DASC-SPT: Towards Self-Supervised Panoramic Semantic Segmentation

Tianlong Tan[1,3*], Bin Chen[1,2*], Hongliang Cao[1,2], Chenggang Yan[3,4], Yike Ma[1], Feng Dai[1†]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3]Shandong University, [4]Hangzhou Dianzi University

Abstract

Self-Supervised Semantic Segmentation, aiming to leverage masses of unlabeled data for boosting semantic segmentation, has been rapidly emerging as an active task in recent years. However, existing self-supervised semantic segmentation approaches mainly focus on planar images, leaving multiple distorted objects encountered in panoramic images unexplored due to the formidable challenge of handling heterogeneous degrees of distortions across different locations. In this paper, we propose a novel Self-Supervised Panoramic Semantic Segmentation model, termed DASC-SPT, built upon the mainstream contrastive learning framework. Towards distortions in panoramic images, we present two structures to better learn from distorted features by applying planar images. For the input images of self-supervision, we design a Spherical Projection Transformation (SPT) strategy that involves randomly projecting planar images onto various locations of the sphere to introduce the distortions. For pixel-wise distorted features, we construct a Deformation-aware Sampling Consistency (DASC) framework to further utilize the shared content and discrepancies caused by different distortions of paired views, where the deformation-aware consistency can be quantified on pixel-wise features. Both of the two components facilitate the model to adapt to distortions and boost panoramic semantic segmentation. Extensive comprehensive experiments on three panoramic datasets demonstrate the effectiveness and superiority of DASC-SPT approach.

1. Introduction

For fully-supervised semantic segmentation, sufficient labeled data is essential. Nevertheless, the process of labeling data is both time-consuming and costly. In recent years, self-supervised learning has attracted increasing attention due to it learns from the easy-to-obtain unlabeled data, and attains promising results in semantic segmentation task. Existing methods [24,27,33,34,37–39,43] mainly focus on representation understanding with undistorted objects in planar images. However, in more complex image representation, such as panoramic images , which usually have high resolutions and objects in images undergo geometric shape changes, resulting in the higher annotation cost. Therefore, self-supervised panoramic semantic segmentation is worth studying.

Compared with planar images, the primary characteristic of panoramic images are distortions, which are introduced by a transformation when mapping from a spherical

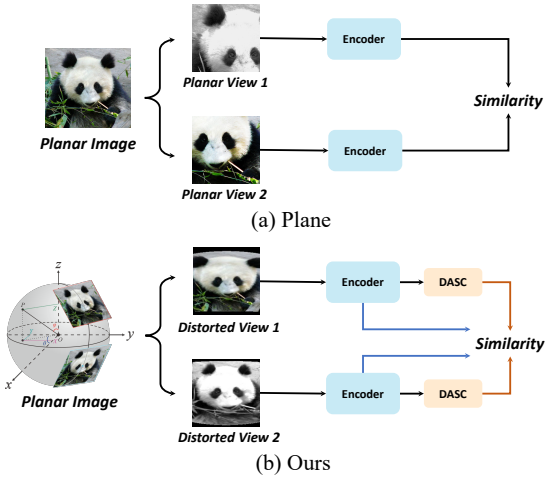


Figure 1. Comparison of two types of self-supervised semantic segmentation methods. (a): The plane-based methods take planar images as input and generate two undistorted views using random crop, where two encoders are added to compute the global similarity between the two paired views. (b): Ours takes planar images as the input and generates two distorted views using the SPT strategy. where the DASC framework is also applied to compute the similarity behind the encoder besides the existing two global branches.

* These authors contribute equally.
† The corresponding author. E-mail: fdai@ict.ac.cn

representation to the commonly used equirectangular projection (ERP) representation in panoramic datasets. The mainstream self-supervised semantic segmentation methods with CNNs on planar images [2, 3, 13–15, 41, 45] are based on the contrastive learning framework, as shown in Figure 1(a). By introducing data transformations, the main idea of the framework is to maximize the similarity between two augmented views from the same image and minimize the similarity between views from different images. This enables the model to learn more generalized features across instances or features that are more robust for specific attributes, such as the rotation and mask prediction. To extend the framework to panoramic semantic segmentation, we believe two key aspects need to be addressed: 1) Distortions are an inherent property of objects in panoramic images. How can we introduce different distortions to the same object to get two views in the self-supervised learning framework. 2) As the backbone network is a complex nonlinear function, we can modify the framework during pre-training to facilitate the backbone in learning panoramic distortions as effectively as possible, thereby mitigating the challenges of segmenting distorted objects in semantic segmentation.

To the best of our knowledge, this paper is the first work to propose the Self-Supervised Panoramic Semantic Segmentation approach focusing on panoramic distortions, termed DASC-SPT. Following SimSiam [4], DASC-SPT is built upon the siamese framework, where one of the views from the same image is input into the prediction head to predict the other view. The two key designs are an transformation method using planar images and a distortions-robust framework at the dense pixel level.

Our methods are shown in Figure 1(b). To be specific, considering that acquiring views of the same object with different distortions in panoramic images is challenging, we propose a Spherical Projection Transformation (SPT) strategy. SPT enables the mapping of planar images to different latitudes of a sphere, generating views with varying degrees of panoramic distortions. In this manner, we can guide the learning of backbone by maximizing the similarity between two different panoramic distorted views from the same image. Furthermore, SPT has the potential to introduce more distorted objects based on large datasets of planar images. In addition, considering that pixels in different views are misaligned after using SPT, we propose a Deformation-aware Sampling Consistency (DASC) framework. DASC predicts offsets for each pixel from one view, and pixels displace with offsets based on their own positions to align with the corresponding pixel from the other view. This approach aims to achieve pixel-wise deformation-aware consistency, enabling the learning of more fine-grained representations for semantic segmentation and improving robustness of backbone to distorted perspectives.

