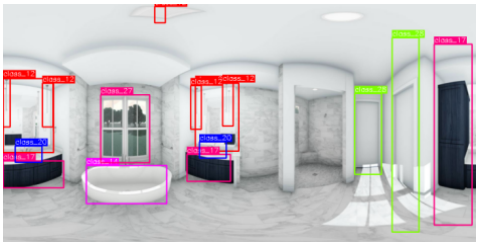


Figure 3. Effect of annotating in spherical space with planar boxes and then converting to an ERP image.

To more accurately enclose objects, the Bounding Field of View (BFoV) and its rotated form (RBFOV), as introduced by Zhao et al. [29], are defined directly on the sphere. The annotation format is $(\theta, \phi, \alpha, \beta, (\gamma))$, where (θ, ϕ) represent the longitude (azimuthal angle) and latitude (polar angle) of the object's center, determining its position on the sphere; (α, β) represent the horizontal and vertical fields of view, defining the object's extent in angular space; and γ represents the object's rotation angle. This representation of object shape is not adversely affected by the object's position on the sphere. However, this shift also brought the core challenge of how to accurately calculate the IoU between two spherical boxes. The technological evolution in this area clearly reflects the trade-offs and breakthroughs among accuracy, efficiency, and differentiability, as shown in Table 1.

Early Approximate Calculations (Sph-IoU [31], FoV-IoU [2]): As early approximate solutions, their core idea was to circumvent complex spherical geometry calculations through simplified mappings. Although somewhat better than the original 2D planar IoU, their approximation strategies were too coarse, leading to significant deviations from the true IoU and insufficient accuracy when dealing with complex distortions in high-latitude regions.

Exact Analytical Solution (Unbiased IoU [8]): To establish an unbiased evaluation standard, Dai et al. (2022) proposed Unbiased-IoU, the first method capable of calculating the exact analytical solution for IoU using spherical geometry formulas. Due to its mathematical completeness, it is recognized as the "gold standard" for spherical IoU calculation and is the most reliable metric for evaluating model performance. However, its complex analytical process leads to high computational costs, and most critically, it is non-differentiable. This fatal flaw prevents it from being used as a loss function in the training of modern detectors, causing an inconsistency between the training objective (e.g., L1 Loss) and the final evaluation metric (IoU).

Practical and Differentiable Breakthrough (Sph2Pob

[17]): To solve the aforementioned disconnect between training and evaluation, Liu et al. (2023) proposed Sph2Pob. Its core contribution is to equivalently transform the complex spherical IoU problem into the well-solved planar oriented box IoU problem. This clever transformation not only yields a highly accurate approximate IoU but, more importantly, inherits the excellent properties of being efficient and differentiable from planar IoU calculators. The emergence of Sph2Pob was a key breakthrough, as it made it possible to apply advanced IoU-based loss functions (like CIoU-Loss) to the end-to-end training of spherical object detectors.

Method	R_all ↑	R_low ↑	R_high ↑	T_cpu ↓	T_cuda ↓
Sph	0.7819	0.9922	0.4274	0.0364	0.0013
Fov	0.9600	0.9974	0.8860	0.0372	0.0014
Sph2Pob	0.9989	0.9990	0.9988	2.2275	0.0038
Unbiased	1.0000	1.0000	1.0000	46.4417	-

Table 1. Comparison of panoramic IoU calculation efficiency. This experiment is from the Sph2Pob paper and was tested in our experimental environment.

It is based on this technological evolution that our Sphere-CenterNet framework stands on the shoulders of giants. By using a differentiable IoU loss based on Sph2Pob during training and Unbiased IoU as the gold standard for evaluation, we achieve a unification of the training objective and the evaluation metric, thereby effectively improving the model's localization accuracy. In comparison, our Sphere-CenterNet provides a more comprehensive and systematic solution. We not only designed a hybrid feature extraction strategy of "low-level SphereConv + deep-level LADCN" but also innovated in multiple dimensions, such as seamless multi-scale feature fusion. Through this end-to-end, highly integrated, and self-consistent design, we aim to build a truly spherical-native detection framework.

2.4. Datasets for Panoramic Object Detection

High-quality, large-scale datasets are the cornerstone of advancing deep learning methods. Compared to the field of general 2D object detection, datasets for panoramic object detection are relatively scarce due to difficulties in collection and annotation, and the lack of a fully unified annotation format, which has somewhat limited the development of the field. Early research often relied on semi-virtual datasets synthesized from existing 2D datasets. For example, Coors et al. [7] and Su et al. [21] rendered object instances onto panoramic backgrounds to create synthetic panoramic equivalent datasets. Although these datasets played a role in the initial validation of algorithms, they could not fully simulate the complex distortions, lighting, and scene layouts of the real world. To bridge this gap, the research community has recently released several panoramic image detection datasets collected from real-world scenes and manually annotated. Among them, two datasets have become recognized as core benchmarks in the field due to their representativeness and challenges:

360-Indoor Dataset: Released by Chou et al. (2020) [5], this is the first large-scale, real-world panoramic object detection dataset specifically for indoor scenes. It contains over 3,300 high-resolution indoor panoramic images and nearly 90,000 object instance annotations, covering 37 categories. The dataset is annotated using the axis-aligned Bounding Field of View (BFoV) format, i.e., $B_S(\theta, \phi, \alpha, \beta)$. As a classic dataset in the field, 360-Indoor provides a standardized platform for validating a model's basic detection capabilities in handling real-scene distortions and complex layouts.

PANDORA Dataset: Released by Hsu et al. (2022) [24], the PANDORA dataset is similar to the 360-Indoor dataset, but its annotations use the Rotated Bounding Field of View (RBFOV) format, i.e., $B_S(\theta, \phi, \alpha, \beta, \gamma)$, adding a rotation angle parameter.

