Self-supervised Learning for Semantic Tooth Segmentation on Cone-Beam CT Image

Anonymous Authors

No Institute Given

Abstract. Semantic tooth segmentation in the Cone-Beam Computed Tomography(CBCT) images is essential is essential for dental diagnosis. The success of deep learning provides a solution to realizing automatic segmentation, while requiring a large-scale labeled CBCT dataset - a condition that is hard to meet because of the challenging nature of manually annotating complex CBCT images. In this paper, we propose a novel self-supervised framework to boost the accuracy of CBCT tooth semantic segmentation. Our method first employs a self-supervised pre-training network, which is regulated by a modified contrastive loss that is computed based on the spatial distances between local regions within each CBCT image, to pre-train a Swin Transformer backbone with a large-scale unlabeled CBCT dataset. Next, we finetune the network with an UPerNet segmentation head on a small expertly annotated CBCT dataset. Compared to fully supervised methods trained by the same amount of annotated samples, our method achieves a superior performance of 91.33% tooth IoU. Moreover, our method can obtain better performance with only 25% of annotated samples of supervised counterparts. Our work presents a possible solution to reducing the human efforts in CBCT image segmentation.

Keywords: Self-supervised Learning · Semantic segmentation · Cone-Beam Computed Tomography · Digital dentistry.

1 Introduction

Reconstructing 3D dental models is important for many dental applications, such as orthodontics and implant. There are mainly two types of 3D dental models: Intraoral scanning and Cone-Beam Computed Tomography (CBCT) [1]. Intraoral scans can provide accurate reflection of the tooth crowns with a very high resolution, while CBCT, though with a lower resolution, can provide comprehensive 3D information for most oral tissues, including tooth crowns, tooth roots, alveolar bone, etc. In modern digital dentistry, it is critical to obtain complete anatomical information for more precise and clinically applicable treatment planning, e.g., avoiding alveolar bone fenestration in orthodontics.

Many methods have been proposed for tooth segmentation in the past decades. Traditional solutions are mainly based on thresholding [5, 6] or level-set methods [2–4, 7], which usually require manual initialization or prior knowledge, e.g., initial seeds or starting slices [7]. Recently, due to the success for segmentation tasks, several deep learning methods have been proposed for tooth instance segmentation on CBCT images [8, 11–14]. ToothNet [9] uses a two-stage network to achieve automatic tooth segmentation from CBCT images. Ezhov et al. [8] propose a coarse-to-fine framework to segment individual teeth in 3D CBCT images, where the model is trained with a large coarsely labeled dataset and subsequently fine-tuned with a smaller downscaled precisely labeled dataset. Wu et al. [10] propose a two-level hierarchical deep neural network which first gets the localization information and then realizes accurate boundary segmentation with the DenseASPP-UNet. All of these works are supervised methods that require precisely labeled samples. However, manually labeling 3D CBCT scans is very labor-intensive, e.g., it takes about 30 to 60 minutes to annotate a single CBCT slice for an experienced dentist, while a CBCT scan usually consists of hundreds of slices. Such a conflict between the demand for annotated samples and the under-supply expertise and human labor motivates us to develop algorithms that can segment tooth with only few annotated slices.

Self-supervised learning (SSL) has shown superior performance in many vision tasks by learning expressive representations on unlabeled data via pretext tasks [15–17, 28, 29]. In many SSL frameworks, it is not uncommon to perform pre-training via a contrastive learning strategy that usually minimizes the distance between positive pairs while pushing away negative samples, where positive pairs are usually different augmentations of the input image and negative samples are sampled in the dataset otherwise. Recently, self-supervised learning has demonstrated its effectiveness on medical images classification or segmentation downstream tasks [18–22]. However, to the best of our knowledge, there exists no prior work of SSL for tooth segmentation on CBCT images due to several domain-specific challenges. While most existing SSL methods, such as BYOL [15], are pre-trained with image-level contrastive loss, which might be suboptimal to generate powerful pixel-level representations for downstream dense semantic segmentation tasks. Besides, many images within a CBCT dataset bear a strong coarse-level resemblance. However, the challenges for CBCT image segmentation remain for complicated anatomical structures, such as indistinguishable boundaries between tooth contacts of upper and lower jaws, ambiguous pixels among tooth, alveoli and alveolar bones, indicating that novel dense representation learning strategies are needed rather than only selecting images-level negative samples for coarse contrastive learning.

