Complementary Bi-directional Feature Compression for Indoor 360° Semantic Segmentation with Self-distillation

Zishuo Zheng[1,2], Chunyu Lin[1,2]*, Lang Nie[1,2], Kang Liao[1,2], Zhijie Shen[1,2], Yao Zhao[1,2]
[1]Institute of Information Science, Beijing Jiaotong University, Beijing, China
[2]Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China
{zszheng, cylin, nielang, kang_liao, zhjshen, yzhao}@bjtu.edu.cn

Abstract

Semantic segmentation on 360° images is a vital component of scene understanding due to the rich surrounding information. Recently, horizontal representation-based approaches outperform projection-based solutions, because the distortions can be effectively removed by compressing the spherical data in the vertical direction. However, these methods ignore the distortion distribution prior and are limited to unbalanced receptive fields, e.g., the receptive fields are sufficient in the vertical direction and insufficient in the horizontal direction. Differently, a vertical representation compressed in another direction can offer implicit distortion prior and enlarge horizontal receptive fields. In this paper, we combine the two different representations and propose a novel 360° semantic segmentation solution from a complementary perspective. Our network comprises three modules: a feature extraction module, a bi-directional compression module, and an ensemble decoding module. First, we extract multi-scale features from a panorama. Then, a bi-directional compression module is designed to compress features into two complementary low-dimensional representations, which provide content perception and distortion prior. Furthermore, to facilitate the fusion of bi-directional features, we design a unique self distillation strategy in the ensemble decoding module to enhance the interaction of different features and further improve the performance. Experimental results show that our approach outperforms the state-of-the-art solutions on quantitative evaluations while displaying the best performance on visual appearance.

1. Introduction

Panoramic images captured by omnidirectional cameras can provide a wide field-of-view (FoV), making it more practical in many crucial scene perception tasks [9], [24], [30], [40]. As a fundamental topic in computer vision, semantic segmentation aims to assign each pixel in the image a category label and is critical for various applications such as pose estimation [26], autonomous vehicles [31], augmented reality [2]. Directly applying normal FoV semantic segmentation methods [16], [22], [23], [44] to 360° images is not satisfactory due to the significant distortions in panoramas (usually produced from equirectangular projection—ERP) and large mismatch of FoV between panoramas and normal Fov images.

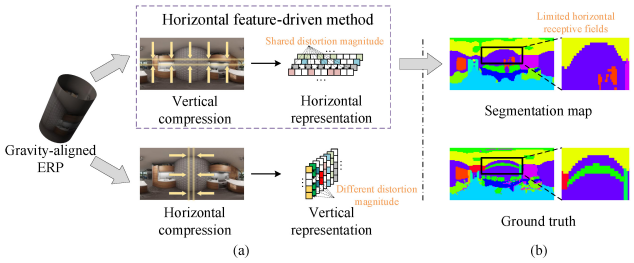


Figure 1. The motivation of the proposed 360° semantic segmentation approach: (a) The horizontal representation in each channel shares the same distortion magnitude while the vertical can perceive the distortion distribution. (b) The horizontal representation is limited by local receptive fields.

To overcome the above limitations, some researchers propose to adopting different projection formats (e.g., cubemap projection and icosahedron groups) [11], [41] or spherical convolutions [8], [12], [17] to decrease the negative effects of panoramic distortions. However, these methods sacrifice either accuracy or efficiency and fail to perceive precise panoramic structures.

Most recently, inspired by the geometric nature of gravity-aligned panoramas, some horizontal representation-driven methods [27], [33] are proposed to address the above problems. They squeeze the ERP image into 1D vector along the vertical direction, making it to be more content-focused, as shown in Fig.1a (top). Such a manner can be regarded as a process of shrinking spherical data towards the equator. Thus each element in the representation shares the same distortion magnitude and removes the negative effects

*Corresponding author

of panoramic distortions effectively. However, due to the fixed compression direction, it is inherently limited to local receptive fields which lack receptive capability in the horizontal direction (Fig.1b). Additionally, if there is no extra guidance during the decoding stage, it will lead to the lack of distortion distribution information in panoramic segmentation results, resulting in unsatisfactory performance.

Motivated by the horizontal features, we observe that compressing spherical data in the horizontal or vertical direction yields different data representations. Essentially, extracting the vertical feature along the horizontal direction is to contract panoramic images towards a certain meridian. Considering the data at the same latitude shares the same magnitude of distortion, this operation gathers data belonging to the same magnitude. Despite this representation may blur the image content, it makes feature distortions more prominent which can provide implicit distortion distribution prior. Compared with existing works that devote to eliminate the effects of distortions, we introduce implicit distortion information to guide the panoramic segmentation. Furthermore, the vertical representation also enhances the receptive capability in the horizontal direction.

In this paper, we present a novel neural network for 360° semantic segmentation consisting of three parts: a feature extraction module, a bi-directional compression module, and an ensemble decoding module. To be more specific, we first extract multi-scale features from an ERP image. Then our bi-directional compression module encodes the features into two complementary 1D flattened sequences. To achieve it, we design a Mix-MLP layer to yield a useful representation before we contract the dimensions. Subsequently, we propose a pyramid pooling compression (PPC) layer to perceive both distortion and content information by aggregating different sub-regions with different receptive fields. During the ensemble decoding process, we adopt *A-Conv* [27] to stretch dimensions and rebuild two different 2D features. Finally, the different features are fused for complementarity to predict the segmentation results.