3. Methodology

3.1. Overview

To address the challenges of geometric distortion and topological discontinuity brought by Equirectangular Projection (ERP) in omnidirectional images, we propose Sphere-CenterNet, a novel framework that systematically embeds spherical awareness into every stage of a single-stage, anchor-free detector. Figure 5 illustrates the overall framework. This section will detail its overall architecture, the core hybrid spherical feature extractor, the geometry-aware supervision paradigm, and the end-to-end processing pipeline.

3.1.1 Panoramic Image Object Detection Dataflow

In a panoramic image, the pixel position and the spherical position of an object have a one-to-one correspondence, which is derived from the non-uniform sampling of the sphere. The conversion formula we use from pixel coordinates to spherical coordinates is as follows, where an offset of half a pixel is added to counteract the half-pixel offset during the conversion process. From ERP $(u + \frac{1}{2}, v + \frac{1}{2})$ to spherical angles:

$$\begin{aligned} \phi &= \frac{360^\circ}{W}\left(u + \frac{1}{2}\right) - 180^\circ, \\ \theta &= 90^\circ - \frac{180^\circ}{H}\left(v + \frac{1}{2}\right) \end{aligned} \tag{1}$$

Our method takes an ERP panoramic image $I(H \times W)$ as input. Each pixel corresponds to a point on the sphere, with longitude $\phi \in (-180^\circ, 180^\circ]$ and latitude $\theta \in [-90^\circ, 90^\circ]$.

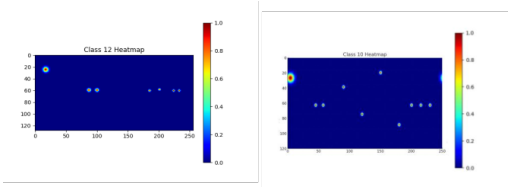


Figure 4. Single-class heatmap of center points and a heatmap showing a cross-seam center point.

The target is represented by a spherical center (θ, ϕ) and angular dimensions (w, h); if the dataset includes rotation, an in-plane angle γ is also included. Detection is performed on a down-sampled feature map, and the output directly falls into the spherical parameter space, avoiding the mapping of distortion to Euclidean pixel dimensions.

Furthermore, the above coordinates are in degrees. Since the center point and angular dimensions often differ greatly in magnitude, direct training can be difficult for the model. Therefore, we apply a special scaling to (w, h) as shown in Equation 2, where r is the down-sampling rate of CenterNet, typically 4. This essentially scales w and h by an average down-sampled width and height.

$$wh_{k,0} = \frac{w_{deg}}{360°/W \cdot r}, \quad wh_{k,1} = \frac{h_{deg}}{180°/H \cdot r} \tag{2}$$

3.2. Sphere-CenterNet Overall Framework

Our method is built upon the paradigm of the advanced anchor-free detector, CenterNet. We chose CenterNet as the foundation because its core idea—treating object detection as a problem of locating keypoints (like center points) and regressing related attributes—is naturally more adaptable to the severe geometric deformations of objects in ERP images than methods relying on fixed anchor boxes. The anchor-free paradigm focuses on the relatively more stable geometric features of objects (like the center point) rather than their easily variable apparent shapes on the 2D plane. Our calculations for the center point and offset are based on pixel coordinates, but the ground truth fed into the model is entirely based on spherical angular coordinates. When drawing the center point heatmap, we naturally plot the center points of objects that cross the seam on both the left and right sides of the seam, as shown in Figure 4. A cross-seam center point is rare and only occurs when the calculated Gaussian radius R is very large.

The core idea of Sphere-CenterNet is to migrate the entire detection pipeline from the traditional 2D pixel coordinate system to the natural spherical coordinate system for direct object localization and regression. As shown in Figure 5, the input ERP image first passes through a feature extractor (composed of a Backbone and a Neck) meticulously designed for spherical geometry to generate a high-resolution feature map. Subsequently, three parallel detection heads perform dense prediction on this feature map, outputting the spherical geometric attributes of the objects:

- **Center Heatmap:** A heatmap of dimension $C \times H' \times W'$, where C is the number of object classes. Each peak in the heatmap corresponds to the spherical coordinate center point (θ, ϕ) of a potential object. An anisotropic Gaussian is placed at (x_n, y_n) on the heatmap for class c, as shown in Equation 3. We also attempted to perform special calculations for σ_x, σ_y based on panoramic distortion pixels (see Section 3.3).

$$Y_c(y, x) = \exp\left(-\frac{(x - x_n)^2}{2\sigma_x^2} - \frac{(y - y_n)^2}{2\sigma_y^2}\right) \tag{3}$$

- **Size Regression:** A feature map of dimension $2 \times H' \times W'$, used to regress the angular dimensions (w_{deg}, h_{deg}) of the bounding box on the sphere corresponding to each object's center point.

- **Local Offset Regression:** A feature map of dimension $2 \times H' \times W'$, used to compensate for the position quantization error caused by network down-sampling, thereby predicting the precise sub-grid offset $(\delta\theta, \delta\phi)$ of the center point.

3.3. Hybrid Spherical Feature Extractor

To equip the network with strong spherical geometry perception capabilities while maintaining flexibility in modeling complex semantic deformations, we did not adopt a single technology but proposed a hierarchical, progressive hybrid feature extraction strategy. This strategy follows the design philosophy of using strict geometric priors at the low level and data-driven adaptation at the high level, and uses different geometric processing modules at different depths of the network.

