Gaussian Label Distribution Learning for Spherical Image Object Detection

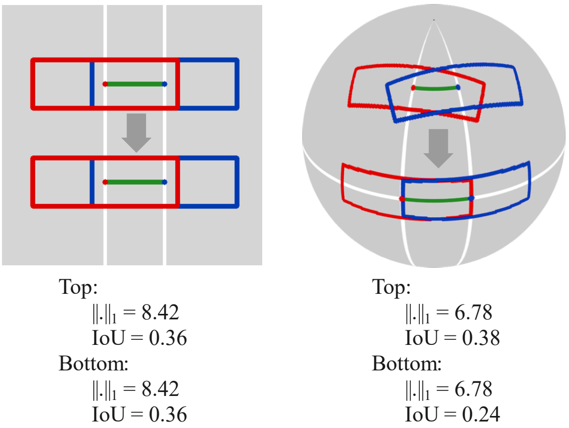
Hang Xu[1,2*] Xinyuan Liu[2] Qiang Zhao[2*†] Yike Ma[2] Chenggang Yan[1] Feng Dai[2†]
[1]Hangzhou Dianzi University, Hangzhou, China
[2]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
{hxu, cgyan}@hdu.edu.cn, {liuxinyuan21s, zhaoqiang, ykma, fdai}@ict.ac.cn

Abstract

Spherical image object detection emerges in many applications from virtual reality to robotics and automatic driving, while many existing detectors use l_n-norms loss for regression of spherical bounding boxes. There are two intrinsic flaws for l_n-norms loss, i.e., independent optimization of parameters and inconsistency between metric (dominated by IoU) and loss. These problems are common in planar image detection but more significant in spherical image detection. Solution for these problems has been extensively discussed in planar image detection by using IoU loss and related variants. However, these solutions cannot be migrated to spherical image object detection due to the undifferentiable of the Spherical IoU (SphIoU). In this paper, we design a simple but effective regression loss based on Gaussian Label Distribution Learning (GLDL) for spherical image object detection. Besides, we observe that the scale of the object in a spherical image varies greatly. The huge differences among objects from different categories make the sample selection strategy based on SphIoU challenging. Therefore, we propose GLDL-ATSS as a better training sample selection strategy for objects of the spherical image, which can alleviate the drawback of IoU threshold-based strategy of scale-sample imbalance. Extensive results on various two datasets with different baseline detectors show the effectiveness of our approach.

1. Introduction

In the past few years, with the numerous development of panoramic cameras with omnidirectional vision, the applications of spherical images and videos are also becoming more extensive, such as virtual & augmented reality [9,17,24], robotics [7,8,18], automatic driving [1,28,31], etc. As these spherical data increase, the demand for spherical vision analysis tasks increases, especially the object detection task of spherical image. However, compared with the large literature on planar image object detection [2, 12, 16, 34, 39], research in spherical image object detection is relatively in its earlier stage, with many open problems to solve.



(a) Planar image (b) Spherical image

Figure 1. Comparison between planar image and spherical image. (a) Moving centers of two bounding boxes in the planar image along the y-axis does not change the distance between the two centers, so IoU and L1 are unchanged. (b) Moving centers of two bounding boxes to the equator along the longitude changes the distance between two centers, which causes IoU decrease sharply whereas the L1 value is unchanged.

In spherical image object detection, a bounding box is represented by a Bounding Field of View (BFoV) [25]. Many existing detection benchmarks [4, 5, 26, 27, 29] use l_n-norms loss for the regression of BFoVs. However, the l_n-norms loss has some intrinsic flaws. First, parameters of the bounding box are optimized independently in the l_n-norms loss, leading to the detection accuracy sensitive to the fitting of any of the parameters. Second, Intersection

[*] This work was done when Hang Xu and Qiang Zhao were at ICT.
[†] Corresponding author.

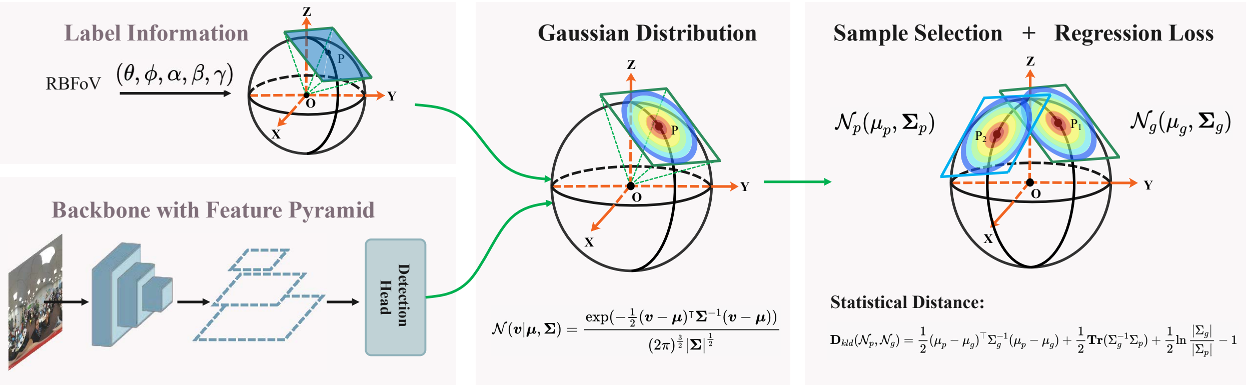

Label Information
RBFoV
$(\theta, \phi, \alpha, \beta, \gamma)$
Z
P
Y
O
X
Backbone with Feature Pyramid
Detection Head
Gaussian Distribution
Z
P
Y
O
X
$\mathcal{N}(\boldsymbol{v}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp(-\frac{1}{2}(\boldsymbol{v}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{v}-\boldsymbol{\mu}))}{(2\pi)^{\frac{3}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}$
Sample Selection + Regression Loss
Z
$\mathcal{N}_p(\mu_p, \boldsymbol{\Sigma}_p)$
$\mathcal{N}_g(\mu_g, \boldsymbol{\Sigma}_g)$
P_2
P_1
Y
O
X
Statistical Distance:
$\mathbf{D}_{kld}(\mathcal{N}_p, \mathcal{N}_g) = \frac{1}{2}(\mu_p - \mu_g)^\top \Sigma_g^{-1}(\mu_p - \mu_g) + \frac{1}{2}\mathbf{Tr}(\Sigma_g^{-1}\Sigma_p) + \frac{1}{2}\ln\frac{|\Sigma_g|}{|\Sigma_p|} - 1$