However, considering the difference between two representations, the feature domains diverge severely, making them difficult to integrate harmoniously. Here, we address it by designing a unique self distillation strategy [43]. Specifically, We divide self distillation into three parallel ones: *horizontal-driven branch (HDB)*, *vertical-driven branch (VDB)*, and *ensembled branch (EB)*, of which *EB* is the fusion result of two representations. *HDB* and *VDB* are regarded as student models, while EB is the teacher model. The knowledge in the fusion portion will be shared with the separate portions. Finally, the retuned features in student models will feedback to the teacher model, which enhances the interaction of different features and further improves the performance.

With abundant experiments, we verify that the proposed solution can significantly outperform state-of-the-art algorithms in panoramic semantic segmentation and achieve great improvement on the quantitative evaluation. Besides, the ablation studies also reveal the effectiveness of our bi-directional representations and specially designed self distillation. In summary, our principle contributions are summarized as follows:

- To enlarge the limited horizontal receptive fields and offer implicit distortion prior, we combine horizontal and vertical representations to establish a novel 360° semantic segmentation network from a complementary perspective.
- To facilitate the fusion of bi-directional representations, we design a unique self distillation strategy to enhance the interaction of different feature and further improve the performance.
- Experiments demonstrate that our method significantly outperforms the current state-of-the-art approaches with great improvement on all metrics.

2. Related Work

2.1. Semantic Segmentation of Panoramic Images

Early approaches [4], [36] were based on the synthesized panoramic dataset or manually labeled samples. Motivated by the style transfer [46] and data distillation [28], Yang *et al.* proposed a framework [37], [38], [39] for reusing the models trained on perspective images by dividing the ERP into multiple restricted FoV sections for predictions. Although quite accurate, their strategy relies on labeled perspective image datasets with similar categories and scenes. Recent works find their solutions on the real-world datasets [1]. Tateno *et al.* [34] presented spherical convolution filters to make the network aware of the distortion from ERP. Compared to solutions that operate directly on ERP, [41] projected spherical signals into subdivided icosahedron mesh to mitigate distortion as well as improve prediction accuracy. Eder *et al.* [11] introduced tangent images, a novel representation that renders the image onto narrow FoV images tangent to a subdivided icosahedron. Sun *et al.* [33] used compressing method to encode latent features and use discrete cosine transform (DCT) to finish holistic scene modeling. Recently, driven by the self-attention mechanism, many transformer-based methods[6], [35], [45] emerge due to the powerful capability to aggregate long-range dependencies. Zhang *et al.* [29] and Shen *et al.* [42] use deformable components to eliminate image distortions and achieve state-of-the-art performance.

2.2. Horizontal Representation

Unlike the most existing methods that use pure 2D features to perform prediction, exploiting 1D horizontal rep-

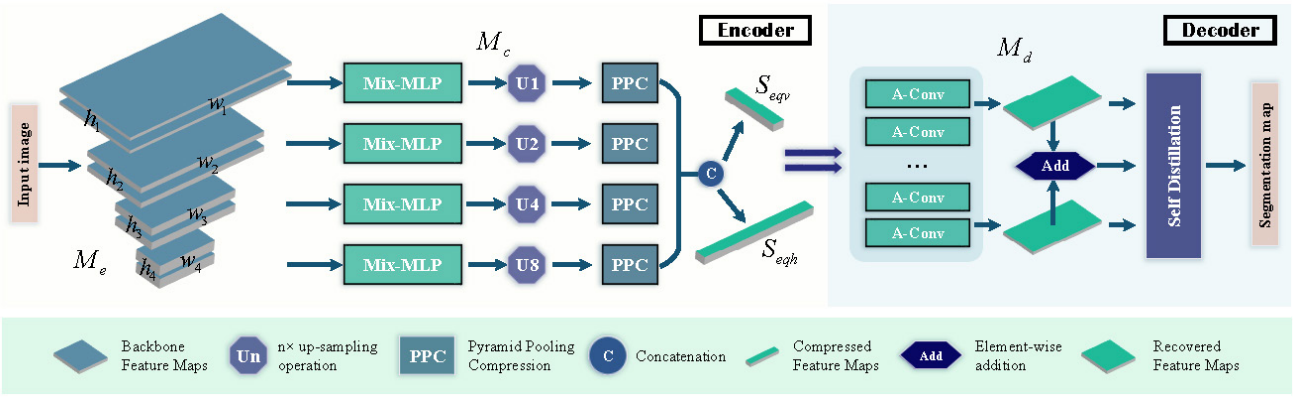


Figure 2. The architecture of the proposed network. This network consists of a feature extraction module M_e, a bi-directional compression module M_c and a ensemble decoding module M_d.

resentation can make the network learn the underlying geometric correlated knowledge. Su *et al.* [32] utilized different kernel sizes of the standard convolution to overcome the distortions. Particularly, the weight can only be shared along the horizontal. With the assumption that the horizontal dimension contains rich contextual information, Yang *et al.* [39] proposed a horizontal-driven attention method to capture omni-range priors in 360° images. Sun *et al.* [33] used 1D horizontal representation to design HorizonNet for the task of estimating room layout. This trend prompts a variety of works on scene understanding. For instance, Pintore *et al.* [27] proposed SliceNet and adopt Long Short-Term Memory (LSTM) to model the long-range dependencies for 360° depth estimation. However, these methods do not consider the latitudinal distortion property and horizontal respective capability, thus leading the accuracy degradation. Our solution solves this issue by integrating horizontal and vertical representations simultaneously which we believe is the optimal manner to eliminate the influence of distortions and preserve details. Moreover, adding extra tensors will not increase the model complexity and computational cost greatly, because our complexity is still $\mathcal{O}(N)$.