Extensive experiments demonstrate that our DASC-SPT approach achieves significantly improved performance compared with existing methods on three panoramic semantic segmentation datasets, including the Stanford2D3DS [1], SUN360 [12], and CVPG [26] datasets. Ablation experiments also verify the effectiveness of SPT and DASC.

In summary, this paper proposes SPT and DASC by analyzing distortions, making an early exploration of self-supervised learning for panoramic semantic segmentation. We hope that this work can provide a good starting point and serve as an effective baseline for future research.

2. Related Work

2.1. Self-Supervised Semantic Segmentation

In recent years, self-supervised learning has achieved impressive performance in image classification and segmentation. Among these methods, contrastive-based methods [3, 4, 14, 15, 25, 31, 35, 41] train models by minimizing the distance between similar object features and maximizing the distance between dissimilar object features, thereby learning powerful representation capabilities. SimSiam [4] maximizes the similarity of positive pairs without using negative sample pairs and large batches. Nonetheless, these methods merely focus on instance discrimination tasks from a global perspective. For semantic segmentation, which require pixel-level classification, emphasizing pixel-level learning can lead to more effective improvements [2, 24, 27, 33, 34, 37–39, 43]. DenseCL [33] implements dense contrastive learning by optimizing a pairwise contrastive (dis)similarity loss at the pixel level between two views of input images. VICRegL [2] learns pixelwise representations by forcing local features to remain constant over different viewing conditions, exploring the fundamental trade-off between learning local and global features. In the field of panoramic images, PPS [18] directly applies PixPro [39] to panoramic image semantic segmentation. 360VAM [9] maximizes the mutual information of different views from the equator and poles of panoramic images. However, none of these works are designed for distortions in panoramic semantic segmentation. This paper aims to fill this blank and offer a starting point for future research.

2.2. Panoramic Semantic Segmentation

Different from planar semantic segmentation, panoramic semantic segmentation focuses on solving the problem of panoramic distortions. Because applying planar segmentation methods directly to panoramic images will not achieve the original expected results due to the existence of distortions. To address this, recently some methods [5, 32] convert panoramic images into the cube representation with lower distortions levels to tackle cube boundary discontinuity. Due to the increased uniformity and discretization of the icosahedron compared to the cube, many methods

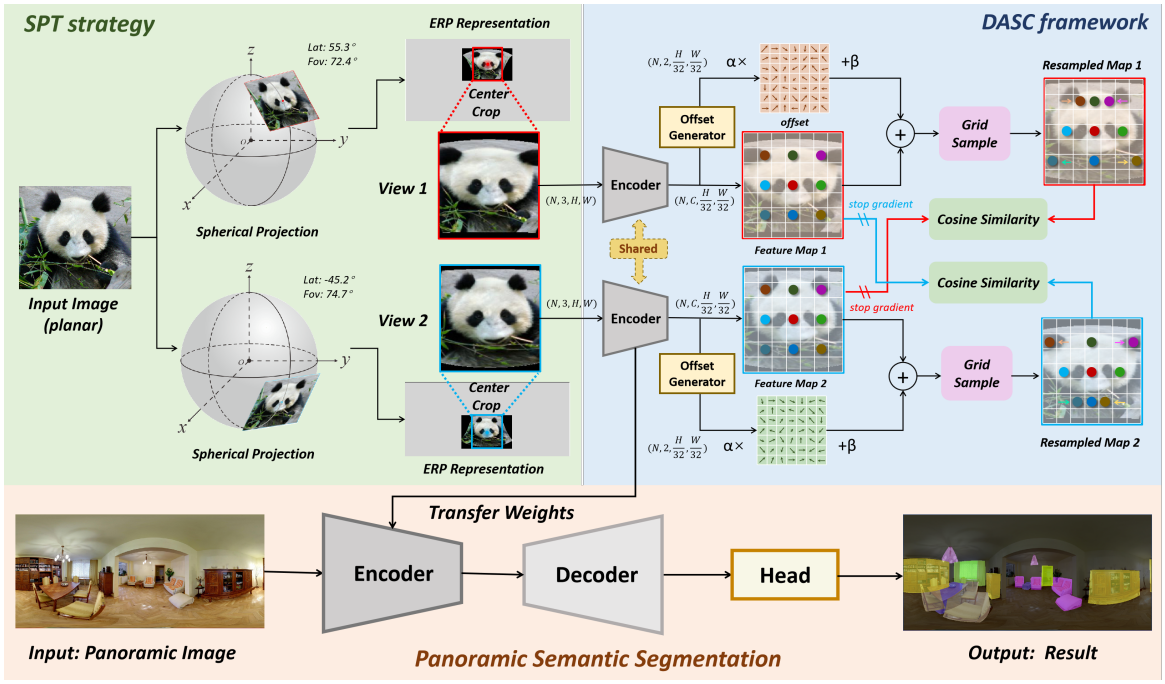


Figure 2. The overview of our DASC-SPT approach, which consists of Self-Supervised module, including the SPT strategy and the DASC framework as the initial upstream self-supervision, and the eventual Panoramic Semantic Segmentation task as the final downstream task.

[11, 19, 22, 42] design convolution and pooling operations on its representation to simulate planar operations. Efforts have been made to use graph neural networks for feature extraction on the sphere, aiming for more flexible connections [20,21,40].Other methods [8,28–30,44] propose the spherical convolutions,which aim to extract features varying in the size on different latitudes. Different from the above works that focus on the supervised paradigm, this paper makes an early exploration of self-supervised panoramic semantic segmentation, which can boost the backbone robust to distortions and reduce the annotation cost.

