DASC-SPT: Towards Self-Supervised Panoramic Semantic Segmentation
Supplementary Material

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

1. Theoretical analysis on consistency measure of DASC

[1]Here we give the deriation of the consistency measure of DASC. Suppose that two original views from SPT are denoted as z_1 and z_2. And the objective of the local pixel-wise contrastive learning is to minimize the distance of two views before and after Encoder. Take z_1 as an example, which is written as

$$
\begin{aligned}
&arg\min_{\theta,o}\mathbb{E}_{p,\mathcal{T}}\left[\left\|\mathcal{T}_\theta\left(z_1(p)\right)-G\left(\mathcal{T}_\theta\left(z_2(p)\right);o\right)\right\|_2^2\right]\\
=&arg\min_{o}\mathbb{E}_{p}\left[\left\|z_1'(p)-G\left(z_2'(p+o)\right)\right\|_2^2\right]\\
=&arg\min_{o}\mathbb{E}_{p}\left[\left\|z_1'(p)-\sum\nolimits_{q\in S}G(q,p+o)\cdot z_2'(q)\right\|_2^2\right]\\
\triangleq&arg\min_{o}\mathbb{E}_{p}\left[\left\|z_1'(p)-z_2'\left(\Gamma(p+o)\right)\right\|_2^2\right]
\end{aligned}
$$

where $\mathcal{T}_\theta$ is Encoder, p is pixel index, o is offset, G is Grid Sampling, S indicates interpolation locations, Γ is round function, and $\triangleq$ suggests nearest-neighbor interpolation is used for simplified derivation. Then the loss is given by

$$
\begin{aligned}
L(z_1',z_2';o)&=\left\|\mathcal{N}\left(z_1'(p)\right)-\mathcal{N}\left(z_2'\left(\Gamma(p+o)\right)\right)\right\|_2^2\\
&=\left\|\frac{z_1'(p)}{\|z_1'(p)\|_2}-\frac{z_2'\left(\Gamma(p+o)\right)}{\|z_2'\left(\Gamma(p+o)\right)\|_2}\right\|_2^2\\
&=2-2\cdot\frac{z_1'(p)}{\|z_1'(p)\|_2}\cdot\frac{z_2'\left(\Gamma(p+o)\right)}{\|z_2'\left(\Gamma(p+o)\right)\|_2}
\end{aligned}
$$

where $\mathcal{N}$ represents the ℓ_2 normalization. Hence, the optimization aligns with the cosine consistency measure, assessing local semantic similarity and aligning pixel-wise locations. Meanwhile, the global loss measures high-level semantic features. And integrating both losses improves the model's semantic learning and adaptability to distortions.

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

2. More Ablation Studies

Different quantities of the labeled data in fine-tuning. We conduct an ablation study by halving the downstream task data, as shown in Table 1, which shows that our approach needs panoramic images for semantic segmentation to learn the differences between them and planar images. However, the mIoU of 51.53 (ours with half labeled data) vs. 51.59 (supervised with whole labeled data) shows that half labeled panoramic set in fine-tuning achieves comparable results.

Ablation on loss weight λ. Apart from the loss design, the loss weight λ is also needed to perform the ablation study. The results are shown in Table 2. We can observe that using smaller or larger weight will lead to the performance degradation of the proposed DASC-SPT model, and thus we choose $\lambda = 2$ as the best setting.

Ablation on center crop mode in SPT. We have also conducted the ablation study on the mode of center crop operation. The results are shown in Table 3, including the random crop operation, the center crop operation with double areas and the center crop operation we used in this paper. As shown in the Table, expanding the center crop area introduces uninformative background, while the random crop can create views lacking information, which both illustrates the decrease of the accuracy.

Comparison for training costs. We give the comparison for training costs of our approach and the baseline shown in Table 4, which demonstrates a slight increase in GFLOPs of our approach due to DASC incorporating consistency. Besides DASC, the projection calculation of our porposed SPT strategy also increases pretraining time. Overall, adding the extra training costs in our proposed approach is worthwhile because it can lead to greater performance gains.

Ablation on the frozen backbone in downstream task. The ablation study is conducted to compare using a frozen backbone versus a non-frozen backbone, and the results are

Table 1. Different Quantities of the Labeled data.

Method	Quantity	mAcc	mIoU
Supervised	half	**60.52**	44.61
Ours	half	66.27	**51.53**
Supervised	all	66.14	51.59
Ours	all	**73.02**	**60.76**

Table 2. Ablation study on loss weight λ.