2.3. Knowledge Distillation

Knowledge distillation [15] is one of the most popular compression approaches. It is inspired by knowledge transfer from teachers to students. And it has shown its superiority in other domains such as data argumentation [3], adversarial attack [25], and model transfer [13]. However, it requires substantial efforts and experiments to build teacher models, and we will spend many datasets and long training time to refine student models. To overcome the setbacks of traditional distillation, Zhang *et al.*[43] proposed a novel training technique named self distillation, which means student and teacher models come from the same networks. Therefore, to facilitate the fusion of bi-directional representations, we redesign this technique to adapt to our framework. Specifically, we regard two 1D representations as students, while their fusion results as teacher, and introduce three loss functions for optimization. The well-designed self distillation can enhance the interaction of different feature and further improve the performance.

3. Approach

In this section, we describe the details of the proposed method for 360° semantic segmentation. We first show the overview of our framework. Then, the bi-directional compression module that reduces the dimensions of feature maps along horizontal and vertical directions is discussed. Finally, to foster the fusion of bi-directional representations and narrow the feature domain gap, we design a special self distillation strategy to adapt our network structure in the ensemble decoding module.

3.1. Network Overview

The framework is depicted in Fig. 2. In our feature extraction module, M_e, an ERP format 360° image with the size of $H \times W$ will be passed into a deep convolutional neural network, such as *ResNet* [14], to progressively decrease the panorama resolution and produce hierarchical feature maps at $\{1/4, 1/8, 1/16, 1/32\}$ of the original image resolution. Then we adopt feature pyramid network [19] to form multi-scale features, denoted as $\{F_i^{h_i \times w_i}\}_{i=1,2,3,4}$. In the next step, these feature maps are fed into the bi-directional compression module M_c in parallel which contains a lightweight Mix-MLP layer to yield useful representations and a PPC layer to contract the dimensions in the horizontal and vertical directions. In particular, we use different pooling operations for different feature maps to aggregate local and global context information. We detail this module in Sec.3.2. Then we concatenate multi-level 1D tensors in a single sequence and obtain two representations: S_{eqh} and S_{eqv}.

During the decoding period, an ensemble decoding mod-

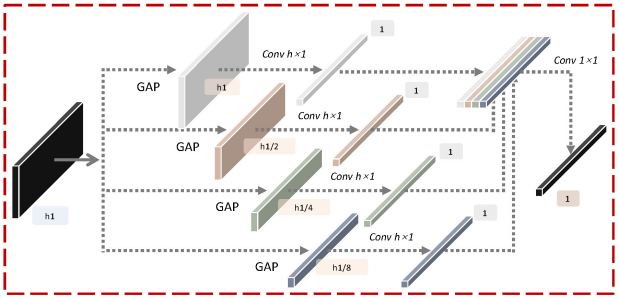


Figure 3. Illustration of pyramid pooling compression (PPC). Given a feature map F_1 with the height of h_1, to generate horizontal representation, we first use global average pooling (GAP) to harvest different sub-region representations, then a Conv2D layer is applied to compress the height to 1, followed by concatenation and convolutions layers to form the final 1D representation, which carries both local and global context information. Note that the sizes of sub-regions only vary in the vertical direction.

ule M_d is employed to reconstruct 2D dense feature maps in Sec.3.3. Finally, the fused features are fed into the segmentation head to predict the final segmentation results. Besides, for most padding operations, we use circular padding for the left-right boundaries of the feature maps.

3.2. Bi-directional Compression Module

To obtain the bi-directional spherical representations in a more effective and efficient way, we introduce a bi-directional compression module. This module compresses features into two complementary low-dimensional representations which provide content perception and distortion perception separately.

3.2.1 Mix-MLP Layer.

Considering the unique structure of ERP, we argue that location information is necessary for our 360° semantic segmentation. However, due to the consistent resolution requirements during training and testing, the positional encoding [10] suffers from the inflexible extension problem. To enable our network the capability of size-free positional encoding, we design a lightweight Mix-MLP layer. Inspired by [7], [35], we mixe a 3×3 convolution with zero padding and two MLP into a unified framework to introduce implicit location information. It can be formulated as:

$$F_{out} = Linear\left(\delta\left(DWC\left(Linear\left(F_{in}\right)\right)\right)\right) + F_{in} \quad (1)$$

where F_{in} is the feature maps from the backbone, and δ is activation function, we use GELU in our experiments. The number of channels in $Linear$ is four times as input. We exploit depth-wise convolutions (DWC) for improving efficiency and reducing the number of parameters. As a result, our backbone-extracted features with location information are useful for bi-directional feature compression.