3. Method

3.1. Preliminary

Panoramic coordinate representations. The ideal representation of panoramic images is an undistorted and complete unit sphere. However, for display and storage purposes, datasets often use an equirectangular representation, and the pixel coordinates between these two representations can be converted correspondingly. Similar to planar images, the equirectangular representation E uses pixel coordinate system to express image pixels $E(x, y)$. We use S to represent the unit sphere, and the geographical location of any point on the sphere can be represented by latitude (*Lat*) $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and longitude (*Lon*) $\theta \in [-\pi, \pi]$, which are also known as the polar angle and azimuthal angle. The equator and poles of the sphere are commonly used to refer to undistorted and heavily distorted regions in panoramic images, respectively. There is a one-to-one correspondence between the $E(x, y)$ and the spherical surface $S(\theta, \phi)$, which can be expressed as $\theta = \frac{2x\pi}{w}, \phi = \frac{y\pi}{h}$. where $x \in [0, w], y \in [0, h]$, w, h are the width and height of the panoramic image. After that each point $P(\theta, \phi)$ can be transformed into $P(X, Y, Z)$ in the 3D Cartesian coordinate system as shown in Figure 1(b), and is defined by

$$\begin{cases} X = \cos\phi\cos\theta \\ Y = \cos\phi\sin\theta \\ Z = \sin\phi \end{cases} \tag{1}$$

Through Equation 1, each pixel on the spherical surface can be applied to the generation of the equirectangular representation via coordinate transformation with sampling. In the following Section 3.2, the spherical projection in 3D space will be more introduced in detail.

Self-Supervised contrastive learning framework. The general framework of self-supervised contrastive learning is based on instance discrimination tasks. Here we take SimSiam [4] as an example to roughly introduce the workflow of self-supervised contrastive learning. Given an unlabeled dataset $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_n\}$, where each image x_i is augmented by two different data transformations T_1 and T_2 to generate two views $z_{i1} = T_1(\mathcal{I}_i), z_{i2} = T_2(\mathcal{I}_i)$. The views are encoded by the encoder $g : \mathcal{I} \to \mathbb{R}^D$ and projector $q : \mathbb{R}^D \to \mathbb{R}^C$, and mapped to a feature space. Next, Sim-

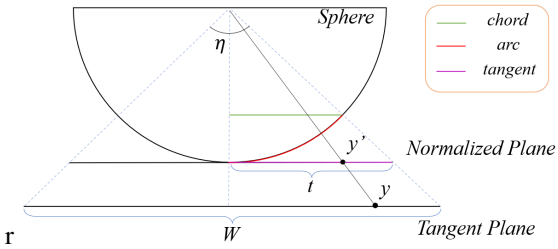


Figure 3. The diagram of different modes for the spherical projection transformation strategy.

Siam uses a predictor $p : \mathbb{R}^C \rightarrow \mathbb{R}^C$ to convert the projected representations to predicted representations. The projector and predictor heads are typically composed of small fully connected layers, ReLU activation functions, and BatchNorm layers. In the training stage, the loss of global instances discrimination L uses cosine similarity, which can be expressed as

$$L(z_{i1}, z_{i2}) = -\frac{p(q(g(z_{i1})))}{\|p(q(g(z_{i1})))\|_2} \cdot \frac{stop(q(g(z_{i2})))}{\|stop(q(g(z_{i2})))\|_2} \quad (2)$$

where $\|\cdot\|_2$ is the l_2-norm, $stop$ indicates the stop-gradient operation. After self-supervised training, the encoder can be transferred to downstream visual tasks, providing the better representation capability for other dense prediction models.

3.2. Spherical Projection Transformation Strategy

As mentioned above, capturing a range of perspectives for a single object while contending with various distortions in panoramas is an inherently complex undertaking, necessitating a comprehensive and nuanced approach. Consequently, we propose a Spherical Projection Transformation (SPT) strategy that generates controllable panoramic distortions from planar images for self-supervised panoramic semantic segmentation to learn distortions.

As shown in Figure 2, our proposed strategy is divided into two steps: the spherical projection and the center crop operation. The spherical projection involves aligning a planar image tangentially with the unit sphere S and rotating the sphere to align the projection point with the x-axis. Then, all points are projected onto the normalized plane. Based on Section 3.1, it is known that each point $\mathbf{P}$ on the sphere S can be represented by $\mathbf{P}(X, Y, Z)$, and the normalized form becomes $\mathbf{P}(1, \frac{Y}{X}, \frac{Z}{X})$, which is denoted as $\mathbf{P}(1, y', z')$. Next, we transform the planar image pixels to the normalized plane through the transformation of the spherical projection using Equation 3. The diagram of the transformation is shown in Figure 3 and the formula can be given by

$$y' = \frac{y}{W} \cdot 2t - t, t = \begin{cases} tan\frac{\eta}{2} & \text{if using } \textit{tangent} \text{ mode} \\ \sin\frac{\eta}{2} & \text{if using } \textit{chord} \text{ mode} \\ \frac{\eta}{2} & \text{if using } \textit{arc} \text{ mode} \end{cases} \quad (3)$$

where y' is coordinate of a pixel in the normalized plane, y represents the coordinate of the pixel in the tangent plane, W is width of the image, $2t$ is the projection limit range based on the set η. And z' dimension follows the same rule.

After that, we filter out the points outside the defined range based on Fov η and different modes, i.e., *tangent*, *chord* or *arc*. By doing this, the relationship is established of the transformation.