Figure 2. Overview of our main contributions. Gaussian distributions of spherical bounding boxes are constructed, and the sample selection strategy (GLDL-ATSS) and regression loss (GLDL loss) are designed in an alignment manner on the basis of K-L divergence. Note that the GLDL-ATSS and GLDL loss are not involved in the inference phase. Therefore, the inference time remains unchanged.

over Union (IoU) has been the standard metric for object detection, so l_n-norms as regression loss cause the negative impact of the inconsistency between metric and loss. These problems are common in planar image detection but more significant in spherical image detection. Fig. 1(b) shows the inconsistency between IoU and L1 Loss in the spherical image. Specifically, moving centers of bounding boxes to the equator along the longitude changes the distance between two centers, which causes IoU decrease sharply while the L1 value is unchanged. In contrast, as shown in Fig. 1(a), moving centers of bounding boxes in the planar image along the y-axis does not change the distance between the two centers, so IoU and L1 are unchanged. Solutions for these problems of l_n-norms loss have become recently popular in planar image detection by using IoU-induced loss, such as IoU loss [32], GIoU loss [23] and DIoU loss [37]. However, when calculating the intersection area of two spherical boxes, the number of intersection points needs to be obtained. When two spherical boxes are completely coincident or one edge is coincident, the number of intersection points will not be fixed and duplicate points will appear. The current SphIoU uses the DFS algorithm to remove these duplicate intersection points. To the best of our knowledge, the DFS algorithm is undifferentiable. Therefore, Spherical IoU (SphIoU) [5, 27] is undifferentiable and these solutions based on IoU loss cannot be migrated to spherical image object detection. The more recent work [30] finds the key to maintaining the consistency between metric and regression loss lies in the trend-level consistency between regression loss and IoU loss rather than value-level consistency, which greatly decreases the difficulty of designing alternatives.

In this paper, we design a simple but effective regression loss based on Gaussian Label Distribution Learning (GLDL) for spherical image object detection. Specifically, in the training phase, we first convert tangent planes of the predicted spherical bounding box and ground truth box into the Gaussian distribution. Then, we devise a dynamic sample selection strategy (GLDL-ATSS) to select positive samples, which can alleviate the drawback of IoU threshold-based strategy of scale-sample imbalance. Finally, we design a regression loss function based on GLDL for spherical object detection task. We observe that GLDL loss achieves a trend-level alignment with SphIoU loss. In the inference phase, we directly obtain the output for the spherical bounding box from the trained model of the parameter weights, so the inference time of the network remains unchanged. The entire framework of the method in this paper is shown in Fig. 2. **The highlights of this paper are as follows:**

- We explore a new regression loss function based on Gaussian Label Distribution Learning (GLDL) for spherical object detection task. It achieves a trend-level alignment with SphIoU loss and thus naturally improves the model.
- We align the measurement between sample selection and loss regression based on the GLDL, and then construct new dynamic sample selection strategies (GLDL-ATSS) accordingly. GLDL-ATSS can alleviate the drawback of IoU threshold-based strategy (i.e., scale-sample imbalance).
- Extensive experimental results on two datasets and popular spherical image detectors show the effectiveness of our approach.

2. Related Work

2.1. Spherical Objection Detection

Spherical image object detection is an emerging direction, which attempts to extend classical planar detectors to the spherical case by adopting the spherical bounding boxes. Campared to planar images, objects are often arbitrary-oriented in spherical images. To this end, Multi-kernel [26] and Reprojection R-CNN [36] are two-stage mainstreamed approach whose pipeline is inherited from Faster RCNN [22], while Multi-projection YOLO [29], SphereNet [4], SpherePHD [11], Sphere-CenterNet [5] and R-CenterNet [27] are based on single-stage methods for faster detection speed. The regression loss of the above algorithms use l_n-norms loss due to the undifferentiable of implementing SphIoU [5, 27].

2.2. Variants of IoU-based Loss

The inconsistency between metric and regression loss is a common issue for the object detection task. The use of IoU-related loss in planar image detection has been extensively considered as a solution to this contradiction. For instance, Unitbox [33] proposes an IoU loss which regresses the four bounds of a predicted box as a whole unit. More works extend the idea of Unitbox by introducing GIoU [23], DIoU [37] and CIoU [38] for bounding box regression. However, their applications to the spherical image object detection are difficult due to the undifferentiable of implementing SphIoU [5, 27].

2.3. Sample Selection Strategies

Sample selection plays an important role in object detection task [10], and many sample selection strategies have been proposed in object detection. Many object detection methods, for instance, Faster RCNN [22], SSD [15], and RetinaNet [13], adopt a fixed max IoU strategy, requiring predefined positive and negative thresholds in advance. To overcome the difficulty in setting fixed thresholds, ATSS [34] uses statistical characteristics to calculate dynamic IoU thresholds and achieves good results. Additionally, other excellent dynamic sample selection strategies exist, for example, DAL [19] dynamically assigns samples according to a defined matching degree, and FreeAnchor [35] dynamically selects labels under the maximum likelihood principle. These methods depend on the IoU as the main metric for evaluating the quality of the sample.

3. The Proposed Method

3.1. Overview

In this section, we present our main approach. Fig. 2 shows an overview of the proposed method. In the training phase, we convert tangent planes of the predicted spherical bounding box and ground truth box into the Gaussian distribution. Then, we devise the dynamic sample selection strategy (GLDL-ATSS) to select positive samples. Finally, we design a regression loss based on Gaussian Label Distribution Learning (GLDL) for spherical object detection task. In the inference phase, we directly obtain the output for spherical bounding box from the trained model of the parameter weights. Compared with the baseline, we only modify the sample selection strategy and loss function in the training phase, which does not change parameters of the network, so the inference time remains unchanged. We elaborate our method in the following sections.

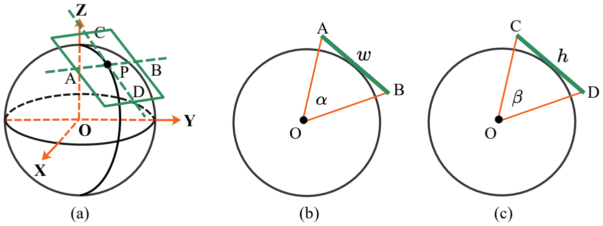


Figure 3. A diagram of the rectangular tangent plane of the spherical bounding box. $P(\theta, \phi)$ is the tangent point of the sphere and rectangular tangent plane. Points A, B, C, and D are the center points of tangent plane edges. α and β are the horizontal and vertical fields of view of the spherical bounding box, respectively.