3.2.2 Pyramid Pooling Compression.

To squeeze the height (h) and width (w), the most straightforward operation is to conduct two Conv2D layers with the kernel sizes of $h \times 1$ and $1 \times w$. However, although the receptive field of ResNet is already enough for the works that utilize the 2D feature, it is shown that it is still small for our method that only uses the 1D representations.

Having observed that the global average pooling is a helpful method as the global contextual prior [21], we design an efficient compression method to overcome the above problems. Moreover, it is more reasonable to introduce a more powerful representation using sub-regions with different sizes instead of the same size [44]. Hence, our PPC layer fuses feature under several different pyramid scales, Fig.3 gives an example. Note that the number of pyramid levels and size of each level can be modified manually. They are related to the size of the feature maps that are fed into the pooling layer. Therefore, combined with our hierarchical structure, the number of pyramid levels decreases with the increase of the network stage. Concretely, given the feature maps $\{F_i^{h_i \times w_i}\}_{i=1,2,3,4}$, the pooling size is $\{\frac{h_i}{2^j} \times w_i\}_{i=1,2,3,4;j=0,\ldots,4-i}$ for horizontal features, and $\{h_i \times \frac{w_i}{2^j}\}_{i=1,2,3,4;j=0,\ldots,5-i}$ for vertical features, respectively.

Besides, because the feature maps in different stages have different sizes, extra upsampling layers (see Fig.2) should be added to align them. For example, we only upsample the feature maps along the horizontal direction to align the horizontal feature. By compressing the features in different directions, our model can implicitly perceive content information and distortion distribution in a panorama from two perspectives.

3.3. Ensemble Decoding Module

To produce per-pixel predictions from 1D representations, Sun *et al.* [33] exploited interpolation operations and inverse discrete cosine transform (IDCT), leading to reduction of the model learnability. Different from the strategy of reshaping the size directly, we utilize n *A-Conv* layers [27], each of which includes an upsampling layer, a Conv2D layer, a BN layer, and a PReLU, to progressively stretch dimensions. Note that we replace the PReLU with ReLU for our segmentation modalities, and the final resolution is $\frac{H}{4} \times \frac{W}{4}$. In this way, we can obtain two distinct reconstructed tensors, denoted as D_h and D_v.

3.3.1 Self Distillation.

Despite we can directly ensemble these features to capture complementary semantic information in horizontal and vertical directions, the direction-dominated characteristics are still underutilized. This is because bi-directional features represent different panoramic characteristics, the apparent

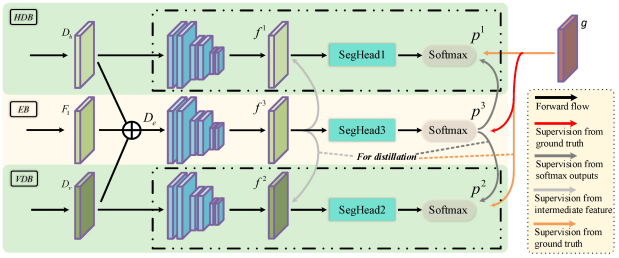


Figure 4. Illustration of our self distillation strategy. During the network's training process, we principally divide our self distillation structure into three sections according to their sources. The D_e is the summation of D_h, D_v, and F_1. In the test stage, these structures in rectangle boxes which only be introduced in training processes will be removed.

feature domain gap will hinder the fusion process. Furthermore, if we directly fuse them, the decoder will undertake a huge burden (responsible for fusion and ultimate classification), making the performance not satisfied.

To facilitate the fusion of different features, we do not design a more complex fusion network that can significantly enlarge the model size. On the contrary, we design a unique self distillation strategy in the decoding stage to narrow the domain gap between different features. As illustrated in Fig. 4, our self distillation structure can be divided into three parts: *HDB*, *VDB*, and *EB*. The *EB* comprises *HDB*, *VDB*, and the backbone feature $F_1^{h_1 \times w_1}$.

During the training process, *HDB* and *VDB* are regarded as student models while *EB* is the teacher model. The students can learn beneficial knowledge from the teacher, and the teacher can obtain good feedback from the students. In this manner, both students and teacher can benefit from each other. We set several convolution layers (bottleneck) and a SegHead as the segmentation part, and both students and teacher share the same network structure. The bottleneck contains three Conv2D layers with kernel sizes of 1×1, 3×3, and 1×1. The SegHead comprises two upsampling layers to reach the original resolution, and two Conv2D layers to predict the segmentation mask at a $H \times W \times N_{cls}$ resolution, where N_{cls} is the number of categories. Due to page limitations, we exhibit a complete example in the supplementary material.

After predicting the segmentation map of a 360° image, it is straightforward to adopt the segmentation labels to supervise the *HDB* and *VDB*, which can produce better D_h and D_v. Nevertheless, if we only use this supervision, the knowledge will not interact between the students and teacher. To this end, we introduce two extra supervisions (supervisions from intermediate features and final softmax outputs of the teacher model) to encourage the student models to learn from the teacher model. In brief, we use cross-entropy loss, kullback-Leibler(KL) divergence loss, and L2 loss as the optimization functions in our self distillation strategy.