A planar image can be projected onto a sphere using spherical projection and transformed into an ERP representation through the above transformation. However, due to the narrower Fov, it contains a amount of irrelevant information. When using random crop, the views may include background regions. Striving to make such views as similar as possible ultimately affects model training. To address this, we further perform center crop on the ERP representation to remove most of the background information and generate views with different distortions as the input. Specifically, the projection tangent center point $e_P(\theta_P, \phi_P)$ can be transformed to the coordinates $e_P(x_P, y_P)$ on the ERP image E. We crop the image around the point e_P to obtain a rescaled clipped image $I(u, v)$, which is defined as

$$I(u, v) = E(i + u, j + v) \quad (4)$$

where $0 \leq u < w, 0 \leq v < h$, and $i = x_P - \frac{h}{2}, j = y_P - \frac{w}{2}$ are the horizontal and vertical coordinates of the top left corner of the clipped region respectively, h and w are the height and width of the rescaled clipped image. Subsequently, the rescaled clipped image is upsampled to the input image size (H, W) using a scale hyper-parameter s, where $H = s^2 \cdot h, W = s^2 \cdot w$. In consequence, the random crop can be replaced by our strategy for the generation of paired views with different distortions to make full use of current large-scale planar image datasets for subsequent self-supervised panoramic semantic segmentation.

3.3. Deformation-aware Sampling Consistency Framework

Since the presented SPT strategy has employed spherical projection and center crop on planar images, the model can possess the adaptability to object distortions in panoramic images to a certain degree. Moreover, on this basis, a deformation-aware sampling consistency framework named DASC is proposed to leverage the paired views with different distortions. The main purpose is to boost the feature representational capabilities of the model and improve its robustness to distorted perspectives by measuring the consistency between two distinct views after undergoing a deformable spatial transformation. So, the DASC not only learns the global consistency introduced in Sec.3.1 but also includes pixel-wise consistency learning for distortions.

As shown in Figure 2, the DASC framework regards the paired views z_{i1}, z_{i2} as the input of the shared encoder g,

and the output can be described as $f : \mathbb{R}^{C\times H\times W}$, where C, H, W are the channels, height and width of the feature map respectively. Furthermore, the projector q, the predictor p and offset generator $c : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^{2\times H\times W}$ are stacked to generate the trainable offset vector, where the channel 2 means the two directions (x, y) of the offset $\mathbf{o}_{ik} = [o_x, o_y]^T_{ik}$ from the k-th view in the i-th image, and $x \in [0, W], y \in [0, H]$, $k = 1$ or 2. Note that each component of the offset vector are constrained to a certain range $o_x \in [-1, 1], o_y \in [-1, 1]$, which represents the deformation to any location $\mathbf{l}_{mn} = [m, n]^T$ within the feature map, where $[m, n]^T = \left[\frac{2x}{W} - 1, \frac{2y}{H} - 1\right]^T \in [-1, 1]$ are the normalized coordinate of the location (x, y). Therefore, each location is transformed to be deformable when the predicted offsets are combined with the original location, and the specific derivation can be given by

$$\mathbf{l}^{'}_{ik}(m, n) = \mathbf{l}_{mn} + \hat{\mathbf{o}}_{ik}(m, n) \tag{5}$$

where $\mathbf{l}^{'}_{ik}(m, n)$ indicates the new location in the (m, n) coordinates from the k-th view in the i-th image after applying the transformed offset $\hat{\mathbf{o}}_{ik}$ on the identical location $\mathbf{l}_{mn}$. Meanwhile, the coordinates of the feature map are normalized to $[-1, 1]$, implying an ideal maximum offset of 2 in any direction. Thus we can apply the linear transformation to adjust the range of the offset, which is formulated as

$$\hat{\mathbf{o}}_{ik}(m, n) = \alpha \cdot \mathbf{o}_{ik}(m, n) + \beta \tag{6}$$

where $\mathbf{o}_{ik}(m, n)$ is the predicted offset of the network, α and β are respectively the scaling and translation hyperparameters, and $\hat{\mathbf{o}}_{ik}(m, n)$ refers to the offsets resulting from the transformation. Once the new deformable locations are computed, the pipeline will perform grid sample with these new coordinates. By doing this, the resampled feature map from one view can be deduced to quantify the consistency with that from the other view. We take $\mathbf{l}^{'}_{ik}(m, n)$ and encoded feature map f as the input of the grid sample G based on a differentiable bilinear interpolation.

By using the bilinear interpolation, the deformation-aware resampled feature map can be acquired to measure the pixel-wise consistency with the one from another view. We treat the cosine similarity in Equation 2 as the consistency quantification. The detailed theoretical analysis of pixel-wise consistency is in the supplementary materials.

Consequently, the complete self-supervised training objective for panoramic images is denoted as

$$\begin{aligned} L(z_{i1}, z_{i2}) = &\frac{1}{2}\left(L_g(z_{i1}, z_{i2}) + L_g(z_{i2}, z_{i1})\right) \\ &+ \frac{\lambda}{2}\left(L_d(z_{i1}, z_{i2}) + L_d(z_{i2}, z_{i1})\right) \end{aligned} \tag{7}$$

where both L_g and L_d are the cosine similarity losses. The former represents the global optimization and the latter indicates the deformation consistency quantification. Compared to the classical cosine similarity loss, the presence of gradient stopping strategies renders the loss asymmetric, necessitating the swapping of paired views to compute the loss twice and average them accordingly. This objective is extremely flexible and extensible, greatly enriching our proposed DASC approach for self-supervised panoramic semantic segmentation. Additionally, λ is the loss weight to balance the two components of Equation 7, which is set to 1.0 by default in our subsequent experiments.