3.2. Representation of Spherical Bounding Box

As shown in Fig. 3(a), in spherical vision, when looking at a point $P(\theta, \phi)$ from the sphere center O with the horizontal and vertical field of view α and β, respectively, we can see a Field of View (FoV) of the sphere surface, and the P is the center of view. Bounding Field of View (BFoV) is defined as a spherical bounding box $(\theta, \phi, \alpha, \beta)$ in the spherical object detection task [3, 5, 36]. Currently, Rotated Bounding Field of View (RBFoV) [27] $(\theta, \phi, \alpha, \beta, \gamma)$ is proposed to better accurately or compactly outline oriented instances in spherical images. The γ represents the angle of the rotation of the BFoV around the axis $\overrightarrow{OP}$. The range of values of γ is $[-90^\circ, 90^\circ)$. The points on a RBFoV can be projected to a rectangular tangent plane $\mathbf{\Pi}$ by gnomonic projection [21]. The tangent plane $\mathbf{\Pi}$ is defined by $(\theta, \phi, w, h, \gamma)$, where $w = 2\tan(0.5\alpha)$ and $h = 2\tan(0.5\beta)$ are the width and height of tangent plane, respectively, as illustrated in Fig. 3(b,c).

3.3. Gaussian Label Distribution Learning

Next, we introduce our the Gaussian Label Distribution Learning (GLDL) approach. Firstly, when the object is in the polar region and its tangent plane is $\mathbf{\Pi}_0(\theta_0, \phi_0, w, h, \gamma)$, it can be converted into a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_0, \mathbf{\Sigma}_0)$ (see Fig. 4(a)) by the following formula:

$$\begin{aligned}
\boldsymbol{\mu}_0 &= [\sin(\phi_0)\cos(\theta_0), \sin(\phi_0)\sin(\theta_0), \cos(\phi_0)] \\
\mathbf{\Sigma}_0 &= \mathbf{R}\mathbf{\Lambda}\mathbf{R}^\top, \\
\mathbf{R} &= \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{\Lambda} = \begin{bmatrix} \frac{w^2}{4} & 0 & 0 \\ 0 & \frac{h^2}{4} & 0 \\ 0 & 0 & 0 \end{bmatrix}
\end{aligned} \tag{1}$$

where $\mathbf{R}$ represents the rotation matrix, and $\mathbf{\Lambda}$ represents the covariance matrix.

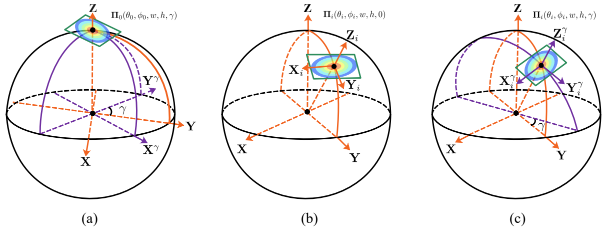

Z
$\Pi_0(\theta_0, \phi_0, w, h, \gamma)$
Y^γ
γ
Y
X^γ
X
(a)
Z
$\Pi_i(\theta_i, \phi_i, w, h, 0)$
Z_i
X_i
Y_i
X
Y
(b)
Z
$\Pi_i(\theta_i, \phi_i, w, h, \gamma)$
Z_i^γ
X_i^γ
Y_i^γ
γ
X
Y
(c)

Figure 4. A schematic diagram of modeling a spherical bounding box by a Gaussian distribution.

Secondly, when the object is in a region other than the polar region, we establish a new coordinate system based on the object's tangent plane $\mathbf{\Pi}_i(\theta_i, \phi_i, w, h, 0)$, as illustrated in Fig. 4(b). The new coordinate system derivation is given in the supplementary material.

$$\begin{cases} \mathbf{X}_i = [\sin(\theta_i), -\cos(\theta_i), 0]^\top \\ \mathbf{Y}_i = [\cos(\phi_i)\cos(\theta_i), \cos(\phi_i)\sin(\theta_i), -\sin(\phi_i)]^\top \\ \mathbf{Z}_i = [\sin(\phi_i)\cos(\theta_i), \sin(\phi_i)\sin(\theta_i), \cos(\phi_i)]^\top \end{cases} \tag{2}$$

At this time, the Gaussian distribution $\mathcal{N}(\mu_i, \mathbf{\Sigma}_i)$ (see Fig. 4(c)) of the tangent plane $\mathbf{\Pi}_i(\theta_i, \phi_i, w, h, \gamma)$ of the object can be obtained from the following formula:

$$\begin{aligned} \boldsymbol{\mu}_i &= (\sin(\phi_i)\cos(\theta_i), \sin(\phi_i)\sin(\theta_i), \cos(\phi_i)) \\ \mathbf{\Sigma}_i &= \mathbf{R}(\mathbf{T}\mathbf{\Lambda}\mathbf{T}^\top)\mathbf{R}^\top, \ \mathbf{T} = [\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i] \end{aligned} \tag{3}$$

where $\mathbf{T}$ represents the rotation matrix from the original coordinate system to the new coordinate system, and $\mathbf{\Lambda}$ is given by Eq. 1.

Finally, the Kullback-Leibler (K-L) divergence is adopted to measure the similarity between the ground-truth distribution $\mathcal{N}_g(\boldsymbol{\mu}_g, \mathbf{\Sigma}_g)$ and predicted distribution $\mathcal{N}_p(\boldsymbol{\mu}_p, \mathbf{\Sigma}_p)$.

$$\begin{aligned} D_{kl}(\mathcal{N}_p, \mathcal{N}_g) &= \frac{1}{2}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_g)^\top \mathbf{\Sigma}_g^{-1}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_g) \\ &+ \frac{1}{2}tr(\mathbf{\Sigma}_g^{-1}\mathbf{\Sigma}_p) + \frac{1}{2}\ln\frac{|\mathbf{\Sigma}_g|}{|\mathbf{\Sigma}_p|} - 1 \end{aligned} \tag{4}$$

Eq. 4 shows that each item of K-L Divergence contains size parameters $\mathbf{\Sigma}(\alpha, \beta)$ and center parameters $\boldsymbol{\mu}(\theta, \phi)$. All parameters of the bounding box form a chain coupling relationship and influence each other. In other words, optimizing one parameter will also promote the optimization of other parameters, which is similar with SphIoU loss and very conducive to optimization of the detector.