3.3.1 Spherical Convolution Frontend

In the initial layer (Stem) of the network, feature extraction is most sensitive to geometric distortion. Standard 2D convolution/pooling here would extract severely distorted features due to their fixed rectangular receptive fields, introducing biases that are difficult to eliminate. To this end, we adopted strict Spherical Convolution (SphereConv) [7] to replace the standard 2D convolution and pooling layers. SphereConv defines the convolution kernel on the unit sphere S^2 and evaluates the kernel response at any spherical position through the action of the rotation group $SO(3)$; its continuous form is shown in Equation 4:

$$y(x) = \int_{S^2} f_c(\xi)\psi^{(c)}(R_x^{-1}\xi)d\Omega(\xi), \quad x \in S^2 \tag{4}$$

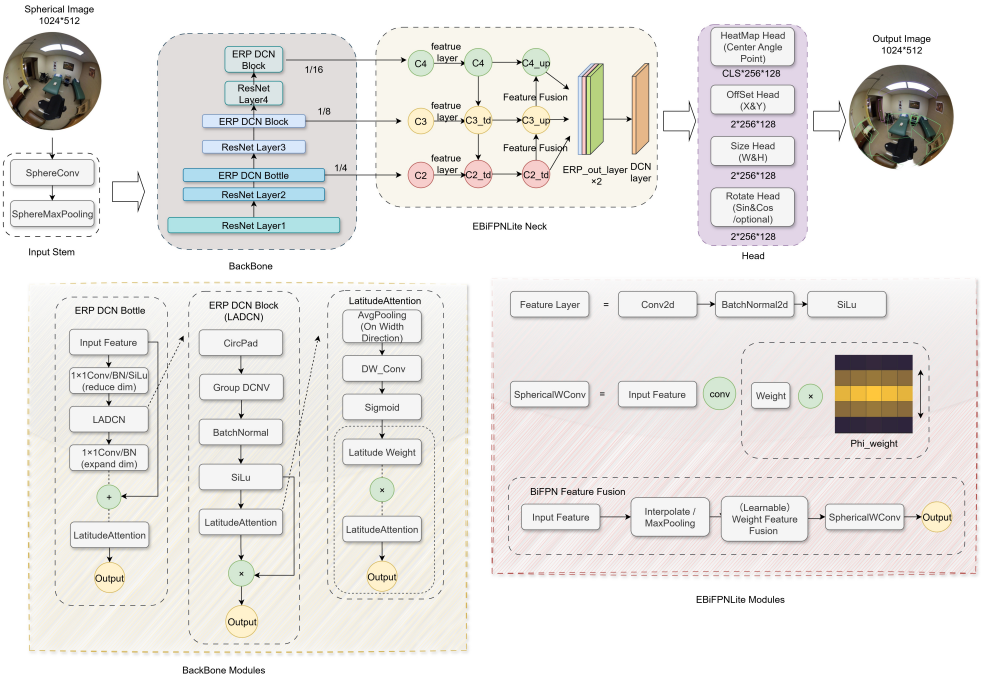


Figure 5. Sphere-CenterNet overall framework. The complete pipeline from the input image passing through the hybrid spherical feature extractor (Backbone + Neck) to the final output of the three prediction heads (heatmap, size, offset). The lower part shows the composition of the LADCN and EBiFPN Lite sub-modules.

where R_x rotates the north pole to position x, ensuring that the receptive field is consistent in terms of geodesic radius and is equivariant/robust to 3D rotations. The SphereConv algorithm is implemented by pre-calculating a non-uniform sampling grid based on spherical geodesic distance, then using the `grid_sample` operation to resample the input features to obtain a spherical grid, and finally performing standard convolution. This process ensures that the receptive field of the convolution is uniform and distortion-free on the sphere, providing a high-quality, geometrically correct initial feature basis for the subsequent network. In our implementation, the first convolutional layer and the max-pooling layer of the backbone network were replaced with 'SphereConv2d' and 'SphereMaxPool2d', respectively.

3.3.2 Latitude-Aware Deformable Convolution

As features are abstracted through multiple network layers, their deformation patterns become more complex and semantic, and relying solely on fixed geometric priors is no longer sufficient to capture all variations. Therefore, in the deeper layers of the backbone network, we designed a novel Latitude-Aware Deformable Convolution module (LADCN), which organically combines spatial adaptability and positional adaptability.

Spatial Adaptability - Deformable Convolution (DCN): We use Deformable Convolution (DCN) as the foundation. DCN learns an additional 2D offset for each convolution sampling point, allowing the network's receptive field to dynamically conform to the actual shape and pose of the object. This endows the model with the ability to handle local geometric deformations of objects. The DCN sampling formula is as follows, where p_0 is the output value at each point, p_k is the predefined offset, and Δp_k^{dcn} is the offset calculated by the DCN convolution.

$$y(p_0) = \sum_k w_k \cdot x(p_0 + p_k + \Delta p_k^{dcn}) \tag{5}$$

Positional Adaptability - Latitude-Aware Attention (LAA): This is one of our core innovations. Standard attention mechanisms (like SE-Net) are spatially invariant and cannot adapt to the distortion patterns that vary with latitude in ERP images. To this end, we designed the Latitude-Aware Attention (LAA) module. As shown in the Backbone Modules in Figure 5, this module first extracts a summary feature vector for each row of the feature map (corresponding to different latitude bands) through average pooling along the width (longitude) dimension. Subsequently, this vector passes through a bottleneck structure similar to SE-Net (composed of two 'Conv1d' layers) for inter-channel information interaction, and finally generates a set of latitudinally variant channel attention weights through a Sigmoid function. The LADCN module combines these two: the input feature first passes through a DCN layer for spa-

tial sampling adaptation, and then its output is element-wise multiplied by the latitude weights generated by the LAA module. This design allows our model to not only adapt to the local shape of the object (achieved by DCN) but also to dynamically adjust its feature response intensity based on its global latitudinal position (achieved by LAA), thus more intelligently and hierarchically handling the non-uniform complex distortions in panoramic images. The computation process of the LADCN_Block can be summarized as follows:

$$F_{out} = \text{DCN}(F_{in}) \otimes \text{LAA}(\text{act}(\text{BN}(\text{DCN}(F_{in})))) \quad (6)$$

where $\otimes$ denotes element-wise multiplication. This design allows our model to adapt not only to the local shape of objects (DCN) but also to adjust its feature response intensity according to its global latitudinal position (LAA), thus more intelligently handling the non-uniform distortion of panoramic images. The calculation process further incorporates latitude gating (position dependence) and channel attention (magnitude dependence) on top of the DCN formula in Equation 5:

$$\Delta p_k = \Delta p_k^{dcn} + g(\phi) \cdot b_k, \quad y(p_0) = a(\phi) \odot y(p_0) \quad (7)$$

where $g(\phi) \subseteq$ is a gate that varies with latitude, and $a(\phi) \in \mathbb{R}^C$ is the channel weight from LAA, which is then broadcast along the width direction to the entire feature row:

$$s(\phi) = \text{GAP}_\theta(x(\phi, \cdot)) \xrightarrow{W_1, \text{ReLU}, W_2, \sigma} a(\phi) \in (0,1)^C \quad (8)$$