Once self-supervised training is completed, the weights of pre-trained model will be transferred to the encoder of panoramic semantic segmentation followed by the decoder and the segmentation head. Due to the multi-class characteristics of our task, the objective for panoramic semantic segmentation L_{pss} adopts the fundamental Cross Entropy Loss, which can be written as

$$L_{pss} = -\sum_{i=1}^{N} p_i \log\left(p_i\right) \tag{8}$$

where p_i represents the predicted probability of each class and L_{pss} indicates the concrete loss for the main task. Note that this loss is identical to the one defined in the FCN [23] model, which is commonly employed as the foundational model for the panoramic semantic segmentation task, to provide fair experimental results.

4. Experimental Results

4.1. Datasets and Implementation Details

Datasets and Metrics. We conduct pretraining on ImageNet1000 dataset, which is a subset of ImageNet-1K containing all categories with about 34K images following existing self-supervised methods on panoramic images [9, 18]. Moreover, in accordance with previous studies [9, 18], we then perform the panoramic semantic segmentation task on three datasets below.

SUN360E [12]. The SUN360E dataset expands the SUN360 [36] dataset by adding segmentation labels to 666 panoramic images (418 bedroom and 248 living room) for semantic segmentation. Each image of 1024×512 resolution is annotated with 14 different categories.

Standford2D3DS [1]. The Stanford2D3DS dataset includes 1413 panoramic images for 2D semantic segmentation across 13 categories, with a resolution of 4096×2048.

CVPG-Pano [26]. The CVPG-Pano dataset is an outdoor panoramic image dataset consisting of 600 images with a resolution of 1664×832, where 524 images are used for training and the rest for testing. The dataset defines 20 categories and groups them into 7 major categories.

We employ mean accuracy (mAcc) and mean Intersection over Union (mIoU) as our evaluation metrics. While mAcc is used to evaluate the overall classification performance, and mIoU handles imbalanced class distributions. Both metrics are higher for better performance.

Table 1. Comparison with different approaches on three panoramic datasets. These approaches are listed as superivsed, planar, panoramic methods and ours. Experiments are evaluated by using mAcc and mIoU metrics.

Method	SUN360E		Standford2D3DS		CVPG-Pano	
	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU
Supervised	66.14	51.59	50.43	40.44	83.32	78.10
Plane						
MoCov2 [15]	69.19 (+3.05)	55.98 (+4.39)	50.46 (+0.03)	40.90 (+0.46)	83.75 (+0.43)	78.60 (+0.50)
SimCLR [3]	69.80 (+3.66)	56.19 (+4.60)	49.47 (–0.96)	40.21 (–0.23)	82.26 (–1.06)	76.61 (–1.49)
BYOL [14]	51.34 (–14.80)	38.46 (–13.13)	46.89 (–3.54)	37.63 (–2.81)	75.40 (–7.92)	70.44 (–7.66)
DenseCL [33]	70.12 (+3.98)	57.05 (+5.46)	51.72 (+1.29)	42.56 (+2.12)	83.88 (+0.56)	78.11 (+0.01)
Barlowtwins [41]	61.30 (–4.84)	46.91 (–4.68)	49.06 (–1.37)	39.53 (–0.91)	79.86 (–3.46)	74.86 (–3.24)
SimSiam [4]	70.37 (+4.23)	57.64 (+6.05)	51.37 (+0.94)	42.42 (+1.98)	85.18 (+1.86)	80.14 (+2.04)
VICRegL [2]	69.05 (+2.91)	54.32 (+2.73)	49.58 (–0.85)	40.96 (+0.52)	80.78 (–2.54)	76.03 (–2.07)
Panorama						
360VAM [9]	65.13 (–1.01)	51.26 (–0.33)	49.25 (–1.18)	40.16 (–0.28)	81.82 (–1.50)	76.17 (–1.93)
PPS [18]	71.35 (+5.21)	58.13 (+6.54)	52.03 (+1.60)	42.56 (+2.12)	83.41 (+0.09)	78.42 (+0.32)
Ours						
DASC-SPT	**73.02** (+6.88)	**60.76** (+9.17)	**52.25** (+1.82)	**42.95** (+2.51)	**86.29** (+2.97)	**80.74** (+2.64)

Implementation Details. We adopt the pretraining setting of SimSiam [4] for our self-supervised pretraining experiments. Specifically, we utilize SGD as our optimizer with initial learning rate of 0.05, and we set the weight decay and momentum to 1e-4 and 0.9. In our SPT, we set the projection latitude range [$\pm$30, $\pm$60], field of view [60, 80] and scale [0.5, 1]. In our DASC-SPT, we set α to 1 and β to 0. Each model is trained on 4 RTX 3090 GPUs with a batch size of 128. Other comparative methods are either implemented through MMSelfSup [7] or based on its source code. As with them, ResNet-50 [16] is used as our encoder. The Supervised method in Table 1 means the encoder is pre-trained on the ImageNet1000 with supervision. For the panoramic semantic segmentation, we follow the strategy of MMSegmentation [6] and fine-tune the FCN [23] with a batch size of 16. We randomly crop images to 512×1024 and train for 20k iterations to get the results.