To further compare the behavior of SphIoU loss and GLDL loss, we conduct several cases. Fig. 6(a) shows the curves of three loss forms for two spherical bounding boxes when moving two spherical bounding boxes along the longitude line. Only the L1 loss curve is constant, while GLDL and SphIoU value will change as the longitude varies. Fig. 6(b) shows the changes of the three losses under different aspect ratio conditions. It can be seen that the L1 value of the two spherical bounding boxes is constant, but GLDL and SphIoU will change as the aspect ratio varies. Fig. 6(c) is the scatter plot between SphIoU loss and GLDL loss, L1 loss when 1000 sample pairs are randomly generated. It can be seen that regardless of the case, GLDL loss can maintain a more consistent trend with SphIoU loss than L1 loss.

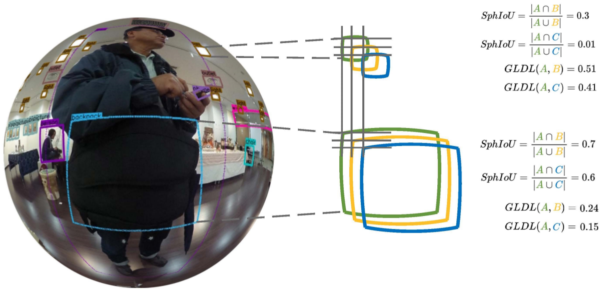

$SphIoU = \frac{|A \cap B|}{|A \cup B|} = 0.3$
$SphIoU = \frac{|A \cap C|}{|A \cup C|} = 0.01$
$GLDL(A, B) = 0.51$
$GLDL(A, C) = 0.41$
$SphIoU = \frac{|A \cap B|}{|A \cup B|} = 0.7$
$SphIoU = \frac{|A \cap C|}{|A \cup C|} = 0.6$
$GLDL(A, B) = 0.24$
$GLDL(A, C) = 0.15$

Figure 5. Compared to GLDL, the sensitivity of SphIoU to objects with different scales is of large variance.

3.4. Regression Loss

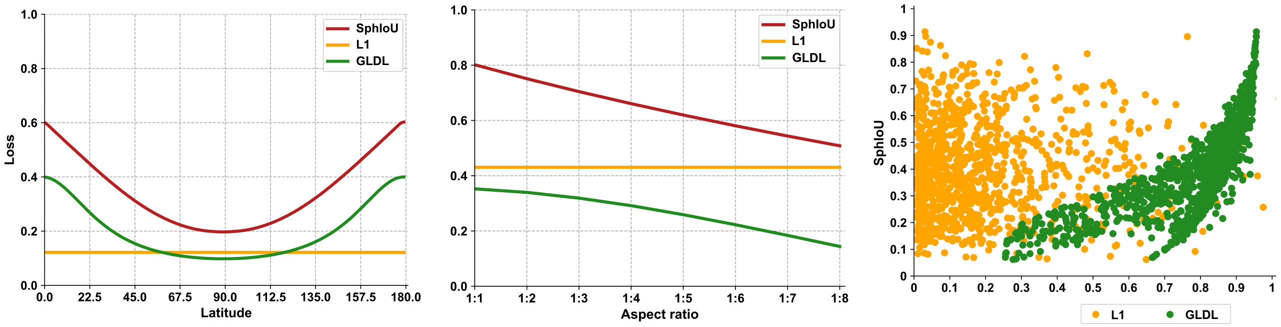
The range of actual value obtained by the Kullback-Leibler divergence between Gaussian distributions is too large to be the regression loss, which leads to difficult convergence. Therefore, normalization is necessary so that the GLDL can be used as regression loss. The normalized regression loss functions for GLDL is defined as

$$\mathcal{L}_{reg} = 1 - \frac{1}{\tau + f(D_{kl})} \tag{5}$$

where $f(\cdot)$ denotes a normalized function to transform the distance D_{kl} to make the loss more smooth and expressive. In the ablation study, we devise a series of small-scale experiments, in which some empirical functions are tested, and the best functions for different metrics are chosen according to the results, details will be shown in Tab. 3. The τ is a hyperparameter, we experimentally observe that its choice is robust in a certain range, details will be shown in Tab. 1.

The regression process of the bounding box is divided into five steps, namely, 1) predict offset $(t_\theta^*, t_\phi^*, t_\alpha^*, t_\beta^*, t_\gamma^*)$, 2) decode the prediction box by the offset value, 3) convert tangent planes of prediction box and target ground-truth box into Gaussian distribution, 4) calculate K-L divergence between two Gaussian distributions, 5) convert the K-L divergence into the final regression loss.

The regression equation of the spherical bounding box is



(a)

(b)

(c)

Figure 6. Behavior comparison of three losses (SphIoU, GLDL and L1) in different cases. (a) Moving center points of two bounding boxes along the longitude changes the distance between two center points, which causes SphIoU and GLDL will decrease sharply while the L1 is unchanged. (b) When center points of two bounding boxes are fixed, SphIoU and GLDL will decrease as the aspect ratio increases, while the L1 is unchanged. (c) When 1,000 examples are randomly generated, GLDL can miantain a more consistent trend to SphIoU than L1.

as follows:

$$\begin{aligned}
t_\theta &= (\theta - \theta_a)/\alpha_a, t_\phi = (\phi - \phi_a)/\beta_a \\
t_\alpha &= \log(\alpha/\alpha_a), t_\beta = \log(\beta/\beta_a) \\
t_\gamma &= \gamma - \gamma_a \\
t^*_\theta &= (\theta^* - \theta_a)/\alpha_a, t^*_\phi = (\phi^* - \phi_a)/\beta_a \\
t^*_\alpha &= \log(\alpha^*/\alpha_a), t^*_\beta = \log(\beta^*/\beta_a), \\
t^*_\gamma &= \gamma^* - \gamma_a
\end{aligned} \tag{6}$$

where $\theta, \phi, \alpha, \beta, \gamma$ denote the spherical bounding box's center coordinates, width, height and angle, respectively. Variables $\theta, \theta_a, \theta^*$ are for the ground-truth box, anchor box, and predicted box, respectively (likewise for $\phi, \alpha, \beta, \gamma$).

3.5. Sample Selection Based on GLDL-ATSS

When training an object detector, we first need to define positive and negative samples for classification, and then use positive samples for regression. Therefore, the sample selection strategy is another key task for object detection. The most popular sample selection strategy is the IoU-based strategy, which selects a sample by comparing IoU values between proposals and ground truth with threshold.