To reduce the model's computational load and maintain the adaptive capability of the feature extraction layers, we first reduce the dimensionality of the output feature map from the Stem stage before feeding it into the LADCN_Block module, and then restore it by increasing the dimensionality. This simplifies the model's process of calculating adaptive dimensional weights, ultimately forming the LADCN_Bottleneck module, which is placed at the first layer of the feature extraction stage.

3.3.3 Seamless Multi-Scale Fusion Neck

In object detection tasks, fusing multi-scale features from different levels of the backbone network is crucial for accurately identifying objects of different sizes. However, standard multi-scale fusion architectures (like FPN) typically use standard 2D convolutions, which can sever the features of objects that cross the $\pm 180°$ meridian boundary when processing ERP images, destroying the circular topological continuity of the scene. To solve this problem, we designed and implemented a seamless multi-scale feature fusion neck. The core idea of this module is to strictly maintain 360° topological continuity throughout the feature fusion process. Our implementation is based on the following two key designs:

Efficient Bi-directional Fusion Architecture: We adopted the efficient Bi-directional Feature Pyramid Network (BiFPN) as the basic architecture. BiFPN achieves fast and effective feature fusion through top-down and bottom-up bi-directional paths and learnable weights. **Global Circular Convolution Modification:** This is our key modification. We replaced all convolution operations in the BiFPN neck, including depthwise separable convolutions and the final upsampling output layer, with our custom Circular Convolution (CircConv). The principle of CircConv is to apply asymmetric padding to the input feature map before the convolution operation. Specifically, it only performs circular padding on the width (longitude) dimension, making the left and right boundaries of the image seamlessly connected at the feature level. In the height (latitude) dimension, it still uses conventional zero padding. By globally and consistently using circular convolution throughout the feature fusion neck, we ensure that the 360° integrity of the scene is always maintained during the flow of information across scales. This enables the model to correctly understand and process objects that are split by the ERP boundary, thereby generating a topologically continuous and semantically consistent multi-scale feature map, providing high-quality input for the subsequent detection heads. The above part addresses the seamless aspect of the neck, but the input features still have dimension-related characteristics. Therefore, we also implemented SphericalWConv convolution to add a weight to different dimensions, allowing the feature fusion layer to also perceive the importance of dimensional features, enhancing robustness to ERP distortion and seam connectivity. The core of SphericalWConv convolution is to introduce a static, symmetric, and separable spatial weight mask Φ to the convolution kernel, with the same shape as the kernel size, for Hadamard element-wise modulation of the learnable weights, as proposed by Horadam [11]: $W_{eff} = W \odot \Phi$. Specifically, given an odd kernel size K and a decay parameter vector $d = (d_1, \ldots, d_M)$ (where $M = \frac{K-1}{2}$), we first construct a one-dimensional symmetric weight:

$$\alpha = [d_1, \ldots, d_M, 1, d_M, \ldots, d_1]^\top \in \mathbb{R}^K \quad (9)$$

Then let the two-dimensional mask be the outer product:

$$\Phi = \alpha\alpha^\top \in \mathbb{R}^{K \times K} \quad (10)$$

Thus, Φ applies decay to the rows and columns of the kernel separately, which is equivalent to applying a fixed proportion to each spatial position of the convolution kernel, highlighting the center and suppressing edge responses. When $K = 3$ and $d = (\alpha)$, this is the commonly used 2D separable extension of $[\alpha, 1, \alpha][\alpha, 1, \alpha]$. In implementation, we first calculate $W_{eff} = W \odot \Phi$ and then call a standard 'conv2d' to complete the forward pass; this step has almost

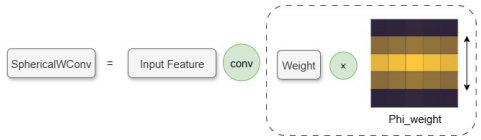


Figure 6. Model diagram of SphericalWConv for dimensional weight convolution.

no additional overhead, just one element-wise multiplication.

3.4. Multi-Task Loss Function

Our total loss function L_{total} is a multi-task loss, composed of weighted loss terms corresponding to the three core prediction heads and an optional rotation loss term:

$$L_{total} = \lambda_{hm}L_{hm}+\lambda_{off}L_{off}+\lambda_{iou}L_{iou}+\lambda_{rot}L_{rot} \quad (11)$$

where the individual loss terms are defined as follows:

Center Point Heatmap Loss (L_{hm}): we use Focal Loss [14], which can effectively handle the extreme class imbalance caused by a large number of background (negative samples) and a small number of object centers (positive samples) in dense prediction. **Local Offset Loss (L_{off}):** For the sub-grid refinement of the center points, we use the standard L1 loss. This loss directly regresses the quantization error caused by the feature map downsampling, making the center point localization more precise. **Size IoU Loss (L_{iou}):** This is key to optimizing object size and localization. Unlike CenterNet's original L1 loss for size, we adopted a differentiable spherical IoU loss (e.g., Sph2PobIoULoss) implemented based on the Sph2Pob idea. The IoU loss optimizes the size and position of the object as a whole, and its optimization objective is highly consistent with the final evaluation metric, thus leading to more accurate localization results. **Rotation Angle Loss (L_{rot}):** To enable our framework to handle oriented object detection tasks, such as on the PANDORA dataset, we introduced an optional rotation loss term L_{rot}. This loss term uses Smooth L1 Loss, as in Fast R-CNN by Girshick et al. [20], specifically for supervising the model's regression of the object's rotation angle γ.