4.2. Main Results

Quantitative Comparison. Given that DASC-SPT is built on SimSiam, our baseline naturally aligns with SimSiam [4]. For a fair comparison, we set three groups of approaches: 1) a supervised approach for reference, 2) self-supervised methods for planar images, 3) self-supervised methods for panoramic images. We conduct a comprehensive comparison between our DASC-SPT and existing methods of three groups on panoramic datasets. As shown in Table 1, our method outperforms the state-of-the-art methods on all three datasets, demonstraing the effectiveness of our method. For mAcc and mIoU on the SUN360E [12], our method is 1.67% on mAcc and 2.63% on mIoU higher than the second-best method PPS [18]. For the Standford2D3DS [1] and the CVPG-Pano [26] with less object distortions, our method can also achieve satisfactory performance compared with the arts, which highlights the competence of our DASC-SPT in handling the inherent distortion problem. To mitigate uncertainty, we run methods 5 times and record standard deviations in Table 2, showing the stability of our method with minimal fluctuations.

Table 2. Comparison of stability between ours and the baseline.

Methods	mAcc	mIoU
Baseline	70.37 $\pm$ 0.20	57.63 $\pm$ 0.33
Ours	**73.02** $\pm$ 0.27	**60.42** $\pm$ 0.41

Table 3. Ablation study on effectiveness of different components.

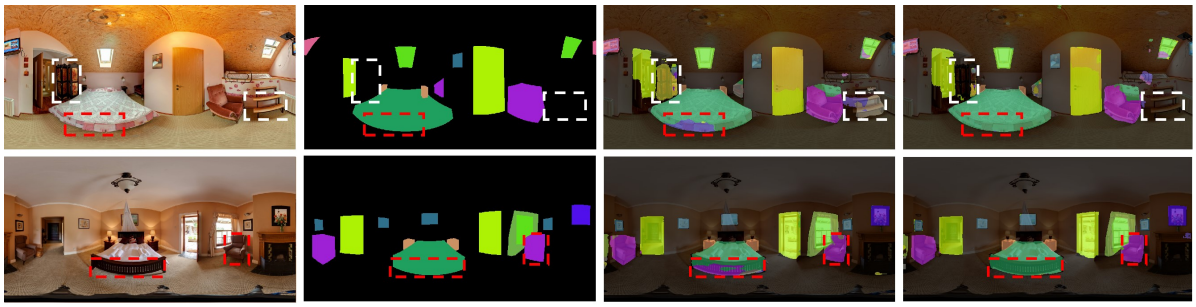
SPT	DASC	mAcc	mIoU
		70.37	57.64
✓		72.46 (+2.09)	59.56 (+1.92)
✓	✓	**73.02** (+2.65)	**60.76** (+3.12)

Qualitative Comparison. We conduct qualitative comparison between the baseline and our method on SUN360E dataset as an example in Figure 4. The baseline tends to produce incomplete masks due to serious distortions (e.g., the bed) as highlighted by the red dashed boxes. However, our method products more complete and precise semantic masks against the baseline. In addition to the distortion problem, there are some texture-closer semantic categories (e.g., the screen). It can be observed that the baseline products some inaccurate masks with similar textures as highlighted by the white dashed boxes. In contrast, our method shows decent performance in the face of such problem.

4.3. Ablation Study

In this section, we conduct extensive ablation studies on different components of DASC-SPT and their hyper-parameters on SUN360E dataset to verify the effectiveness.

Effects of proposed SPT and DASC. To validate the effect of the SPT strategy and the DASC framework, we evaluate two components and results are shown in Table



(a) Images (b) Ground Truth (c) Baseline (d) DASC-SPT

Figure 4. Qualitative Results between the baseline and our proposed DASC-SPT on the SUN360E dataset.

Table 4. Ablation study on comparison of different spherical projection schemes.

scheme	mAcc	mIoU
chord	**73.18**	60.26
arc	72.42	59.74
tangent	73.02	**60.76**

Table 5. Ablation study on comparison of different Scale intervals in the SPT strategy.

Scale	mAcc	mIoU
[0.1, 1.0]	72.61	60.29
[0.5, 1.0]	**73.02**	**60.76**
[0.1, 0.5]	71.48	58.77
[0.5, 1.2]	71.92	59.39

Table 6. Ablation study on comparison of different Fov intervals in the SPT strategy.

Fov	mAcc	mIoU
[50, 80]	**73.48**	60.62
[60, 80]	73.02	**60.76**
[60, 90]	71.89	58.94
[50, 90]	71.27	59.11

Table 7. Ablation study on paradigm of offset convolution settings.

Offset Conv	mAcc	mIoU
shared	71.86	59.87
separated	**73.02**	**60.76**

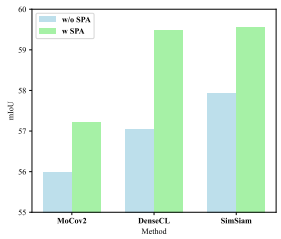


Figure 5. Histogram of comparison of different methods w or w/o our proposed SPT strategy.

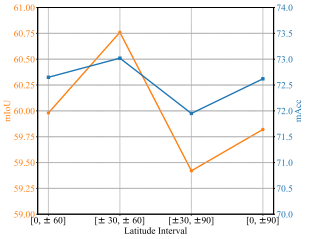


Figure 6. Evaluation on comparison of the different latitude range of the SPT strategy.

Table 8. Ablation study on comparison of DASC losses.

Loss	mAcc	mIoU
MSE	72.67	60.06
CE	70.36	57.72
Cosine	**73.02**	**60.76**

Table 9. Ablation study on different paradigms of offset mapping w or w/o the original location $\mathbf{I}_{mn}$.

Identity	mAcc	mIoU
zero (w/o $\mathbf{I}_{mn}$)	72.51	60.16
zero (w $\mathbf{I}_{mn}$)	**73.02**	**60.76**
border (w/o $\mathbf{I}_{mn}$)	71.71	59.65
border (w $\mathbf{I}_{mn}$)	72.46	59.88

Table 10. Ablation study on comparison of deformation hyper-parameters.