However, some significant problems exist with the sample selection strategy based on SphIoU in spherical image object detection. Firstly, there are many objects of different sizes and categories in a spherical image since the spherical image has a 360° view. Therefore, the scale of the object in a spherical image varies greatly. The huge differences among objects from different categories make the sample selection strategy based on SphIoU challenging. As shown in Fig. 5, we observe that the sensitivity of SphIoU to objects with different scales is of significant variance. Specifically, a minor location deviation for the tiny object will lead to a notable SphIoU drop (from 0.3 to 0.01), resulting in inaccurate positive and negative sample selection. For the larger object, the SphIoU changes slightly (from 0.7 to 0.6) with the same location deviation. Secondly , SphIoU fails to reflect the positional relationship of two spherical rectangles when they have no overlap or are mutually inclusive (i.e., the value of IoU keeps constant), which is often the case for tiny spherical bounding boxes. Finally, the metrics in sample selection and the regression loss have a misalignment when the regression loss is based on the GLDL.

In this paper, we propose a new sample selection strategy based on GLDL. Specifically, GLDL being a distance metric with a value range $[0, +\infty)$ cannot be directly used as a similarity metric for selecting samples. To obtain a value range similar to SphIoU (i.e., between 0 and 1), we select to normalize its value range to $(0, 1]$.

$$f(D_{kl}) = \frac{1}{c + D_{kl}(\mathcal{N}_g, \mathcal{N}_p)}, \tag{7}$$

where c is a hyperparameter, we experimentally observe that its choice is robust in a certain range, details will be shown in Tab. 2.

However, the sample selection strategy based on GLDL is not as intuitive as SphIoU in determining the threshold. To avoid the difficulty of selecting the optimal threshold, inspired by ATSS [34], the threshold for selecting samples is calculated dynamically according to the statistical characteristics of all the normalized distance. For the i-th ground truth, the dynamic threshold t_g is calculated as

$$\begin{aligned}
m^i_g &= \frac{1}{N}\sum_{j=1}^{N} f^{i,j}(D_{kl}) \\
v^i_g &= \sqrt{\frac{1}{N}\sum_{j=1}^{N}(f^{i,j}(D_{kl}) - m^i_g)^2} \\
t^i_g &= m^i_g + v^i_g
\end{aligned} \tag{8}$$

where N is the number of candidate samples, and $f^{i,j}(D_{kl})$ is the normalized GLDL between the i-th ground truth and

Dataset	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	**baseline**
360-Indoor	20.3	**20.5**	20.4	20.0	19.4	17.6
PANDORA	20.1	**20.3**	20.1	19.9	19.7	17.2

Table 1. The AP of different τ on different datasets. The τ is used for adjusting the GLDL loss.

Dataset	$c = 1$	$c = 2$	$c = 3$	$c = 4$	$c = 5$	**baseline**
360-Indoor	21.5	**21.8**	21.7	21.4	21.2	20.1
PANDORA	20.9	**21.3**	21.2	21.1	22.8	19.6

Table 2. The AP performance under different c on different datasets. The c is used for adjusting the number of positive samples in GLDL-ATSS.

the j-th proposal. Finally, positive samples are selected using the general assignment strategy, that is, candidates are selected whose similar values are greater than or equal to the threshold t_g^i.

4. Experiments

4.1. Datasets and Implementation Details

Datasets. *360-Indoor* [3] is the first released real-world spherical object detection dataset up to now. It consists of 3,335 indoor spherical images and 89,148 Bounding FoVs (BFoVs) annotations among 37 categories. Before 360-Indoor was presented, evaluations were made with synthetic data alone, which did not reflect the complex scenes of the real world. *PANDORA* [27] is the first real-world dataset to use the Rotated Bounding FoVs (RBFoVs) annotations in spherical object detection. It consists of 3,000 indoor spherical images and 94,353 instances RBFoVs annotations among 47 categories. All images of two dataset are with 960 × 1920 resolution. The proportion of the training set, validation set, and testing set is 1/2, 1/6, and 1/3, respectively. For training and testing, the images' resolution is all resized to 512 × 1024.

Metric. The widely used mAP [14] is adopted to evaluate the performance of detectors in all our experiments. Furthermore, AP values are calculated based on SphIoU to adapt to the spherical bounding box and produce an accurate result.

Training details. All approaches are implemented in PyTorch [20], and training is done on 8 GeForce RTX 2080Ti GPUs with a batch size of 32 and input resolution is 512×1024. We train detectors by updating their regression loss and sample selection strtegy using the proposed our method. Since the detector is optimized by classification and regression losses, we can easily replace the regression one with GLDL loss while keeping the original classification loss. SGD is adopted to optimize the models with momentum set to 0.9 and weight decay set to 0.0005. All evaluated models are trained for 120 epochs with an initial learning rate of 0.001 which is then divided by 10 at 60 epochs and again at 90 epochs. To make it fair, we keep all the approaches' settings and hyper parameters the same as depicted in corresponding papers.

Dataset	**Function of $\mathcal{S}_{\text{ss}}$**	**Function of $\mathcal{L}_{\text{reg}}$**	**AP$_{50}$**
360-Indoor	$\frac{1}{2+D_{kl}}$	$1 - \frac{1}{2+\sqrt{D_{kl}}}$	**25.0**
	$\frac{1}{2+D_{kl}}$	$1 - e^{-\sqrt{D_{kl}}}$	23.1
	$\frac{1}{2+D_{kl}}$	$1 - e^{-D_{kl}}$	18.36
	$\frac{1}{2+\sqrt{D_{kl}}}$	$1 - \frac{1}{2+\sqrt{D_{kl}}}$	13.6
PANDORA	$\frac{1}{2+D_{kl}}$	$1 - \frac{1}{2+\sqrt{D_{kl}}}$	**25.2**
	$\frac{1}{2+D_{kl}}$	$1 - e^{-\sqrt{D_{kl}}}$	23.9
	$\frac{1}{2+D_{kl}}$	$1 - e^{-D_{kl}}$	20.7
	$\frac{1}{2+\sqrt{D_{kl}}}$	$1 - \frac{1}{2+\sqrt{D_{kl}}}$	15.3

Table 3. Experiment results of different normalized function of $\mathcal{S}_{\text{ss}}$ and $\mathcal{L}_{\text{reg}}$ on 360-Indoor and PANDORA dataset.

Loss	**Normalized Function**	**360-Indoor**	**PANDORA**
		AP$_{50}$	**AP$_{50}$**
Smooth L1	w/	13.7	12.9
	w/o	**17.6**	**17.2**

Table 4. Analysis of normalized function. The based detector is RetinaNet.