3.4. Objective Function

Our objective function comprises three kinds of loss as the objective functions for optimizing predictions.

Cross-Entropy Loss. The first supervision is the cross-entropy loss. Almost all CNNs for this task exploit cross-entropy loss. It is computed with the ground truth (GT) from the training samples and the predictions of the softmax layer. We deploy it not only to the teacher's branch but also to two student branches. Through cross-entropy loss, the knowledge hidden in the training set is introduced directly from GT to all the branches. It can be written as:

$$L_{ce} = \langle -g \log(p^i) \rangle \tag{2}$$

where g, p^i denote the GT and predicted values from softmax layer's output, respectively; $i \in \{1, ..., N\}$, where N denotes the number of SegHeads in the training period, and $N = 3$ in our experiments (refer to Fig4). Moreover, class-wise weighted [41] is utilized to balance different classes.

KL Divergence Loss. The second supervision is the KL divergence loss. We use KL divergence to measure the difference between two distributions. It can be obtained through the computation of softmax outputs between students and teachers. Under the teacher's guidance, the distributions of students' SegHeads can approximate the teacher's, which indicates the supervision from distillation. It can be obtained by:

$$L_{kl} = \langle p^N \log(\frac{p^N}{p^i}) \rangle \tag{3}$$

L2 Loss. The last supervision is L_2 loss which works by decreasing the distance between feature maps in the student branches and the teacher branch. In this way, the knowledge in feature maps is distilled to students bottleneck layers.

$$L_{l2} = \| f^i - f^N \|_2^2 \tag{4}$$

Note that the last two losses for the teacher are zero, which means the supervision in the teacher model only comes from GT. Most importantly, we denote it as base loss L_b in our network without distillation. For students, we collect all supervision to obtain self distillation loss L_s. Meanwhile, to make the fusion process more interactive, we adopt three hyper-parameters to balance them.

$$\begin{aligned} L_{total} &= L_b + L_s, \\ &= \sum_{i=N} L_{ce} + \sum_{i=1}^{N-1} (\alpha * L_{ce} + \beta * L_{kl} + \gamma * L_{l2}). \end{aligned} \tag{5}$$

4. Experiments

We evaluate the effectiveness of our model in this section by carrying out comprehensive experiments on a real-world dataset. In the following subsections, we first introduce the dataset and implementation details, then report quantitative and qualitative results compared with the state-of-the-art approaches. Finally, we perform a series of ablation experiments for the proposed components.

4.1. Dataset

We evaluate our method on the Stanford 2D-3D-S dataset [1], which consists of 1413 real-world equirectangular RGB-D images over 13 categories. The dataset contains six large-scale indoor areas and provides semantic labels with the ERP format as annotations. Besides, the panoramas have a resolution of 2048 $\times$ 4096 and contain black void regions at the top and bottom. Following the prior works, we report averaged quantitative results from the 3-fold cross-validation splits.

4.2. Implementation Details

We conduct our experiments on three resolutions: 64 × 128, 256 × 512, and 1024 × 2048. We train our learning model using Adam [18] optimizer on a GTX 3090 GPU, and the batch sizes are set to 16, 8, and 2. For the low-resolution (the first two) inputs, we use the residual UNet-style architecture as backbone [8], [17], [41] and replace the specialized kernels with planar one. For the high-resolution (the last one) inputs, we adopt ResNet-101 pre-trained on COCO [20] as backbone [11], [33] to capture the larger receptive field. Inspired by [5], [16], we employ the poly learning rate policy where the base learning rate is multiplied by $(1 - \frac{iter}{max\_iter})^{power}$ with $power = 0.9$. The learning rate is set to 1×10^{-3} with $max\_iter = 300$ for low-resolution and 1×10^{-4} with $max\_iter = 60$ for high resolution. To prevent overfitting, we adopt a strategy of randomly cutting a patch of the input image and padding this region with a black mask, where the sizes of the hole are chosen from the set $\{20 \times 40, 80 \times 160, 320 \times 640\}$. In our loss function, we set $\alpha = 0.7$, $\beta = 0.3$, and $\gamma = 0.003$.

4.3. Results and Analysis

4.3.1 Compared with State-of-the-arts

In this subsection, we compare our method with the state-of-the-art methods on 360° semantic segmentation in both quantitative and qualitative evaluations, for which the numerical results or segmentation map on the same dataset is available. Furthermore, we analyze the model complexity to demonstrate that our method achieves a better efficiency tradeoff between model complexity and performance.

Table 1. Quantitative evaluation on Stanford2D3D dataset. Note that the results are averaged over the 3-folds. Reasons for different high resolutions, refer to Sec.4.3.