Index	α	β	mAcc	mIoU
(a)	1	0	73.02	**60.76**
(b)	$\frac{1}{2}$	$\frac{\lvert x\rvert+\lvert y\rvert}{4}$	**73.05**	60.60
(c)	$\sqrt{x^2+y^2}+\frac{1}{2}$	0	72.42	59.86
(d)	$\frac{1}{2}$	$-\frac{x}{2}/-\frac{y}{2}$	71.27	59.12

3.We can observe that the SPT component can gain 2.09 mAcc and 1.92 mIoU improvements compared to the baseline. When the DASC and the SPT are combined, the performance is significantly boosted to 73.02 mAcc and 60.76 mIoU, demonstrating the superiority of each component.

Ablation study on the SPT strategy. As illustrated in Section 3.2, there are three main schemes of the SPT strategy called chord, arc and tangent. To compare the effect among them, we give the ablation study shown in Table 4. The spherical projection by tangent has reached 73.02 mAcc and 60.76 mIoU, which is comparable to chord and surpasses arc with a large margin. This may be because the tangent projection closely approximates the true distortions and provides greater information content compared to the corresponding arc, particularly for large projection angles. We also attempt to verify that the SPT strategy can bring gains among different self-supervised learning methods for panoramic images. Figure 5 is provided to present the performance with or without the SPT strategy. It obviously attains +1.24, +2.43, +1.92 mIoU gains for Mocov2, DenseCL and Simsiam, which illustrates that the SPT strategy is beneficial when applied to general methods.

Ablation study on hyper-parameters of SPT. For the SPT strategy, there are three hyper-parameters, i.e., *Lat*, *Fov* and *Scale*, where *Lat* denotes the latitude range that dominates the extent of distortions, and *Fov* and *Scale* determines the size of the projected view and the resolution of the region by center crop. The ablation study on *Lat* is shown in

Figure 6, where the results show that $[\pm 30, \pm 60]$ *Lat* setting yields higher performance on SUN360E dataset, which implies it necessary to constrain the latitude within a certain range for a better model. In addition, the ablation study on $Scale$ in Table 5 shows that using a scale range that is either too small or too large for center crop yields poor results. On one hand, a scale range like $[0.1, 0.5]$ causes the distorted view to contain insufficient information from the image. On the other hand, the result of the scale range $[0.5, 1.2]$ in the distorted view includes excessive background regions, which interferes with training. Compared to these conditions, setting the scale range to $[0.5, 1]$ achieves higher performance. Besides, the ablation study on Fov in Table 6 indicates that the [60, 80] Fov setting yields a higher mIoU, while the [50, 80] Fov setting achieves a higher mAcc. However, when the Fov range is expanded to 90, the performance significantly decreases. This is because increasing the Fov from 80 to 90 results in the cropped image not containing sufficient information from the original image with more invalid features that leads to the degration.

Ablation study on the DASC framework. We conduct ablation studies to delve into different parts of our DASC framework. First, the ablation study is conducted to different types of the offset convolution. In Table 7, the offset prediction with the separated is superior to the shared, which is because the separated one allows independent optimization for the paired views with discrepancies of distortions. Second, the loss of the DASC framework is substituted by MSE loss or CE loss and the results are shown in Table 8, which demonstrates that the Cosine Similarity Loss is still indispensable and surpasses other two losses by 0.70 and 3.04 mIoU. At last, Table 9 compares alternative paradigms of offset mapping in grid sample. We employ two sampling modes named *zero* and *border*, where *zero* sets pixels outside the boundary to zero while *border* takes the pixel values at the boundary. We also carry out the ablation study on whether to perform offset in the original location, i.e., w or w/o $\mathbf{I}_{mn}$. The results show that using *zero* mode in the original location achieves better, as it imposes constraints on the offset to facilitate learning more meaningful knowledge.

Ablation study on hyper-parameters of DASC. As mentioned in Section 3.3, we can affect the offset range by manipulating α and β. The different settings of the hyper-parameters is given in Table 10. From results we can conclude that adopting a larger deformation range for the center region and a smaller range for edges can yield better results to self-supervised panoramic semantic segmentation, as shown in (a) and (b) in contrast to settings in (c) and (d).

Different quantities of data. Focusing on the planar images for pretraining in panoramic images, we naturally wonder the generalization ability of our method in such expansion and then explore by further manipulating the pretrained dataset. Thus we conduct the evaluation by setting different

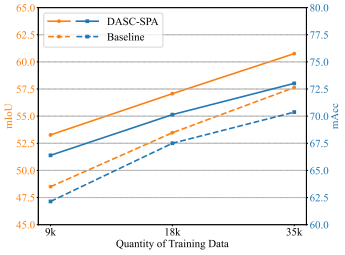


Figure 7. Evaluation on comparison of different quantities of the training data.

Table 11. Ablation study on the effectiveness of applying our approach to CMAE.

Method	mAcc	mIoU
CMAE	54.81	40.66
CMAE + Ours	**55.85** (+1.04)	**41.92** (+1.26)

quantities of data between the baseline and our DASC-SPT, i.e., 9k, 18k and 35k. As shown in Figure 7, our method comprehensively exceeds the baseline at all quantities for both mAcc and mIoU metrics, revealing the scalability and robustness of our method, which has the potential to effectively mitigate the impact of distortions on panoramic semantic segmentation in various practical applications.