Dataset	**Backbone**	$\mathcal{S}_{\text{ss}}$	$\mathcal{L}_{\text{reg}}$	**AP$_{50}$**
360-Indoor	R-101	$\mathcal{S}_{\text{IoU}}$ (Fixed)	$\mathcal{L}_{\text{L1}}$	17.6
	R-101	$\mathcal{S}_{\text{IoU}}$ (Fixed)	$\mathcal{L}_{\text{GLDL}}$	20.7 **(+3.1)**
	R-101	$\mathcal{S}_{\text{GLDL}}$ (Fixed)	$\mathcal{L}_{\text{GLDL}}$	22.8 **(+5.2)**
	R-101	$\mathcal{S}_{\text{IoU}}$ (ATSS)	$\mathcal{L}_{\text{L1}}$	20.1
	R-101	$\mathcal{S}_{\text{IoU}}$ (ATSS)	$\mathcal{L}_{\text{GLDL}}$	22.3 **(+2.2)**
	R-101	$\mathcal{S}_{\text{GLDL}}$ (ATSS)	$\mathcal{L}_{\text{GLDL}}$	25.0 **(+4.9)**
PANDORA	R-101	$\mathcal{S}_{\text{IoU}}$ (Fixed)	$\mathcal{L}_{\text{L1}}$	17.2
	R-101	$\mathcal{S}_{\text{IoU}}$ (Fixed)	$\mathcal{L}_{\text{GLDL}}$	21.4 **(+4.2)**
	R-101	$\mathcal{S}_{\text{GLDL}}$ (Fixed)	$\mathcal{L}_{\text{GLDL}}$	22.7 **(+5.5)**
	R-101	$\mathcal{S}_{\text{IoU}}$ (ATSS)	$\mathcal{L}_{\text{L1}}$	19.6
	R-101	$\mathcal{S}_{\text{IoU}}$ (ATSS)	$\mathcal{L}_{\text{GLDL}}$	23.4 **(+3.8)**
	R-101	$\mathcal{S}_{\text{GLDL}}$ (ATSS)	$\mathcal{L}_{\text{GLDL}}$	25.2 **(+5.6)**

Table 5. Ablation study of each component. $\mathcal{S}_{\text{ss}}$ and $\mathcal{L}_{\text{reg}}$ represent the sample selection strategy and the regression loss function, respectively.

4.2. Ablation Study

To verify the effectiveness of the proposed modules individually and to exclude the randomness of hyper-parameters, we design the following ablation study. We use

Method	Backbone	$\mathcal{S}_{ss}$		$\mathcal{L}_{reg}$		360-Indoor			PANDORA		
		$\mathcal{S}_{\text{IoU}}$	$\mathcal{S}_{\text{GLDL}}$	$\mathcal{L}_{\text{L1}}$	$\mathcal{L}_{\text{GLDL}}$	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Multi-Kernel [26]	R-101	✓		✓		4.7	11.1	2.8	4.2	10.8	2.2
	R-101	✓			✓	7.2(**+2.5**)	14.2(**+4.1**)	5.4(**+2.4**)	7.8(**+3.6**)	15.6(**+4.8**)	4.3(**+2.1**)
	R-101		✓	✓		6.8(**+2.1**)	13.9(**+2.8**)	4.7(**+1.9**)	6.2(**+2.0**)	14.5(**+3.7**)	3.9(**+1.7**)
	R-101		✓		✓	9.3(**+4.6**)	17.2(**+6.1**)	6.6(**+3.8**)	10.2(**+6.0**)	17.6(**+6.8**)	6.9(**+4.4**)
Sphere-SSD [4]	R-101	✓		✓		2.9	7.8	1.4	2.3	7.7	1.5
	R-101	✓			✓	5.6(**+2.7**)	10.8(**+3.0**)	4.2(**+2.8**)	5.9(**+3.6**)	12.3(**+4.6**)	4.9(**+3.4**)
	R-101		✓	✓		4.9(**+2.0**)	10.2(**+2.4**)	3.7(**+2.3**)	4.1(**+1.8**)	9.8(**+2.1**)	3.2(**+1.7**)
	R-101		✓		✓	7.8(**+4.9**)	12.6(**+4.8**)	5.4(**+4.0**)	8.0(**+5.7**)	13.8(**+6.1**)	6.8(**+5.3**)
Reprojection R-CNN [36]	R-101	✓		✓		5.0	15.3	1.9	4.2	14.7	1.8
	R-101	✓			✓	7.5(**+2.5**)	18.2(**+2.9**)	3.8(**+1.9**)	7.9(**+3.7**)	18.7(**+4.0**)	4.5(**+2.7**)
	R-101		✓	✓		7.1(**+2.1**)	17.8(**+2.5**)	3.2(**+1.3**)	6.8(**+2.6**)	17.4(**+2.7**)	3.0(**+1.2**)
	R-101		✓		✓	10.8(**+5.8**)	22.5(**+7.2**)	5.3(**+3.4**)	11.1(**+6.9**)	22.8(**+8.1**)	5.8(**+4.0**)
Sphere-CenterNet [5]	R-101			✓		10.0	24.8	6.0	-	-	-
	R-101				✓	11.2(**+1.1**)	26.1(**+1.3**)	7.4(**+1.4**)	-	-	-
R-CenterNet [27]	R-101			✓		-	-	-	7.3	22.7	2.6
	R-101				✓	-	-	-	8.7(**+1.4**)	24.3(**+1.6**)	4.5(**+1.9**)

Table 6. Comparison of the performance of different methods of the Gaussian distances as metrics for sample selection and regression loss on 360-Indoor and PANDORA dataset. Compared with Sphere-CenterNet, R-CenterNet only adds an Angle regression branch to regress the RBFoV. Therefore, we did not do experiments for Sphere-CenterNet on PANDORA and R-CenterNet on 360-Indoor.

the one-stage detector RetinaNet [13] as the baseline in ablation study. Different from the original RetinaNet, anchor boxes now use spherical bounding boxes, and IoU of sample selection uses SphIoU.

Different Hyper-parameter Value. There are two hyper-parameters in our design. First, when designing GLDL loss, we use a constant τ (Eq. 5) to modulate the regression loss. Note that the sample selection uses the SphIoU-based strategy in this experiment. The results are shown in Tab. 1. We observe that when changing τ in a certain range (from 1 to 5), the value of AP waves marginally and is much higher than baseline. It indicates that the choice of τ is robust in this range. From Tab. 1, the overall performance of using $\tau = 2$, the model achieves the best performance. Second, there is a hyper-parameter c (Eq. 7) in GLDL-ATSS, which is used to adjust the number of positive samples assigned to each instance. Note that the regression loss uses the L1 loss in this experiment. We set c to 1, 2, 3, 4 and 5 to test its performance. From the results in Tab. 2, we can see that when setting c to 2, the best performance can be attained, so $c = 2$ is chosen as the default setting.