λ	mAcc	mIoU
0.5	72.23	59.67
1.0	73.02	**60.76**
2.0	**73.03**	60.23

Table 3. Ablation study on the Crop mode.

Mode	mAcc	mIoU
Random Crop	69.56	56.94
Center Crop (x2)	71.92	59.77
Center Crop (Ours)	**73.02**	**60.76**

Table 4. Ablation Study on Training costs.

Method	GFLOPs	Time/iter
Baseline	1049	609ms
Ours	1129	1066ms

Table 5. Ablation study on the Frozen backbone.

Backbone	mAcc	mIoU
Frozen	71.52	58.25
Unfrozen (Ours)	**73.02**	**60.76**

shown in Table 5. From the table, we can conclude that the method with the frozen backbone learns good representations with only a slight decrease, as what we have expected before, which we ascribe to the differences between distortions learning and segmentation.

3. Detailed Results on Three Datasets

We report the overall results in the main paper. To further demonstrate the effectiveness of our method processing the distortions on panoramic images, we include more detailed results between our method and different self-supervised approaches on three datasets in Table 6, 7 and 8. It can be observed that our method achieves the best results on all three datasets (e.g., 2.63% on mIoU higher than the second-best method PPS [7] in Table 6). Besides, our method achieves the highest mIoU results on the majority of individual categories. Even for some challenging categories, our method also achieves satisfactory performance (e.g., LT in SUN360 in Table 6 and CH in Standford2D3D in Table 8). Due to the SPT module, our method could introduce more distortions in the pretraining stage and could learn more consistency from the paired views based on the DASC framework. Therefore, these detailed results verify the promotion of our method is statistically significant.

4. More Qualitative Results

In this section, we provide more qualitative results between the baseline and our proposed DASC-SPT. As shown in Figure 1, our method could produce more superior masks against baseline from an overall perspective. Specifically, the baseline may suffer from the distortions on panoramic images and the texture-closer semantic categories, thus producing incomplete semantic masks (e.g., the 3rd row and the 4th row of Figure 1) and incorrect dense predictions (e.g., the 1st row and the 6th row of Figure 1). As a comparison, our framework DASC based on the SPT could further leverage the shared content and discrepancies caused by different distortions of the paired views, which could produce more accurate dense predictions. Even in complex scenes having severe distortions and large differences in scale (e.g., the 4th row of Figure 1), our DASC-SPT still produces satisfactory results (e.g., the window), demonstrating the robustness of our method.

Table 6. Detailed comparisons with different self-supervised approaches on SUN360 dataset. The abbreviations for the following categories represent: Bed (BD), Painting (PG), Table (TB), Mirror (MR), Window (WN), Curtain (CT), Chair (CH), Light (LT), Sofa (SF), Door (DR), Cabinet (CB), Bedside (BS), TV, Shelf (SH).

Method	BD	PG	TB	MR	WN	CT	CH	LT	SF	DR	CB	BS	TV	SH	mAcc	mIoU
Supervised	70.7	63.9	43.5	38.0	56.7	72.9	34.5	64.1	53.0	56.3	36.4	55.6	66.7	10.1	66.1	51.6
MoCov2 [6]	73.8	72.8	53.8	49.8	58.3	73.2	37.9	60.2	62.4	60.4	40.1	59.1	68.8	13.2	69.2	56.0
SimCLR [2]	73.8	69.7	51.4	53.1	53.7	73.8	33.9	61.8	58.1	59.2	38.6	63.3	67.7	**28.7**	69.8	56.2
BYOL [5]	58.5	55.4	34.8	22.5	44.0	69.8	18.0	30.1	45.1	46.7	25.2	34.4	46.1	8.0	51.3	38.5
DenseCL [8]	75.2	73.2	53.5	49.9	57.6	72.6	37.8	67.1	62.9	59.8	38.4	66.3	72.0	12.4	70.1	57.1
Barlowtwins [9]	67.8	60.3	43.0	35.4	47.5	71.2	28.9	53.8	51.6	50.5	30.4	45.6	59.9	10.9	61.3	46.9
SimSiam [3]	76.9	70.7	52.5	50.0	61.1	73.9	33.8	**76.2**	60.7	62.3	41.4	65.0	70.0	12.7	70.4	57.6
VICRegL [1]	73.1	70.4	52.7	47.4	58.8	72.5	33.5	62.8	58.9	57.4	35.8	57.3	67.7	12.1	69.1	54.3
360VAM [4]	69.6	65.3	46.2	49.3	56.8	72.9	28.8	57.4	54.7	58.8	37.4	49.0	62.9	8.5	65.1	51.3
PPS [7]	**77.5**	73.4	53.3	51.4	**61.3**	**75.1**	36.6	67.1	63.4	63.0	40.0	62.0	71.4	18.3	71.4	58.1
DASC-SPT	77.3	**74.2**	**57.8**	**54.7**	60.6	74.5	**39.3**	70.7	**65.5**	**63.9**	**45.1**	**70.5**	**73.2**	21.2	**73.0**	**60.8**