4. Experiments

This section aims to validate the effectiveness of our proposed Sphere-CenterNet framework through a series of rigorous experiments. We first introduce the datasets, evaluation metrics, and implementation details used in our experiments. Then, through detailed ablation studies, we quantitatively analyze the performance gains brought by each of our proposed core components (including the hybrid spherical feature extractor, geometry-aware heatmap supervision, etc.). Finally, we compare the performance of our best model with existing state-of-the-art methods in the field.

4.1. Datasets and Evaluation Metrics

4.1.1 Datasets

To comprehensively evaluate our model's performance, our experiments were mainly conducted on two recognized panoramic object detection benchmark datasets collected from real-world scenes, primarily using the 360-Indoor dataset as the benchmark:

360-Indoor: This is the first large-scale, real-world panoramic object detection dataset for indoor scenes, released by Chou et al. (2020) [5]. It contains over 3,300 high-resolution indoor panoramic images and nearly 90,000 object instance annotations, covering 37 categories. The dataset is annotated using the axis-aligned Bounding Field of View (BFoV) format, i.e., $B_S(\theta, \phi, \alpha, \beta)$. As a classic benchmark in the field, 360-Indoor provides a standardized platform for validating a model's basic detection capabilities in handling real-scene distortions and complex layouts.

PANDORA: Released by Hsu et al. (2022) [24], the PANDORA dataset raises the detection challenge to a new level. It is also a dataset containing 3,000 real-world panoramic images and over 90,000 annotations, but its core feature is the provision of oriented bounding box annotations. The dataset is annotated using the Rotated Bounding Field of View (RBFOV) format, i.e., $B_S(\theta, \phi, \alpha, \beta, \gamma)$, adding a rotation angle parameter. Experiments on this dataset can more comprehensively test the generalization ability and robustness of our framework, especially its effectiveness in more complex oriented object detection tasks.

4.1.2 Evaluation Metrics

We follow the standard COCO evaluation protocol and report the following core metrics: mAP (mean Average Precision over IoU thresholds from 0.5 to 0.95, denoted as AP in the paper), AP50 (IoU threshold of 0.5), and AP75 (IoU threshold of 0.75), as well as Recall (evaluating the proportion of predicted positive samples). A crucial difference is that, to ensure the geometric correctness, fairness, and comparability of the evaluation, all Intersection over Union (IoU) calculations for final performance evaluation use the high-precision Unbiased Spherical IoU (Unbiased IoU) proposed by Dai et al. (2022) [8]. Unbiased IoU is the first method that can calculate the exact analytical solution for IoU using spherical geometry formulas and is recognized as the "gold standard" in the field. Using this metric for evaluation ensures that our experimental results are not affected by any approximate calculation errors, thus most truly and fairly reflecting the model's performance in spherical geometric localization accuracy.

4.2. Implementation Details

Our model is based on ResNet-101 [10] as the backbone network, with weights pre-trained on ImageNet [9]. All models were trained end-to-end for 120 epochs using the AdamW optimizer. We used a Cosine Annealing learning rate strategy with a 5-epoch linear warmup, with an initial learning rate of 1.25e-4. The input image resolution was uniformly adjusted to 512x1024. All experiments were conducted on NVIDIA A40 GPUs. Unless otherwise specified, our data augmentation pipeline included random circular translation, color jitter, and horizontal flipping. The λ weights for our loss function were: $\lambda_{hm} : \lambda_{off} : \lambda_{iou} : \lambda_{rot} = 1 : 0.8 : 1 : 0.2$. The final model took approximately 8 hours to train.

4.3. Comparison with State-of-the-Art

We compared our final Sphere-CenterNet model with other state-of-the-art methods published on the 360-Indoor and PANDORA datasets. As shown in the tables, our method demonstrated competitive performance on all core metrics. The final model size reached 389.9 MB, with a parameter count of 102.02M. The estimated compute at 1024×512 is approximately 135 GFLOPs (forward-only), obtained by area scaling from ResNet-101 and adding DCN/BiFPN/heads overhead. The FPS is approximately 38–42.

Performance on the 360-Indoor Dataset: As shown in Table 2, our Sphere-CenterNet demonstrated excellent performance on the 360-Indoor validation set. We selected recent available panoramic image object detection frameworks for comparison. Our model ultimately achieved 16.8% mAP, 32.8% AP50, and 15.2% AP75, significantly surpassing existing methods on all core metrics.

Methods	Backbone	AP	AP50	AP75
CenterNet(`base_line`)	ResNet-101	8.6	20.5	5.8
CenterNet	Hourglass	13.4	29.5	10.2
Multi-Kernel	ResNet-101	4.7	11.1	2.8
Sphere-SSD	ResNet-101	2.9	7.8	1.4
Reprojection R-CNN	ResNet-101	5.0	15.3	1.9
Unbiased IOU	ResNet-101	10.0	24.8	6.0
Sph2Pob IOU	ResNet-101	11.6	26.1	8.4
Ours	**ResNet-101**	**16.8**	**32.8**	**15.2**

Table 2. Performance comparison on the 360-Indoor validation set.

It is worth noting that since we sample spherical coordinates as the input and output data flow of the model, our baseline did not reach the baseline accuracy of the latest IoU-based detectors. Specifically, compared to the powerful baseline model also trained with the advanced Sph2Pob IoU loss, our method achieved a significant improvement of +5.2% in mAP (16.8% vs. 11.6%) and nearly doubled the performance on the more stringent AP75 metric (15.2% vs. 8.4%). This huge advantage under strict measurement standards strongly proves that our model's localization accuracy in spherical space far exceeds previous methods.