$H \times W$	Input	Method	Pub. & Year	mIoU ↑	mAcc ↑
Low-resolution input					
64×128	RGB-D	Gauge Net [8]	ICML'19	39.4	55.9
	RGB-D	UGSCNN [17]	ICLR'19	38.3	54.7
	RGB-D	HexRUNet [41]	ICCV'19	43.3	58.6
	RGB-D	SWSCNN [12]	NeruIPS'20	43.4	58.7
	RGB-D	TangentImg [11]	CVPR'20	37.5	50.2
	RGB-D	HoHoNet [33]	CVPR'21	40.8	52.1
	RGB-D	Ours	-	**47.2**	**61.2**
256×512	RGB-D	TangentImg [11]	CVPR'20	41.8	54.9
	RGB-D	HoHoNet [33]	CVPR'21	43.3	53.9
	RGB-D	PanoFormer [29]	ECCV'22	48.9	64.5
	RGB-D	Ours	-	**53.8**	**66.5**
High-resolution input					
512×1024	RGB	Trans4PASS [42]	CVPR'22	52.1	-
1024×2048	RGB	HoHoNet [33]	CVPR'21	52.0	65.0
2048×4096	RGB	TangentImg [11]	CVPR'20	45.6	65.2
512×1024	RGB	Ours	-	52.2	65.6
1024×2048	RGB	Ours	-	**52.4**	**65.9**
2048×4096	RGB-D	TangentImg [11]	CVPR'20	51.9	69.1
1024×2048	RGB-D	HoHoNet [33]	CVPR'21	56.3	68.9
1024×2048	RGB-D	Ours	-	**57.1**	**69.9**

Quantitative Evaluation: Table.1 shows the quantitative comparison results with the current state-of-the-art methods on different input resolutions. It is evident that our approach substantially outperforms the compared approaches in all metrics. From these evaluations on the lowest resolution, we can conclude that:

(i) Compared with the spherical CNNs methods [8], [12] [17] which aim to directly learn distortion-aware representation from the sphere, our approach avoids complex convolutional design on the transfer between planar and sphere, showing more promising generality and flexibility.

(ii) Compared with the distortion-tolerate approaches [11], [41], which project 360° images into icosahedron format, our approach only need ERP as input and omit the process of transformation. For example, our approach outperforms HexRUNet [41] which equip a specially non-rectangular kernel by a significant margin, with approximately 9% improvement on mIoU and 4% on mAcc.

(iii) As benefits of introducing vertical representation which provides guidance of distortion distribution and enlarges receptive fields in horizontal direction during the learning stage, our approach achieves a 17% improvement on both mIoU and mAcc on the lowest resolution compared with the [33] that only use horizontal representation.

To further demonstrate the generality of our method, we conduct experiments on other resolutions. It can be observed that our network has achieved satisfactory results on the 256 $\times$ 512 resolution with at least 24% improvements on mIoU and 23% on mAcc compared with CNN-based methods. Besides, we achieve 10% improvements on mIoU

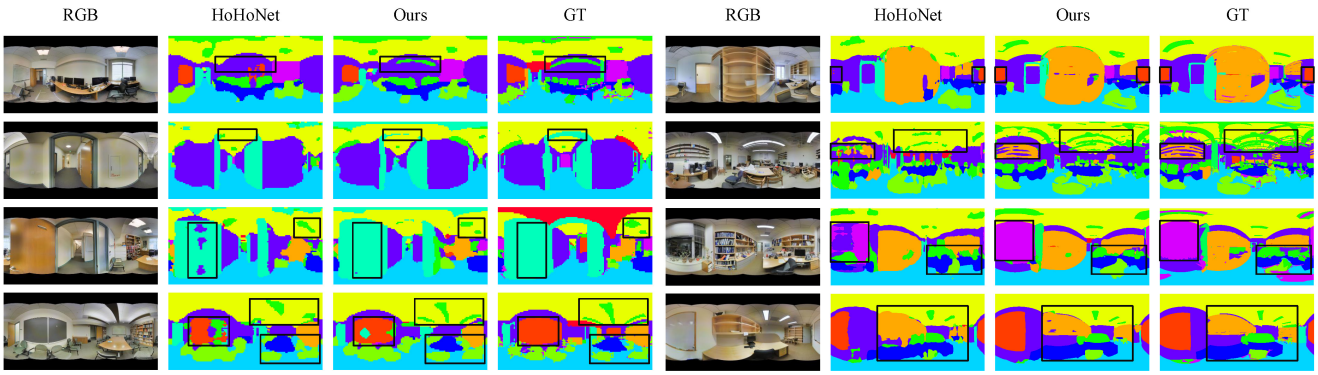


Figure 5. Qualitative evaluations of the segmentation results on 64 × 128 (left) and 256 × 512 (right) resolutions. Black rectangles are used to highlight difference.

Table 2. Number of parameters, FLOPs, and frame per second (FPS).

Method	FLOPs(G)	FPS	Params(M)
HoHoNet [33]	2.15	49	12.75
Ours	2.71	37	21.84

and 3% on mAcc compared with recent transformer-based methods which are good at aggregating long-range information. Unfortunately, due to the limitation of our device, we failed to train our network on higher resolutions (2048 × 4096) as [11]. Similar to [33] and [42], we can only train our network with ResNet-101 as the backbone on a lower 1024 × 2048 and 512 × 1024 resolutions with a small batch size and channel dimension. Empirically, these settings could decrease our performance, but we still achieve SOTA as reported in the table. Finally, we also argue that with the increase of resolution, the performance improvement decreases because we compress the image to 1D, making it challenging to recover 2D information. But compared with these methods, our architecture is more simpler and more efficient without the embedding or projection process.