Ablation study on CMAE. Since the backbone of our experiments is based on CNNs, in the era where transformer methods prevail, we are also very curious whether our idea can be applied to the recent MIM-based methods. We apply the SPT strategy and the DASC framework to the CMAE [17] method. For DASC, we improve the way of predicting offsets by adding a prediction head to predict soft labels for each patch after the feature decoder. Each patch is then linearly combined with its soft label to induce offset, aligning with the features from the Momentum Encoder. Finally, we optimize the L1 distance between the two features. The experimental results are shown in Table 11, which shows that our proposed method brings significant improvements to the CMAE. However, we notice that the overall metrics of the method are not high enough. Based on results from the ViT [10], we reckon that the issue may be related to the dataset size not reaching the millions level.

5. Conclusion

In this work, we propose a practicable and effective solution for self-supervised panoramic semantic segmentation. To be specific, a Spherical Projection Transformation (SPT) strategy is introduced by randomly projecting planar images onto various locations of the panoramic sphere with center crop to enhance the learning of representations through contrastive learning. Besides, a Deformation-aware Sampling Consistency (DASC) framework is constructed to quantify the consistency on the paired views with different distortions for training an adaptive and scalable model. With these components, our DASC-SPT achieves the state-of-the-art performance for self-supervised panoramic semantic segmentation. In the future, we hope that our proposed approach will be further applied in various fields.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *ArXiv preprint arXiv:1702.01105*, 2017. 2, 5, 6

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. VICRegl: Self-supervised learning of local visual features. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2, 6

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 2, 6

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2, 3, 6

[5] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun. Cube padding for weakly-supervised saliency prediction in 360 videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1420–1429, 2018. 2

[6] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 6

[7] MMSelfSup Contributors. MMSelfSup: Openmmlab self-supervised learning toolbox and benchmark. https://github.com/open-mmlab/mmselfsup, 2021. 6

[8] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–533, 2018. 3

[9] Yasser Abdelaziz Dahou Djilali, Tarun Krishna, Kevin McGuinness, and Noel E O'Connor. Rethinking 360deg image visual attention modelling with unsupervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15414–15424, 2021. 2, 5, 6

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8

[11] Marc Eder and Jan-Michael Frahm. Convolutions on spherical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–5, 2019. 3

[12] Clara Fernandez-Labrador, Jose M Facil, Alejandro Perez-Yus, Cédric Demonceaux, Javier Civera, and Jose J Guerrero. Corners for layout: End-to-end layout recovery from 360 images. *IEEE Robotics and Automation Letters*, 5(2):1255–1262, 2020. 2, 5, 6

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *ArXiv preprint arXiv:1803.07728*, 2018. 2

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 2, 6

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 6

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

[17] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 8

[18] Alexander Jaus, Kailun Yang, and Rainer Stiefelhagen. Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised contrastive learning. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 1421–1427. IEEE, 2021. 2, 5, 6

[19] Chiyu Jiang, Jingwei Huang, Karthik Kashinath, Philip Marcus, Matthias Niessner, et al. Spherical cnns on unstructured grids. *ArXiv preprint arXiv:1901.02039*, 2019. 3

[20] Renata Khasanova and Pascal Frossard. Graph-based classification of omnidirectional images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 869–878, 2017. 3

[21] Renata Khasanova and Pascal Frossard. Geometry aware convolutional filters for omnidirectional images representation. In *International Conference on Machine Learning*, pages 3351–3359. PMLR, 2019. 3

[22] Yeonkun Lee, Jaeseok Jeong, Jongseob Yun, Wonjune Cho, and Kuk-Jin Yoon. Spherephd: Applying cnns on a spherical polyhedron representation of 360deg images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9181–9189, 2019. 3

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 5, 6

[24] Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems*, 33:4489–4500, 2020. 1, 2

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv preprint arXiv:1807.03748*, 2018. 2

[26] Semih Orhan and Yalin Bastanlar. Semantic segmentation of outdoor panoramic images. *Signal, Image and Video Processing*, 16(3):643–650, 2022. 2, 5, 6

[27] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1144–1153, 2021. 1, 2

[28] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360 imagery. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[29] Yu-Chuan Su and Kristen Grauman. Kernel transformer networks for compact spherical convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9442–9451, 2019. 3

[30] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 707–722, 2018. 3

[31] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 2

[32] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 462–471, 2020. 2

[33] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 1, 2, 6

[34] Zhaoqing Wang, Qiang Li, Guoxin Zhang, Pengfei Wan, Wen Zheng, Nannan Wang, Mingming Gong, and Tongliang Liu. Exploring set similarity for dense self-supervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16590–16599, 2022. 1, 2

[35] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2

[36] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702. IEEE, 2012. 5

[37] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10539–10548, 2021. 1, 2

[38] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021. 1, 2

[39] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 1, 2

[40] Qin Yang, Chenglin Li, Wenrui Dai, Junni Zou, Guo-Jun Qi, and Hongkai Xiong. Rotation equivariant graph convolutional network for spherical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4303–4312, 2020. 3

[41] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 2, 6

[42] Chao Zhang, Stephan Liwicki, William Smith, and Roberto Cipolla. Orientation-aware semantic segmentation on icosahedron spheres. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3533–3541, 2019. 3

[43] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. Dense siamese network for dense unsupervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 464–480. Springer, 2022. 1, 2

[44] Qiang Zhao, Chen Zhu, Feng Dai, Yike Ma, Guoqing Jin, and Yongdong Zhang. Distortion-aware cnns for spherical images. In *the 27th International Joint Conference on Artificial Intelligence*, pages 1198–1204, 2018. 3

[45] Yucheng Zhao, Guangting Wang, Chong Luo, Wenjun Zeng, and Zheng-Jun Zha. Self-supervised visual representations learning by contrastive mask prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10160–10169, 2021. 2