In this paper, we propose a novel SSL pre-training method that employs dense contrastive learning on large-scale unlabeled CBCT images in order to boost the performance of the downstream semantic segmentation task. We built a CBCT image dataset that consists of 123,904 unlabeled images and 2,903 annotated images from 400 patients. To better align the pre-training network with the downstream segmentation tasks, we inherit the siamese architectures that are designated to learn image-level representations, but further integrate a dense contrastive learning branch to calibrate pixel features based on the spatial distance between the local regions within every image view. With Swin Transformer [24] as the backbone, our tooth segmentation network is first pre-trained with the SSL framework and dense contrastive loss, and subsequently finetuned with an UPerNet [32] segmentation head on a small labeled CBCT dataset. Compared to fully supervised networks trained by the same amount of annotated samples, our method achieves state-of-the-art performance of 91.33% of IoU on a hold-out set of 903 images from 43 patients. Moreover, our method can achieve better performance than the supervised counterparts with only 25% of annotated samples. Our work demonstrates that it is highly possible to achieve accurate CBCT segmentation with a limited number of human laborers.

2 Method

2.1 Overview

Given a dataset $\mathbf{B} = \{\mathbf{x}_i\}_{i=1}^{N}$, where $\mathbf{x}_i$ denotes a 2D Cone Beam CT image with a resolution of 512×512 or 640×640, we aim to develop a model that automatically annotates each pixel in the image with two categories, i.e., tooth or background. The core of our method is a two-stage network. The first stage involves pre-training a Swin Transformer backbone by feeding two augmented views of each CBCT image into the encoder network for representation learning. The network is trained with a novel dense contrastive loss function in a self-supervised manner. The second stage finetunes the pre-trained backbone with a subsequent UPerNet segmentation network supervised by a small-scale annotated CBCT dataset.

2.2 Self-supervised Pretraining

Image Augmentation The overall self-supervised pre-training framework is illustrated in Fig. 1. We adopt the general siamese network architectures which take two different augmented views of the same input images for representation learning. As for the augmentation strategies, we first randomly select a patch of input images and then resize them to a resolution of 224 × 224. Subsequently, a random horizontal flip is applied, followed by random crop and resize, color jittering, grayscale conversion, as well as Gaussian blur. Finally, optional solarization is applied. Noted that during the random crop-resize stage, the coordinates of the left bottom and the top right pixel of the cropped view will be recorded, so that the pixels in the output feature map can be easily mapped back to the original image for representation learning.

Segmentation backbone Swin Transformer is a general-purpose backbone for computer vision that achieved state-of-the-art performance on various vision tasks [27]. It hierarchically computes the representation with shifted windows while preserving great efficiency [24]. In this work, we adopt the tiny version of Swin Transformer (Swin-T) as our default backbone. Specifically, given an image $\mathbf{x}_i$ and its augmented view v, the backbone f encodes v into a feature vector

$\mathbf{y}_i = f(v) \in \mathbb{R}^{768}$ via a 4-stage network, where each stage consists of a patch merging module and different numbers of Swin Transformer blocks(2/2/6/2, respectively). The hierarchical architecture in Swin Transformer can learn both local and global representations. The local features help to identify the boundary between the background and the tooth classes, while the global features provide richer context information for robust classification. Though we also demonstrate the effectiveness of our framework with ResNet backbone [23], we experimentally find Swin Transformer to be the best option. More details are provided in [24].

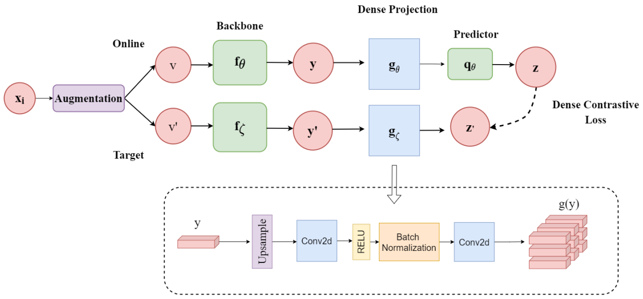

Fig. 1: Architecture of the dense contrastive learning.