Qualitative Evaluation: Fig.5 shows qualitative results on Stanford2D3DS dataset, compared to [33]. From the figure, we can observe that our approach performs well on all indoor scenes, while the horizontal representation method shows inferior segmentation results, especially in regions with distortions or regions with complex contextual information. With the complementary relationship between two representations, our method has a larger receptive field and sufficient distortion information. For example, the class with a strong distribution along horizontal direction while weaker along the vertical direction (see Fig.5 (left) first two rows) has an inferior performance in HoHoNet. Because these pixels occupy a small proportion in each column, they will be omitted when compressing the height dimension and are difficult to recover. In contrast, our vertical representation perceives this distribution in another dimension and supplement it to the decoding module. In general, our approach achieves a better performance from local details (receptive fields) to global distribution (distortion shape), which benefits from the designed modules. More qualitative results are included in the supplementary material.

Table 3. Ablation study with the key components on our 360° semantic segmentation approach. Experiment resolution: 64 × 128.

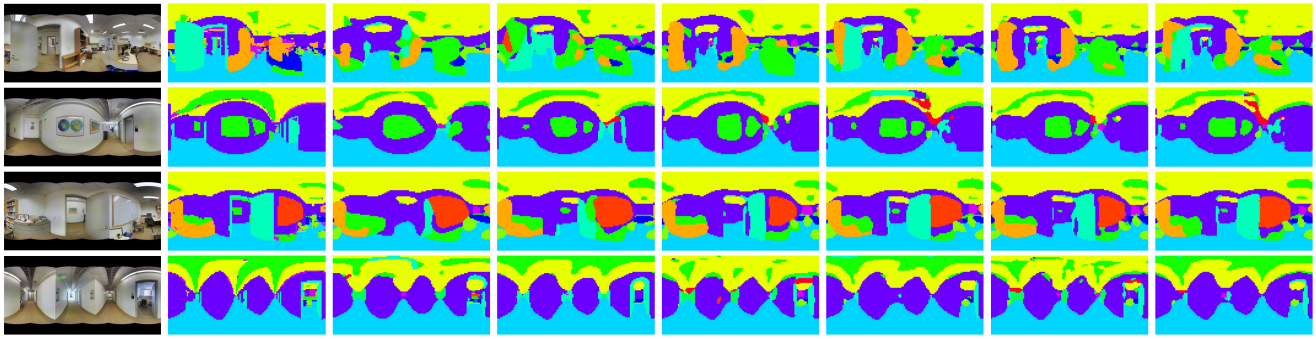
$\mathcal{H}$	$\mathcal{V}$	$\mathcal{M}$	$\mathcal{P}$	mIoU	mAcc
✓				41.50	53.27
	✓			40.63	52.65
✓	✓			42.76	55.00
✓	✓	✓		43.35	55.84
✓	✓	✓	✓	**44.71**	**57.03**

Model Complexity Analyses: We further show comparisons with [33] on three metrics to evaluate the model complexity and efficiency. The comparisons are conducted in one GTX 2080Ti GPU. As listed in the Table.2, our method offers acceptable higher computational complexity but achieves better performance.

4.3.2 Ablation Studies

To validate the effectiveness of different components in our approach, we conduct ablation studies and as illustrated in Table.3 and Table.4. Note that all the experiment results are evaluated on the lowest resolution input.

Effectiveness of bi-directional representations: We remove the self distillation to explore the effectiveness of combining two representations (horizontal ($\mathcal{H}$), vertical ($\mathcal{V}$)). Concretely, we implement our model without other key schemes (Mix-MLP ($\mathcal{M}$), PPC ($\mathcal{P}$)) and provide the quantitative results of variants equipped with different representations. From Table.3 (first three rows), we can observe that the joint representation performs better than only



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 6. Visual ablation comparison on our 360° semantic segmentation approach. (a) panoramic image. (b) ground truth. (c) *VDB w/o* self distillation. (d) *VDB w/* self distillation. (e) *HDB w/o* self distillation. (f) *HDB w/* self distillation. (g) *EB w/o* self distillation. (h) *EB w/* self distillation.

Table 4. Ablation study with the self distillation strategy. Resolution: 64 × 128.

Fold	*w/o* self distillation						*w/* self distillation					
	VDB		*HDB*		*EB*		*VDB*		*HDB*		*EB*	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
1	38.68	50.16	44.91	56.11	47.08	58.05	45.12	57.29	49.34	61.06	50.48	61.93
2	33.48	48.72	36.19	51.68	38.23	52.37	37.22	53.90	40.41	57.00	40.87	57.83
3	41.19	53.71	46.59	59.39	48.82	60.68	46.30	61.12	49.20	63.57	50.35	63.84

using one directional representation, which indicates that our network gains information from two complementary perspectives to facilitate the accuracy. In addition, we can observe that the $\mathcal{H}$ performs well than the $\mathcal{V}$, which proves that the vertical representation contains implicit distortion prior and blurs the content.