Table 7. Detailed comparisons with different self-supervised approaches on CVPG-Pano dataset.

Method	Flat	Construction	Object	Nature	Sky	Person	Vehicle	mAcc	mIoU
Supervised	97.8	90.4	46.9	86.0	98.8	41.0	85.7	83.3	78.1
MoCov2 [6]	98.0	90.7	48.8	85.3	**98.9**	41.3	87.3	83.8	78.6
SimCLR [2]	97.8	90.0	42.7	85.8	98.8	35.8	85.4	82.3	76.6
BYOL [5]	96.9	87.2	29.9	84.0	98.6	18.1	78.5	75.4	70.4
DenseCL [8]	97.9	90.5	48.3	85.3	98.8	39.5	86.4	83.9	78.1
Barlowtwins [9]	97.6	89.0	37.9	84.6	98.7	32.8	83.4	79.9	74.9
SimSiam [3]	98.0	91.2	51.8	**86.7**	**98.9**	46.9	88.4	85.2	80.3
VICRegL [1]	97.9	90.3	43.1	85.6	98.8	31.4	85.3	80.8	76.0
360VAM [4]	97.7	89.7	44.2	85.1	98.8	32.7	85.1	81.8	76.2
PPS [7]	**98.1**	91.1	47.7	86.3	**98.9**	40.1	86.8	83.4	78.4
DASC-SPT	**98.1**	**91.3**	**53.2**	86.3	**98.9**	**48.7**	**88.6**	**86.3**	**80.7**

References

[1] 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. 3, 4

[2] 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. 3, 4

[3] 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. 3, 4

[4] 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. 3, 4

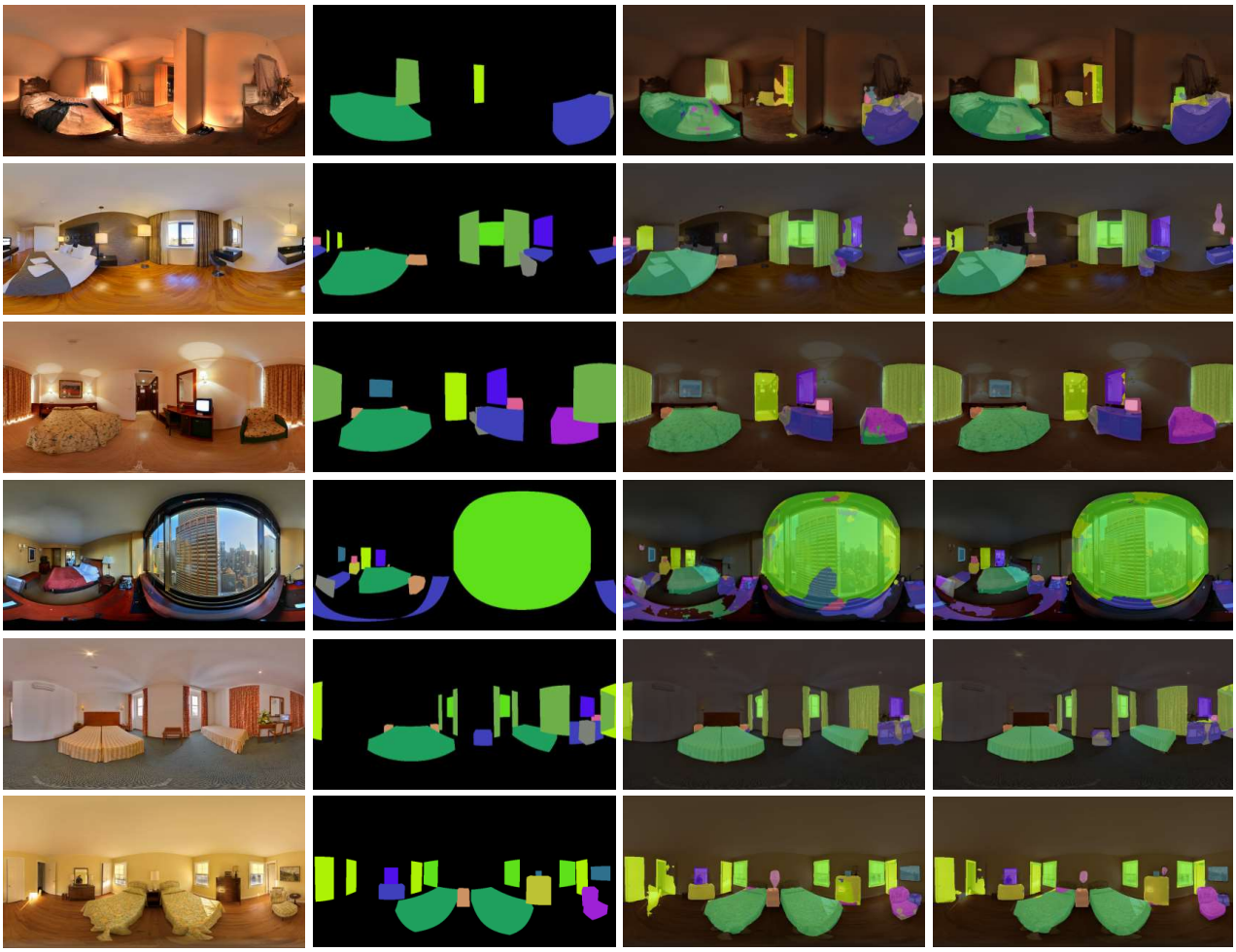
[5] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 3, 4