Different Normalized Function. GLDL being a distance metric with a value range $[0, +\infty)$ cannot be directly used as a similarity metric for the regression loss and selecting samples. Therefore, normalization is necessary for GLDL. As shown in Tab. 3, some empirical functions, such as $\sqrt{(\cdot)}$ and $e^{(\cdot)}$, are tested as the normalized function in the experiments. Finally, the normalized function of sample selection and loss is $\frac{1}{2+\mathbf{D}_{kl}}$ and $1 - \frac{1}{2+\sqrt{\mathbf{D}_{kl}}}$ according to the best results of the experiments.

Analysis of Normalized Function. As mentioned above, the purpose of using the normalized function is to normalize and smooth the quick increase trend of GLDL. If the normalization operation is not used, the range of loss value of GLDL is too large, resulting in NAN in training. To verify that the effectiveness of our method does not come from the normalized function, we perform a normalization operation on the Smooth L1 loss to get rid of the interference brought by the normalized function. As shown in Tab. 4, the performance of Smooth L1 suffers significantly when the normalized function is used. Tab. 4 proves that the effectiveness of GLDL loss does not come from the normalized function.

Analysis of Each Component. Tab. 5 compares the performances when different evolution metrics, SphIoU, GLDL and L1, are used in fixed and dynamic sample selection strategies and regression loss. The predefined positive and negative thresholds in the fixed strategy for SphIoU is both 0.5 and 0.4. Even if only the L1 loss is replaced by $\mathcal{L}_{\text{GLDL}}$, the performance of our method is better than that of baseline (**20.7** vs. **17.6**, **21.4** vs. **17.2**). The performances based on fixed sample selection strategies varied greatly as a result of the hand-crafted hyperparameters. Therefore, experiments based on dynamic sample selection strategies are constructed to objectively compare the performances of the metrics. Additionally, the superiority of our method is clearly demonstrated when the sample selection strategies ($\mathcal{S}_{\text{GLDL-ATSS}}$) are used (**25.0** vs. **22.3**, **25.2** vs. **23.4**). The experimental results demonstrated that the overall performances of the proposed methods surpassed that of the base-

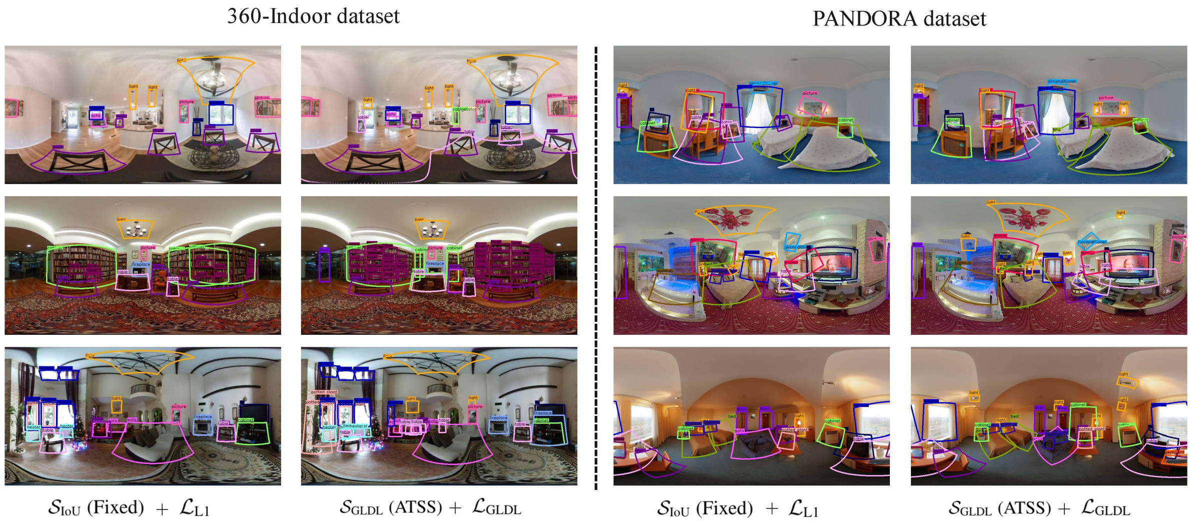

360-Indoor dataset
PANDORA dataset
$\mathcal{S}_{\text{IoU}}$ (Fixed) $+$ $\mathcal{L}_{\text{L1}}$
$\mathcal{S}_{\text{GLDL}}$ (ATSS) $+$ $\mathcal{L}_{\text{GLDL}}$
$\mathcal{S}_{\text{IoU}}$ (Fixed) $+$ $\mathcal{L}_{\text{L1}}$
$\mathcal{S}_{\text{GLDL}}$ (ATSS) $+$ $\mathcal{L}_{\text{GLDL}}$

Figure 7. Visual comparisons on the 360-Indoor and PANDORA dataset. The detector is RetinaNet.

line. (**25.0** vs. **17.6**, **25.2** vs. **17.2**).

We visualize some of the predicted results shown in Fig. 7, which includes 360-Indoor dataset and PANDORA dataset. From the results, we can observe that our methods can get more precise predicted results for both 360-Indoor and PANDORA dataset. The visualization results are consistent with the data results in Tab. 5.

4.3. Evaluations

Baseline Methods. To verify that our method can be applied into any spherical image detector and boost the detector performance, we select five state-of-the-art spherical image detectors for testing, including one-stage anchor-based detectors: Sphere-SSD [4], two-stage anchor-based detectors: Multi-Kernel [26] and Reprojection R-CNN [36], one-stage anchor-free detectors: Sphere-CenterNet [5] and R-CenterNet [27]. To make it fair, we keep all the experiments settings and hyper parameters the same as depicted in corresponding papers. The backbone networks in all the methods are all the same ResNet-101 [6] architecture.

Quantitative Results. Tab. 6 compares the detection results of using GLDL-ATSS and GLDL loss on different datasets and detectors. We verify the effectiveness of GLDL-ATSS and GLDL loss by respectively replacing the SphIoU-based sample selection and Smooth L1-based regression loss in the original network of baseline methods. For each dataset, we provide the AP, AP^{50} and AP^{75} performance. Tab. 6 shows that detectors based on GLDL-ATSS and GLDL loss improve the AP metric of Multi-Kernel, Sphere-SSD, and Reprojection R-CNN on 360-Indoor and PANDORA datasets. Since Sphere-CenterNet and R-CenterNet are anchor-free methods, we only use the GLDL loss instead of Smooth L1 loss. It can be seen that the performance of Sphere-CenterNet and R-CenterNet based on GLDL loss is better than that of baseline on 360-Indoor and PANDORA dataset.