Self-supervised learning framework The main architecture of our pre-training framework is illustrated in Fig.1. It inherits the general asymmetric siamese network architectures like BYOL, with modifications to adapt to dense contrastive learning that is conducted among local regions within each CBCT image. The pre-training pipeline consists of an online network f_θ and a target network f_ζ, which share the same architecture, except that they are defined by different sets of weights: θ and ζ, and the online network has an extra predictor q_θ. The target network is trained with the regression targets provided by the online network. For each training step, with target decay rate τ, the target parameters ζ are updated by the slow-moving average of online parameters θ:

$$\zeta \leftarrow \tau\zeta + (1-\tau)\theta \tag{1}$$

Given a single CBCT image $\mathbf{x}_i$, we first generate two augmented views: v and v' which are respectively fed into the online network and the target network. For the target network, v' will go through the Swin Transformer backbone f_ζ to obtain a feature map $y' = f_\zeta(v') \in \mathbb{R}^{768}$. Subsequently, the the dense projection network g_ζ will upsample y' into a dense dense feature map $z' = g_\zeta(y') \in \mathbb{R}^{768\times7\times7}$. Similarly, the view v in the online network will also go through the backbone network f_θ and the projection network g_θ, but has an extra MLP predictor q_θ to output a feature map $z = q_\theta(g_\theta(y)) \in \mathbb{R}^{768\times7\times7}$.

A critical part in our framework is the projection network. It consists of two 1×1 convolutional layers with a RELU layer and a batch normalization layer in between. After the feature map y and y' are obtained from the Swin backbone, they are upsampled by a factor of 7 to be transformed into a dense manner. The projection network g_θ and g_ζ will then respectively produce two nonlinear projections of the feature maps: $g_\theta(y)$, $g_\zeta(y') \in \mathbb{R}^{768\times7\times7}$, with spatial resolution of 7×7, where a local contrastive learning pretext task can be constructed. Compared to most existing contrastive learning frameworks where the contrastive loss is computed at image level, the projection head network in this paper is modified to produce a dense projection, and thus enables the contrastive loss to be calculated based on the local regions within the image views.

Dense Contrastive Learning Given the dense feature map z and $z' \in \mathbb{R}^{768\times7\times7}$, we define the negative and positive pairs based on the spatial distance of local regions in the feature maps. Both z and z' have 49 local feature regions. As we record the coordinates of corner pixels of the cropped views in the image augmentation period, we can map the feature maps back to the coordinates in the original CBCT image and calculate the distance between each feature point in the two views:

$$dist(i,j) = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2} \tag{2}$$

where i, j stand for feature regions in z, z' respectively, and X_i, X_j, Y_i and Y_j are the corresponding positions of the i- and j-th feature regions in the 2D Cartesian coordinate. Here we set a distance threshold $\mu = 0.8$ to define the positive and the negative pairs of the pixels as in [26]. The pair (i,j) is defined to be positive if $dist(i,j) < \mu$, and negative if $dist(i,j) \geq \mu$. The dense contrastive loss of a single feature point in the feature map is defined similar to the InfoNCE loss [25]:

$$\mathcal{L}_i = -\frac{1}{\mathcal{D}} log(\frac{\sum e^{cos(h_i,h_j^+)/\lambda}}{\sum e^{cos(h_i,h_j^+)/\lambda} + \sum e^{cos(h_i,h_j^-)/\lambda}}) \tag{3}$$

where $\mathcal{D}$ is the diagonal length of the feature map; h_i is the feature vector of feature point i; h_j^+ and h_j^- are the feature vectors of j that is denoted as the positive or negative pair of i; λ is a temperature hyper-parameter that is set to 0.3 by default. The dense contrastive loss of a single view is the averaged loss of all the 49 feature regions in the view, and the final loss is averaged across all the views in the batch. Optimizing over the dense contrastive loss would maximize the dissimilarity of the feature vectors generated by spatially distant pixels, while encouraging the spatially close pixels to output similar features, leading to a good initialization of pixel-wise representations in CBCT images for the subsequent tooth semantic segmentation.