Generalization Ability Verification on the PANDORA Dataset: To further test the generalization ability and robustness of our framework on more complex tasks, we conducted experiments on the PANDORA dataset. This dataset requires the detection of oriented bounding boxes (RBFOV) with rotation angles, placing higher demands on the model's geometric understanding ability. As shown in Table 3, Sphere-CenterNet maintained its leading position in this more challenging task, achieving 15.3% mAP, 31.5% AP50, 12.7% AP75, and a Recall of 24.9%. Compared to the previous best method on this dataset (Sph2Pob CenterNet), our model achieved a substantial improvement of +4.7% in mAP (15.3% vs. 10.6%) and an even greater improvement of +5.6% in AP75 (12.7% vs. 7.1%). Achieving consistent and significant performance advantages on two datasets with different annotation paradigms fully demonstrates that our Sphere-CenterNet framework is not an overfitted design for a specific task but a universal solution with a general understanding of spherical geometry.

Methods	Backbone	AP	AP50	AP75
Multi-Kernel	ResNet-101	3.8	13.7	1.0
Sphere-SSD	ResNet-101	3.2	12.0	0.6
Reprojection R-CNN	ResNet-101	4.3	16.6	0.7
Sph-CenterNet	ResNet-101	5.5	19.9	1.1
R-CenterNet	ResNet-101	7.3	22.7	2.6
Sph2Pob CenterNet	ResNet-101	10.6	25.7	7.1
Ours	**ResNet-101**	**15.3**	**31.5**	**12.7**

Table 3. Performance comparison on the PANDORA validation set.

We believe the performance improvement mainly stems from three points: First, the use of Sph2Pob-IoU and IoU-type regression/sample assignment during training significantly reduced the gap between the training and evaluation objectives, making the optimization more directly oriented towards maximizing overlap, especially providing stronger gradients in high-overlap regions. Second, the spherical-aware feature extraction and multi-scale fusion showed positive complementarity: replacing with spherical convolution alone would change the sampling statistics and introduce instability, while adding EBiFPN Lite strengthened cross-layer alignment and feature reuse, allowing distortion correction information to be stably transmitted to the detection head, improving both localization and recall, which is fully consistent with our ablation results. Third, we engineered the handling of ERP's periodic continuity and high-latitude distortion (horizontal circular processing and latitude-aware enhancement), effectively reducing duplicate/missed detections across the left-right boundaries and center/shape devi-

ations in the polar regions, thereby improving the matching rate under strict thresholds.

4.4. Ablation Studies

To systematically validate the effectiveness of our proposed components and to deeply investigate their interactions, we conducted a series of detailed ablation studies based on the 360-Indoor dataset. We first established a strong baseline model, then analyzed the independent contributions of each core component and the synergistic effects produced when they were combined, observing the performance changes as shown in Table 4.

First, we independently evaluated the contribution of each core architectural component. As shown in the table, individually introducing the SphereConv frontend, our original LADCN module, or the EBiFPN Lite neck all led to performance improvements. Among them, the effect of the LADCN module was particularly significant, bringing a +1.8% AP gain. Similarly, the EBiFPN Lite neck also brought a +2.9% AP improvement, highlighting the importance of efficient and topologically continuous multi-scale feature fusion. These preliminary results validate the independent value of each of our module designs. **Feature Fusion Conflict:** When we began to explore combinations of modules, we observed more complex and profound phenomena. We unexpectedly found that directly combining SphereConv with LADCN led to a catastrophic performance drop (AP decreased by 2.8%). We speculate that this stems from a negative synergistic effect (antagonism) between the two design philosophies: the strict geometric prior of the low-level (SphereConv) may have constrained the learning space of the data-driven dynamic adaptability of the deep-level (LADCN), making it difficult for the model to optimize. **Positive Synergistic Effect:** In stark contrast, the combination of SphereConv and EBiFPN Lite showed a strong positive synergistic effect. This combination achieved a 13.2% AP, an improvement of +2.6% over the baseline. We believe that SphereConv acted as a "feature purifier" here, providing the powerful "fusion master" BiFPN with higher-quality, more geometrically regular low-level features, thus enabling it to perform cross-scale information fusion more effectively, ultimately achieving a 1+1>2 effect. Finally, this series of explorations led us to the ultimate victory. Our final model, which not only integrates all core architectural components but also adopts advanced training strategies including IoU loss and a Warmup learning rate strategy, ultimately achieved 16.8% AP, a +6.2% improvement over the baseline model. This result decisively proves that the excellent performance of our model does not come from any single component but from the powerful overall effect produced by the systematic integration of all geometry-aware modules with advanced training strategies. It proves that our core idea of building an end-to-end, highly self-consistent spherical-native detection framework is correct and extremely effective.

4.5. Qualitative Results

To more intuitively demonstrate the effectiveness of our Sphere-CenterNet framework, we provide some visualizations of detection results on the 360-Indoor validation set in Figure 7. These images were carefully selected to showcase our model's robust performance in handling the challenging scenarios unique to panoramic images. As shown, our model can not only accurately detect various objects in ordinary indoor scenes (Figure 7(a)) but, more importantly, it can successfully cope with the extreme geometric distortions caused by ERP projection. In Figure 7(b), the person near the bottom of the image (high-latitude region) is severely stretched, but our model, thanks to the Latitude-Aware Deformable Convolution (LADCN) module, can still generate compact and accurate bounding boxes. Furthermore, Figure 7(c) demonstrates our model's ability to handle circular topological continuity. The door in the figure spans the $\pm 180^\circ$ meridian boundary of the ERP image and is split into left and right parts in the 2D image. However, because our framework maintains topological continuity in both the feature fusion (circular convolution) and post-processing (circular NMS) stages, the model can correctly identify it as a single complete object instance and output a unified detection result. These qualitative results strongly prove that our Sphere-CenterNet has powerful 360° seamless perception capabilities. Figure 7(d) shows that objects with imbalanced aspect ratios in mid-to-high latitude regions can also be correctly detected.