Effectiveness of components in M_c: Subsequently, we gradually add the removed components to show the different segmentation performances. Note that we utilize a Conv2D layer to compress without $\mathcal{M}$. As seen from Table.3 (last three rows), the mIoU is improved from 42.76 to 44.71 with a percentage gain of 4.6%, and the mAcc is boosted from 55.00 to 57.03 with the percentage gain of 3.7%. It also can be found that our network with useful position information derived from $\mathcal{M}$ achieves pleasing results. For the compression strategy, $\mathcal{P}$ provides large receptive fields and sufficient contextual information, making our model gains further improvements and outperforms a single Conv2D layer with 3.1% on mIoU and 2.1% on mAcc. Finally, the completed framework achieves the best results proving the effectiveness of our proposed components.

Effectiveness of self distillation: Since the different representations have a severe feature domain gap, it is difficult to integrate them harmoniously. Thus the self distillation plays the role of facilitating the fusion of bi-directional representations in our method. Furthermore, different from other knowledge distillation methods that pre-training a large teacher model, we exploit self distillation by directly dividing our network into student models (*VDB* and *HDB*) and teacher model (*EB*). To validate the effects of this strategy, we experiment with removing all supervision for student models, which means the knowledge from the teacher and dataset are obstructed. The quantitative results are shown in Table.4, we report detailed semantic segmentation results on three folds. From Table.4, we can conclude that via self distillation, all branches gain a significant improvement, which indicates that the well-designed training technique can foster the interaction of bi-directional representations and notably improve the segmentation performance. We also present the qualitative comparison results in Fig.6.

5. Conclusions

In this paper, a novel panoramic semantic segmentation network is presented from a complementary perspective by combining horizontal and vertical representations, which is capable of expanding the limited horizontal receptive fields and offering implicit distortion prior. To integrate complementary bi-directional representations, we design a unique self distillation strategy to enhance the interaction of different representations and make the predicted segmentation map more accurate. As the benefit of the proposed complementary representation, our approach significantly outperforms state-of-the-art solutions on the real-world dataset.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (Nos. 62172032, 62120106009).

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] Ronald T Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4):355–385, 1997.

[3] Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. Label refinery: Improving imagenet classification through label progression. *arXiv preprint arXiv:1805.02641*, 2018.

[4] Ignas Budvytis, Marvin Teichmann, Tomas Vojir, and Roberto Cipolla. Large scale joint semantic re-localisation and scene understanding via globally unique instance coordinate regression. *arXiv preprint arXiv:1909.10239*, 2019.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021.

[7] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021.

[8] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *International conference on Machine learning*, pages 1321–1330. PMLR, 2019.

[9] Greire Payen de La Garanderie, Amir Atapour Abarghouei, and Toby P Breckon. Eliminating the blind spot: Adapting 3d object detection and monocular depth estimation to 360 panoramic imagery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 789–807, 2018.

[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.

[11] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12426–12434, 2020.

[12] Carlos Esteves, Ameesh Makadia, and Kostas Daniilidis. Spin-weighted spherical cnns. *Advances in Neural Information Processing Systems*, 33:8614–8625, 2020.

[13] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

[14] 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.

[15] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[16] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019.

[17] Chiyu Max Jiang, Jingwei Huang, Karthik Kashinath, Philip Marcus, Matthias Niessner, et al. Spherical cnns on unstructured grids. In *International Conference on Learning Representations*, 2018.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[20] 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. Springer, 2014.

[21] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[22] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[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.

[24] Guangxiao Ma, Shuai Li, Chenglizhao Chen, Aimin Hao, and Hong Qin. Stage-wise salient object detection in 360 omnidirectional image via object-level semantical saliency ranking. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3535–3545, 2020.

[25] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016.

[26] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019.

[27] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11536–11545, 2021.

[28] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128, 2018.

[29] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360 {\deg} depth estimation. *arXiv preprint arXiv:2203.09283*, 2022.

[30] Zhijie Shen, Chunyu Lin, Lang Nie, Kang Liao, and Yao Zhao. Distortion-tolerant monocular depth estimation on omnidirectional images using dual-cubemap. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[31] Mennatullah Siam, Mostafa Gamal, Moemen Abdel-Razek, Senthil Yogamani, Martin Jagersand, and Hong Zhang. A comparative study of real-time semantic segmentation for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 587–597, 2018.

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

[33] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2573–2582, 2021.

[34] 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.

[35] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

[36] Yuanyou Xu, Kaiwei Wang, Kailun Yang, Dongming Sun, and Jia Fu. Semantic segmentation of panoramic images using a synthetic dataset. In *Artificial Intelligence and Machine Learning in Defense Applications*, volume 11169, page 111690B. International Society for Optics and Photonics, 2019.

[37] Kailun Yang, Xinxin Hu, Luis M Bergasa, Eduardo Romera, and Kaiwei Wang. Pass: Panoramic annular semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 21(10):4171–4185, 2019.

[38] Kailun Yang, Xinxin Hu, Hao Chen, Kaite Xiang, Kaiwei Wang, and Rainer Stiefelhagen. Ds-pass: Detail-sensitive panoramic annular semantic segmentation through swaftnet for surrounding sensing. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 457–464. IEEE, 2020.

[39] Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, and Rainer Stiefelhagen. Capturing omni-range context for omnidirectional segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1376–1386, 2021.

[40] Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, and Yinda Zhang. Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12632–12641, 2021.

[41] 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.

[42] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16917–16927, 2022.

[43] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019.

[44] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[45] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.