2.3 Finetune

The pre-trained backbone is finetuned for the downstream tooth semantic segmentation task. A Unified Perceptual Parsing Network (UPerNet) [32] is cascaded to the network and trained in a supervised manner with a tiny-scaled annotated CBCT dataset. The image x_i and the corresponding annotated mask are identically augmented before they are fed to the network. The UPerNet segmentation head is trained to capture hierarchical image information from the feature map encoded by the backbone, and to predict texture labels for semantic segmentation, with the supervision from the annotated samples.

3 Experiment

3.1 Experimental Setup

The dataset for SSL pre-training consists of 123,904 unlabeled CBCT images from 400 patients, with a resolution of 512×512 or 640×640. The dataset for finetuning consists of 2,903 CBCT images with annotations by experienced dentists, split into a training set with 2,000 images from 102 patients and a test set with 903 images from 43 patients. In the pre-training period, we use Swin-T as the backbone, trained by the AdamW optimizer with learning rate $\eta_1 = 0.001$, cosine decay rate=0.05, and batch size $b = 64$ for 300 epochs. We use the AdamW optimizer with the learning rate $\eta_2 = 6e - 05$ and a polynomial decay rate of 0.001 during finetuning. The network in the finetuning stage is trained for 160 epochs with batch size $b = 2$. We comprehensively evaluate the performance of our method with various metrics, including Intersection-over-Union(IoU), Dice Similarity Coefficient, precision, and recall over the predicted tooth masks. Given the predicted tooth masks P and the label T, the IoU and Dice are computed by $\frac{P \cap T}{P \cup T}$ and $\frac{2|P \cap T|}{|P|+|T|} = \frac{2 \times precision \times recall}{precision + recall}$,respectively.

3.2 CBCT semantic segmentation results

Table 1: Segmentation performance of our method and baselines.

Method	IoU	Dice	Recall	Precision
FCN	88.87	94.11	98.08	90.44
Deeplabv3	86.87	92.98	90.9	**95.15**
Swin	90.29	94.9	97.85	92.12
BYOL	90.78	95.17	**98.45**	92.10
Ours(Resnet-101)	91.01	95.29	96.65	93.98
Ours(Swin)	**91.33**	**95.47**	98.33	92.77

We evaluate the performance of our method (with Swin-T and Resnet-101 as the backbones) and compare with several supervised baselines, i.e., Fully Convolutional Networks(FCN) [30], Deeplabv3 [31], Swin Transformer, as well as the image-level SSL method BYOL, with results reported in Table.1. Note that BYOL uses Swin-T as the backbone, while FCN and Deeplabv3 use ResNet-101 as the backbone. While both BYOL and our method significantly outperform all supervised baselines, our method further boosts the performance over BYOL, achieving an IoU of 91.33%. This demonstrates that our method can achieve non-trivial improvement compared to its supervised and self-supervised counterparts. Furthermore, our method is also effective over different backbones, with larger performance gain for the less powerful backbone ResNet-101.

We further evaluate the performance of our method when using different amount of unlabeled data in the pre-training period. Unless otherwise indicated, we use Swin-T as the backbone. Our method can obtain 91.00% of IoU and 95.29% of Dice even with only 1% unlabeled data, while increasing the amount of unlabeled data during pretraining can lead to constant performance improvement, which further demonstrates the effectiveness of our method.

Table 2: Performance with different ratios of unlabeled data during pre-training.

Unlabeled Data Ratio	IoU	Dice	Recall	Precision
1%	91.00	95.29	98.12	92.62
50%	91.15	95.37	98.26	92.65
100%	**91.33**	**95.47**	**98.33**	**92.77**

Table 3: Performance with different ratios of labeled data.

Data Ratio	Training Strategy	IoU	Dice	Recall	Precision
1%	From scratch	40.90	58.06	47.04	75.81
	Ours	59.94	74.95	68.02	83.45
5%	From scratch	82.53	90.43	96.28	85.24
	Ours	85.82	92.37	96.93	88.22
10%	From scratch	86.63	92.84	98.13	88.09
	Ours	89.18	94.28	97.37	91.38
25%	From scratch	87.81	93.51	98.67	88.86
	Ours	91.06	95.32	97.31	93.40
50%	From scratch	88.16	93.71	98.81	89.11
	Ours	91.27	95.44	98.00	93.01
100%	From scratch	90.29	94.90	97.85	92.12
	Ours	**91.33**	**95.47**	**98.33**	**92.77**

We also carry out multiple experiments to evaluate the effect with different amounts of labeled data during finetuning, with results reported in Table 3. We compare our method with the network that employs the same Swin-UPerNet architecture but trained from scratch with fully supervised settings. Our method can consistently outperform its supervised counterparts, especially for settings with limited annotated samples. When using only 25% of the annotated CBCT images, our method achieves competing or even superior performance compared to its counterpart that is trained with the entire labeled dataset.