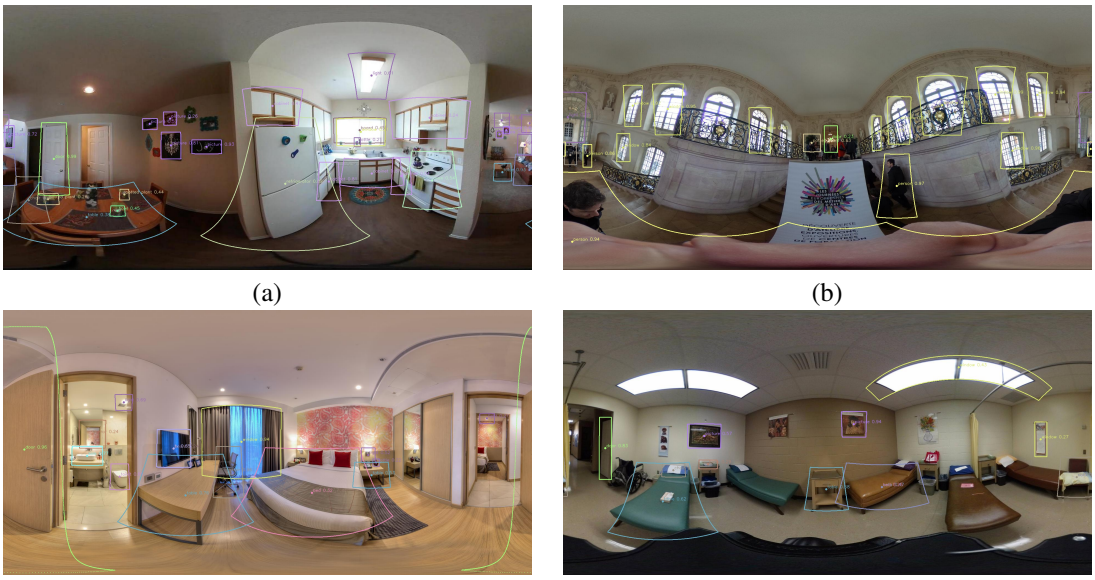
Figure 8 shows the visualization results for the Pandora dataset, which is annotated with rotated RBFOVs. Figure 8(a) shows that a large number of small objects in the dataset can be detected, reflecting the superiority of the seamless multi-scale fusion module EBiFPN in our work. Figure 8(b) shows that rotated windows can also be correctly detected, and the rotated RBFOV boxes better enclose the objects in the image.

5. Discussion and Limitations

While Sphere-CenterNet achieves state-of-the-art performance, there are limitations to be addressed. First, our evaluation is primarily conducted on indoor datasets (360-Indoor, PANDORA). Extending the framework to outdoor scenarios, such as autonomous driving, remains a future work. Second, although we alleviate distortion, detecting small objects near the poles is still challenging due to extreme stretching. Third, as noted in recent surveys [30], integrating multi-modal information (e.g., depth) could further resolve geometric ambiguities.

Methods			Metrics				vs. Base (%)		
SphConv	LADCN	EBiFPN Lite	AP	AP50	AP75	Recall	ΔAP	ΔAP50	ΔAP75
Baseline (ResDCN-101)			10.6	22.2	8.3	20.7	/	/	/
✓			9.4	19.5	7.8	19.9	↓1.2	↓2.7	↓0.5
	✓		12.4	26.9	9.8	23.1	↑1.8	↑4.7	↑1.5
		✓	13.5	28.9	10.9	22.8	↑2.9	↑6.7	↑2.6
✓	✓		7.8	19.2	5.1	25.7	↓2.8	↓3.0	↓3.2
✓		✓	13.2	28.1	10.9	22.1	↑2.6	↑5.9	↑2.6
✓	✓	✓	**16.8**	**32.8**	**15.2**	**23.4**	**↑6.2**	**↑10.6**	**↑6.9**

Table 4. Ablation study of core components. This table clearly shows the complete path from a simple baseline to our final high-performance model.



(a) (b) (c) (d)

Figure 7. Visualization on the 360-Indoor dataset.

6. Conclusion and Future Work

In this paper, we systematically investigated the core challenges of high-precision object detection in Equirectangular Projection (ERP) omnidirectional images, namely the severe geometric distortion and topological discontinuity caused by the projection from a sphere to a plane. To address these challenges, we proposed an end-to-end spherical-native detection framework called SphereCenterNet. The core of our work lies in designing a novel hybrid spherical feature extraction architecture. This architecture follows the design philosophy of using strict geometric priors at the low level and data-driven adaptation at the high level. It establishes geometrically correct initial features using strict Spherical Convolution (SphereConv) at the network frontend and adaptively models complex feature deformations using our original Latitude-Aware Deformable Convolution (LADCN) in the deep network layers. In addition, we constructed a seamless multi-scale feature fusion neck based entirely on circular convolution to ensure 360° topological continuity. The entire framework is optimized in an end-to-end process through a differentiable spherical IoU loss, achieving a high degree of unity between the training objective and the evaluation metric. Through extensive experiments on two representative public datasets, 360-Indoor and PANDORA, we have fully validated the effectiveness of the proposed components. The results of the ablation studies clearly demonstrate the performance gains brought by each of our innovations, while the comparison with existing state-of-the-art methods proves the compet-

(a) (b) (c) (d)

Figure 8. Visualization on the Pandora rotation dataset.

itiveness of our final model. The qualitative results also intuitively show the superior performance of the model in handling challenging scenarios such as extreme distortion and cross-boundary objects.