5. Conclusion

This paper first elaborates on the flaws of regression loss of current spherical detectors, i.e., independent optimization of parameters and inconsistency between metric and loss. Then, to address these issues, we design a simple but effective regression loss, named GLDL loss. Besides, we observe that the scale of the object in a spherical image varies greatly. The huge differences among objects from different categories make the sample selection strategy based on SphIoU challenging. Therefore, we propose GLDL-ATSS as a better training sample selection strategy, which can alleviate the drawback of IoU threshold-based strategy of scale-sample imbalance. Finally, extensive experiments on two spherical dataset show that the our method brings significant and consistent improvements with a number of state-of-the-art models.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62072438, U1936110, 61931008, U21B2024, 62071415), Strategic Priority Research Program of Chinese Academy of Sciences (XDA28040000, XDA28120000), Natural Science Foundation of Shandong Province (ZR2021MF094), Key R&D Plan of Shandong Province (2020CXGC010804), Central Leading Local Science and Technology Development Special Fund Project (YDZX2021122) and Science & Technology Specific Projects in Agricultural High-tech Industrial Demonstration Area of the Yellow River Delta (2022SZX11).

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End object detection with transformers. In *ECCV*, 2020. 1

[3] Shih-Han Chou, Cheng Sun, Wen-Yen Chang, Wan-Ting Hsu, Min Sun, and Jianlong Fu. 360-indoor: Towards learning real-world objects in 360 indoor equirectangular images. In *IEEE Winter Conference on Applications of Computer Vision*, pages 834–842, 2020. 3, 6

[4] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *European Conference on Computer Vision*, pages 525–541, 2018. 1, 3, 7, 8

[5] Feng Dai, Bin Chen, Hang Xu, Yike Ma, Xiaodong Li, Bailan Feng, Chenggang Yan, and Qiang Zhao. Unbiased iou for spherical image object detection. In *AAAI*, 2022. 1, 2, 3, 7, 8

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 8

[7] Y. Heshmat, B. Jones, X. Xiong, C. Neustaedter, A. Tang, B. E. Riecke, and L. Yang. Geocaching with a beam: Shared outdoor activities through a telepresence robot with 360 degree viewing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018. 1

[8] Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, and Min Sun. Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1396–1405, 2017. 1

[9] J. Huang, Z. Chen, Adobe Research, U. D. Ceylan, and U. Hailin. 6-dof vr videos with a single 360-camera. In *Vis. Res.*, 2017. 1

[10] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *ECCV*, 2020. 3

[11] Yeonkun Lee, Jaeseok Jeong, Jongseob Yun, Wonjune Cho, and Kuk-Jin Yoon. Spherephd: Applying cnns on a spherical polyhedron representation of 360 images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9173–9181, 2019. 3

[12] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 7

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 6

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37, 2016. 3

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1

[17] Wen-Chih Lo, Ching-Ling Fan, Jean Lee, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 360° video viewing dataset in head-mounted virtual reality. In *Proc. ACM Conf. on Multimedia Syst.*, 2017. 1

[18] Yoshiki Masuyama, Yoshiaki Bando, Kohei Yatabe, Yoko Sasaki, Masaki Onishi, and Yasuhiro Oikawa. Self-Supervised Neural Audio-Visual Sound Source Localization via Probabilistic Spatial Modeling. In *IROS*, 2020. 1

[19] Qi Ming, Zhiqiang Zhou, Lingjuan Miao, Hongwei Zhang, and Linhao Li. Dynamic anchor learning for arbitrary-oriented object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2355–2363, 2021. 3

[20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[21] F. Pearson. Map projections: Theory and applications. *Crc Press*, 1990. 3

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 3

[23] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 658–666, 2019. 2, 3

[24] Maximilian Speicher, Jingchen Cao, Ao Yu, Haihua Zhang, and Michael Nebeling. 360anywhere: Mobile ad-hoc collaboration in any environment using 360 video and augmented reality. *Proceedings of the ACM on Human-Computer Interaction*, 2018. 1

[25] Yuchuan Su, Dinesh Jayaraman, and Kristen Grauman. Pano2vid: automatic cinematography for watching 360° videos. In *ACCV*, 2016. 1

[26] Kuan-Hsun Wang and Shang-Hong Lai. Object detection in curved space for 360-degree camera. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3642–3646, 2019. 1, 3, 7, 8

[27] Hang Xu, Qiang Zhao, Yike Ma, Xiaodong Li, Peng Yuan, Bailan Feng, Chenggang Yan, and Feng Dai. Pandora: A panoramic detection dataset for object with orientation. In *ECCV*, 2022. 1, 2, 3, 6, 7, 8

[28] Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, and Rainer Stiefelhagen. Capturing omni-range context for omnidirectional segmentation. In *Proc. CVPR*, pages 1376–1386, 2021. 1

[29] Wenyan Yang, Yanlin Qian, Joni-Kristian Kämäräinen, Francesco Cricri, and Lixin Fan. Object detection in equirectangular panorama. In *International Conference on Pattern Recognition*, pages 2190–2195, 2018. 1, 3

[30] Xue Yang, Yue Zhou, Gefan Zhang, Jirui Yang, Wentao Wang, Junchi Yan, Xiaopeng Zhang, and Qi Tian. The kfiou loss for rotated object detection. *arXiv preprint arXiv:2201.12558*, 2022. 2

[31] Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9308–9318, 2019. 1

[32] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 516–520, New York, NY, USA, 2016. Association for Computing Machinery. 2

[33] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520, 2016. 3

[34] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9756–9765, 2020. 1, 3, 5

[35] Xiaosong Zhang, Fang Wan, Chang Liu, Xiangyang Ji, and Qixiang Ye. Learning to match anchors for visual object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 3

[36] Pengyu Zhao, Ansheng You, Yuanxing Zhang, Jiaying Liu, Kaigui Bian, and Yunhai Tong. Spherical criteria for fast and accurate 360 object detection. In *AAAI*, pages 12959–12966, 2020. 3, 7, 8

[37] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *AAAI*, pages 12993–13000, 2020. 2, 3

[38] Zhaohui Zheng, Ping Wang, Dongwei Ren, Wei Liu, Rongguang Ye, Qinghua Hu, and Wangmeng Zuo. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. 2021. 3

[39] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv:1904.07850*, 2019. 1