3.3 Ablation Study and Visualization

The ablation study here focuses on the image augmentation strategies in pre-training. As shown in Table.4, Random Crop Resize brings about the most prominent improvement in performance. Canceling both random crop Resize and flip brings about a significantly inferior performance, indicating the importance of these two augmentation strategies. The visualization of 3 cases is shown in Fig.2, where our method commits much fewer mistakes for false positive and false negative predictions, see more visualizations in the Appendix.

Table 4: The result on CBCT dataset with different augmentation.

Random Crop Resize	Flip	Color Jittering	IoU	Dice	Recall	Precision
✓	✓	✓	**91.33**	**95.47**	98.33	**92.77**
✓	✓		90.69	95.12	**98.69**	91.79
	✓	✓	88.01	93.32	97.40	89.57
		✓	87.23	93.83	93.76	93.90

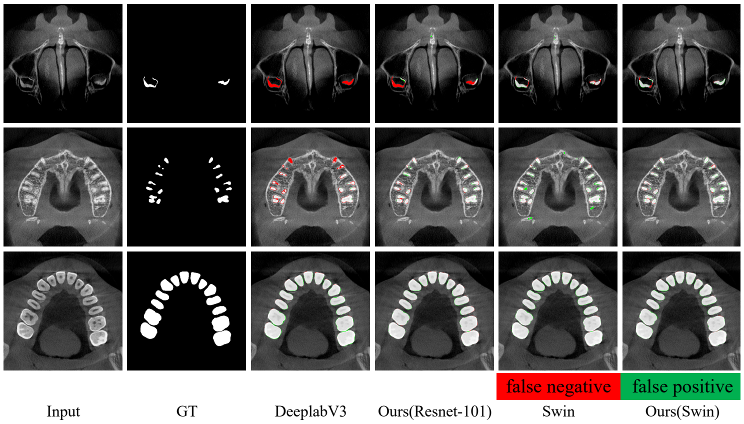

Fig. 2: Visualization of the segmentation result of different methods. The backbone of Deeplabv3 is Resnet101. More results are presented in Appendix.

4 Conclusion

In this paper, we propose a method that improves the performance of semantic segmentation for CBCT images by the means of a SSL pre-training strategy.

Compared to fully supervised networks trained by the same amount of annotated samples, our method achieves a superior performance of 91.33% tooth IoU. Moreover, at low annotated dataset settings, our method can obtain better performance with only 25% of annotated samples of supervised counterparts. The extensive experiments convincingly illustrate the effectiveness of the proposed Self-supervised pre-training strategy for reducing the necessity of manually annotated data in CBCT image segmentation.