6.1. Future Work

We believe that the current work opens up several promising research directions for panoramic image understanding:

Extension to more complex detection tasks: The proposed hybrid feature extraction architecture can be more deeply integrated with the regression targets of Oriented Bounding Box (OBB) detection to fully unleash its potential on datasets like PANDORA. Furthermore, extending the Sphere-CenterNet framework to more complex downstream tasks, such as panoramic instance segmentation or object tracking in 360° videos, is a natural direction. **Model lightweighting and efficiency optimization:** The current model uses a deeper backbone network to pursue high accuracy. Exploring more lightweight and efficient implementations of spherical convolution, as well as model compression and knowledge distillation techniques, to promote the deployment of this technology on mobile and edge devices is a research direction with great practical application value. **Integration of multi-modal information:** Future research can explore the fusion of panoramic images with other modal information (such as depth maps, audio, LiDAR point clouds). For example, using depth information can help the model better understand the 3D structure of the scene, thereby further alleviating the ambiguity caused by projection distortion.

Acknowledgement

This work was supported by the Key Research and Development Project in Xinjiang Uygur Autonomous Region (No.2022B01006). We also explicitly thank the authors of Sph2Pob [17] and Unbiased IoU [8] for making their code and datasets publicly available, which greatly facilitated this research.

References

[1] H. Ai, Z. Cao, J. Zhu, H. Bai, Y. Chen, and L. Wang. Deep learning for omnidirectional vision: A survey and new perspectives. *arXiv preprint arXiv:2205.10468*, 2022. 3

[2] M. Cao, S. Ikehata, and K. Aizawa. Field-of-view iou for object detection in 360° images. *IEEE Transactions on Image Processing*, 2023. 4

[3] J. Chen, H. Mai, L. Luo, X. Chen, and K. Wu. Effective feature fusion network in bifpn for small object detection. In *2021 IEEE international conference on image processing (ICIP)*, pages 699–703. IEEE, 2021. 2

[4] S. Cho, R. Jung, and J. Kwon. Spherical transformer. *arXiv preprint arXiv:2202.04942*, 2022. 2

[5] S.-H. Chou, C. Sun, W.-Y. Chang, W.-T. Hsu, M. Sun, and J. Fu. 360-indoor: Towards learning real-world objects in 360deg indoor equirectangular images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 845–853, 2020. 5, 9

[6] T. Cohen, M. Weiler, B. Kicanaoglu, and M. Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *International conference on Machine learning*, pages 1321–1330. PMLR, 2019. 2

[7] B. Coors, A. P. Condurache, and A. Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the Eu-*

ropean conference on computer vision (ECCV), pages 518–533, 2018. 2, 4, 5, 6

[8] F. Dai, B. Chen, H. Xu, Y. Ma, X. Li, B. Feng, P. Yuan, C. Yan, and Q. Zhao. Unbiased iou for spherical image object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 508–515, 2022. 4, 9, 13

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 10

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 10

[11] A. Hedayat and W. D. Wallis. Hadamard matrices and their applications. *The annals of statistics*, pages 1184–1238, 1978. 8

[12] P. Jia, Y. Tie, L. Qi, and F. Zhu. Pv-yolo: An object detection model for panoramic video based on yolov4. In *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*, pages 56–61. IEEE, 2022. 4

[13] R. Khasanova and P. Frossard. Graph-based classification of omnidirectional images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 869–878, 2017. 1

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 9

[15] X. Lin, X. Ge, D. Zhang, Z. Wan, X. Wang, X. Li, W. Jiang, B. Du, D. Tao, M.-H. Yang, and L. Qi. One flight over the gap: A survey from perspective to panoramic vision, 2025. 1, 2, 3

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 3

[17] X. Liu, H. Xu, B. Chen, Q. Zhao, Y. Ma, C. Yan, and F. Dai. Sph2pob: Boosting object detection on spherical images with planar oriented boxes methods. In *IJCAI*, pages 1231–1239, 2023. 5, 13

[18] J. Miao, Y. Liu, K. Wang, J. Liu, A. Argyriou, Y. Han, and Z. Xu. Six-to-one: Cubemap-guided feature calibration for panorama object detection. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1320–1327. IEEE, 2022. 2

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[20] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3, 9

[21] Y.-C. Su and K. Grauman. Learning spherical convolution for fast features from 360 imagery. *Advances in neural information processing systems*, 30, 2017. 5

[22] Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 3

[23] R. Wang, R. Shivanna, D. Cheng, S. Jain, D. Lin, L. Hong, and E. Chi. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*, pages 1785–1797, 2021. 2, 4

[24] H. Xu, Q. Zhao, Y. Ma, X. Li, P. Yuan, B. Feng, C. Yan, and F. Dai. Pandora: A panoramic detection dataset for object with orientation. In *European conference on computer vision*, pages 237–252. Springer, 2022. 5, 9

[25] W. Yang, Y. Qian, J.-K. Kämäräinen, F. Cricri, and L. Fan. Object detection in equirectangular panorama. In *2018 24th international conference on pattern recognition (icpr)*, pages 2190–2195. IEEE, 2018. 2, 3

[26] D. Yu and S. Ji. Grid based spherical cnn for object detection from panoramic images. *Sensors*, 19(11):2622, 2019. 4

[27] R. Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019. 1, 2

[28] P. Zhao, A. You, Y. Zhang, J. Liu, K. Bian, and Y. Tong. Reprojection r-cnn: A fast and accurate object detector for 360 images. *arXiv e-prints*, pages arXiv–1907, 2019. 4

[29] P. Zhao, A. You, Y. Zhang, J. Liu, K. Bian, and Y. Tong. Spherical criteria for fast and accurate 360 object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12959–12966, 2020. 4

[30] X. Zheng, C. Liao, Z. Weng, K. Lei, et al. Panorama: The rise of omnidirectional vision in the embodied ai era. *arXiv preprint arXiv:2509.12989*, 2025. 1, 11

[31] D. Zhou, J. Fang, X. Song, C. Guan, J. Yin, Y. Dai, and R. Yang. Iou loss for 2d/3d object detection. In *2019 international conference on 3D vision (3DV)*, pages 85–94. IEEE, 2019. 4

[32] X. Zhou, V. Koltun, and P. Krähenbühl. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021. 3

[33] X. Zhou, D. Wang, and P. Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 3