[6] 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. 3, 4

[7] Alexander Jaus, Kailun Yang, and Rainer Stiefelhagen. Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised con-

Table 8. Detailed comparisons with different self-supervised approaches on Standford2D3D dataset. The abbreviations for the following categories represent: Board (BD), Bookcase (BK), Beam (BM), Ceiling (CG), Chair (CH), Clutter (CT), Column (CN), Door (DR), Floor (FL), Sofa (SF), Table (TB), Wall (WL), Window (WN).

Method	BD	BK	BM	CG	CH	CT	CN	DR	FL	SF	TB	WL	WN	mAcc	mIoU
Supervised	65.8	56.6	0.7	70.9	47.8	24.3	10.2	17.3	92.2	19.9	54.5	69.8	35.0	50.4	40.4
MoCov2 [6]	69.0	57.2	0.2	71.2	51.3	25.2	9.4	22.6	90.9	19.5	45.9	69.9	38.4	50.5	40.9
SimCLR [2]	68.2	53.2	0.2	69.8	46.9	24.2	10.8	21.6	**92.7**	13.8	48.5	67.3	**44.6**	49.5	40.2
BYOL [5]	63.1	50.7	**0.6**	70.4	39.1	22.7	5.1	17.2	90.3	11.3	46.8	67.3	40.3	46.9	37.6
DenseCL [8]	68.0	57.3	0.1	68.9	49.6	25.3	10.7	32.2	92.4	25.4	51.9	70.4	42.6	51.7	42.6
Barlowtwins [9]	65.2	52.8	0.5	71.3	44.8	24.0	10.8	24.7	90.6	14.9	47.3	67.5	38.1	49.1	39.5
SimSiam [3]	72.0	56.9	0.1	71.0	51.0	24.7	8.4	24.7	92.3	**31.9**	54.2	**71.2**	38.6	51.4	42.7
VICRegL [1]	67.9	55.2	0.0	71.2	47.7	22.7	8.3	**27.7**	92.6	21.4	52.1	68.6	36.1	49.6	41.0
360VAM [4]	66.4	53.8	0.3	**72.1**	46.8	24.7	8.5	18.0	91.9	17.4	49.8	69.0	42.8	49.3	40.2
PPS [7]	69.2	**58.1**	0.1	71.9	**54.0**	**25.9**	11.0	24.7	92.1	23.8	**55.1**	71.1	37.1	52.0	42.6
DASC-SPT	**70.3**	57.1	0.2	71.1	53.6	24.8	**11.1**	25.2	91.3	**31.9**	53.5	68.7	40.1	**52.3**	**43.0**



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

Figure 1. More qualitative results are provided between the baseline (SimSiam) and our proposed DASC-SPT approach.

trastive learning. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 1421–1427. IEEE, 2021. 2, 3, 4

[8] 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. 3, 4

[9] 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. 3, 4