References

1. S. J. Merrett, N.A.Drage, and P. Durning.: Cone beam computed tomography: A useful tool in orthodontic diagnosis and treatment planning. J. Orthodontics, vol. 36, no. 3, pp. 202—210 (2009)
2. D. X. Ji, S. H. Ong, and K. W. C. Foong.:A level-set based approach for anterior teeth segmentation in cone beam computed tomography images. In: Comput. Biol. Med., vol. 50, pp. 116–128 (2014)
3. H. Gao and O. Chae. : Individual tooth segmentation from CT images using level set method with shape and intensity prior. In: Pattern Recognit., vol. 43, no. 7, pp. 2406–2417 (2010)
4. M. Hosntalab, R.A. Zoroofi, A. A. Tehrani-Fard, and G.Shirani.: Segmentation of teeth in ct volumetric dataset by panoramic projection and variational level set. In: International Journal of Computer Assisted Radiology and Surgery, 3(3-4): 257–265 (2008)
5. H. Heoand O.-S. Chae.: Segmentation of tooth in CT images for the 3D reconstruction of teeth. In: Proc. SPIE, vol. 5298, pp. 455–467 (2004)
6. H. C. Kang, C. Choi, J. Shin, J. Lee, and Y.-G. Shin.: Fast and accurate semiautomatic segmentation of individual teeth from dental CT images. In: Comput. Math. Methods Med., vol. 2015, pp. 1–12 (2015)
7. Y. Gan, Z. Xia, J. Xiong, G. Li, and Q. Zhao.: Tooth and alveolar bone segmentation from dental computed tomography images. In: IEEE journal of biomedical and health informatics, 22(1): 196–204 (2018)
8. Ezhov M, Zakirov A, et al.: Coarse-to-fine Columetric Segmentation of Teeth in Cone-beam CT. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 52–56 (2019)
9. Cui Z, Li C, et al.: ToothNet: Automatic Tooth Instance Segmentation and Identification from Cone Beam CT Images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6368–6377 (2019)
10. X. Wu, H. Chen, Y. Huang, H. Guo, T. Qiu, L. Wang. Center-sensitive and Boundary Aware Tooth Instance Segmentation and Classification From Cone-beam CT, In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp.939–942 (2020)
11. Chung, Minyoung, et al.: Pose-Aware Instance Segmentation Framework From Cone Beam CT Images for Tooth Segmentation. Computers in Biology and Medicine 120, 103720 (2020)
12. Wang, H., et al.: Multiclass CBCT Image Segmentation for Orthodontics with Deep Learning. In: Journal of Dental Research100.9 (2021), pp. 943–949 (2021)
13. Lahoud, Pierre, et al.: Artificial intelligence for fast and accurate 3-dimensional tooth segmentation on cone-beam computed tomography. Journal of Endodontics 47.5, pp.827–835 (2021)

14. Lee, S., et al.: "Automated CNN-Based tooth segmentation in cone-beam ct for dental implant planning. In: IEEE Access 8 (2020), pp. 50507–50518 (2020)
15. Grill, Jean-Bastien, et al.: Bootstrap Your Own Latent-a New Approach to Self-supervised Learning. In: Advances in Neural Information Processing Systems 33, pp. 21271–21284 (2020)
16. He, Kaiming, et al.: Momentum Contrast for Unsupervised Visual Representation Learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020)
17. Chen, Ting, et al. A Simple Framework for Contrastive Learning of Visual Representations. In: International conference on machine learning (2020)
18. Zhuang, Xinrui, et al.: Self-supervised Feature Learning for 3d Medical Images by Playing a Rubik's Cube." International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham (2019)
19. Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S.E., Guo, Y., Matthews, P.M., Rueckert, D.: Self-supervised Learning for Cardiac MR Image Segmentation by Anatomical Position Prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 541–549. Springer (2019)
20. Chaitanya, Krishna, et al.: Contrastive Learning of Global and Local Features for Medical Image Segmentation with Limited Annotations. In: Advances in Neural Information Processing Systems 33, pp. 12546–12558 (2020)
21. Zhang, Pengyue, Fusheng Wang, and Yefeng Zheng.: Self Supervised Deep Representation Learning for Fine-Grained Body Part Recognition. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (2017)
22. Jamaludin, Amir, Timor Kadir, and Andrew Zisserman.: Self-supervised Learning for Spinal MRIs. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 294–302. Springer, Cham (2017)
23. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
24. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
25. Aaron van den Oord, Yazhe Li, and Oriol Vinyals.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
26. 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: IEEE Conference on Computer Vision and Pattern Recognition (2021)
27. Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu: Self-Supervised Learning with Swin Transformers. arXiv preprint arXiv:2105.04553 (2021)
28. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018)
29. He, Kaiming, et al. "Masked autoencoders are scalable vision learners." arXiv preprint arXiv:2111.06377 (2021).
30. E. Shelhamer, J. Long and T. Darrell: Fully Convolutional Networks for Semantic Segmentation, In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.39, no.4, pp.640-651 (2017) https://doi.org/: 10.1109/TPAMI.2016.2572683.

31. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
32. Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun: Unified Perceptual Parsing for Scene Understanding. ECCV (5) pp. 432–448 (2018)