Computer Fraud and Security
ISSN (online): 1873-7056

Pre-Training Study on Image Segmentation based on Multi-Scale Contrast Learning

JinHui Zhang,Wensong Liu,YaYun Sun,YuXuan Zhao,Yuxiao Zhao,Xiang Wang,YuQiong Wang, Peng Wang*

NARI Group Corporation (State Grid Electric Power Research Institute);China Realtime Database Co., Ltd., Nanjing 211106, China

***Corresponding Author:**Peng Wang

Email:zhaoyuxiao@sgepri.sgcc.com.cn

Abstract: the current popular contrast self-supervised pre-training method is mainly more suitable for the downstream task of global image classification, and for image segmentation of spatial information demanding intensive prediction task, the effect is not satisfactory, so we proposed a multi-scale contrast learning based on image segmentation model, the method in the global, local and pixel three contrast learning, fill the gap between self-supervised pre-training and intensive prediction task. The method for the problem of limited annotation data, efficient use of a small amount of annotation data, data through color enhancement, and then use multiscale contrast loss function training can accurately extract the image features of training model, the training model can in image segmentation this task, according to different specific tasks, fine-tune the overall segmentation network. We demonstrate the effectiveness of the proposed training strategy using the Cityscapes and PASCAL VOC 2012 segmentation datasets. Our results show that pre-training with the proposed contrast loss can achieve high performance gain for the intensive prediction task of image segmentation with limited amount of labeled data, outperforming existing technical methods.

Keywords: contrast learning; image segmentation; pre-training

1. Introduction

Image segmentation is an important research task in the field of computer vision and plays an important role. Image segmentation is a pixel-level classification of the image into several different sub-regions. Among them, the pixels in the same area have a certain correlation, and there are certain differences in the different areas of the pixels, that is, the process of giving the same label to the pixels with the same nature in the picture. Image segmentation technology is widely used in intelligent security, driverless vehicles, satellite remote sensing, medical image processing and other fields. Because the accuracy of segmentation will affect the effectiveness of subsequent image analysis, recognition and other tasks, so it is of great significance.

Image pre-training (Pre-train) refers to a model pre-trained model in image processing tasks or the process of pre-trained model, which has been widely used in the field of computer vision in recent years. Since the

datasets that people have are small scale, it is difficult to train a CNN network from scratch, so the general operation is to train a model on a large dataset and then fine-tune some of the parameters of the model on their own tasks. The pre-trained model allows us not to repeatedly train large-scale models and can rely on less data, saving us a lot of time and computational resources. By training on the ImageNet dataset, many pre-trained models were generated, including LeNet-5[1]、AlexNet[2]、GoogLeNet[3]、VGG-16[4]、VGG-19[4]And ResNet-50[5]Such as network model.

Contrast learning is a self-supervised learning method for learning the general features of a dataset by letting the model learn which data points are similar or different without labels. By maximizing the difference between positive samples (the corresponding samples in the input and target modes) and negative samples (the randomly selected samples in the input and target modes), contrast learning enables the backbone network to distinguish related and irrelevant samples, so as to better serve downstream tasks. Due to the outstanding performance of contrast learning in the field of self-supervised learning, in recent years, more and more researchers have applied contrast learning to pre-training tasks, fully utilizing contrast learning to improve the ability of backbone networks to distinguish between relevant and irrelevant samples, thus improving the performance of downstream tasks.

At present, most of the mainstream image segmentation pre-training techniques are supervised learning methods, which rely on manually annotated labels, and the information provided by the data itself is richer than sparse labels. The pre-training model using supervised learning methods cannot include comprehensive data features. Therefore, in order to solve the above problems, this paper proposes a pre-training method combined with multi-scale contrast learning, by extracting the original image and enhance the image of multi-scale contrast loss, so as to use multi-scale contrast loss, in the case of limited marker data can also efficient training network, improve the performance of the image segmentation model.

2. Related work

2.1 Contrast learning

After Turing Award winner Yann Lecun delivered a lecture on contrast learning in AAAI 2020, contrast learning technology became one of the most popular ideas in self-supervised learning. The core idea of contrast learning is to learn the high-dimensional features of samples by comparing positive and negative sample pairs[6], Specifically, the input is expressed as a feature vector in parameter form, and close the distance between positive samples in the feature space while making the distance as far as possible, so as to achieve the goal of distinguishing different inputs. The positive sample pair is the sample associated with the input data, such as taking the input enhanced by the data as the positive sample, while the negative sample is any data within the batch that is unrelated to the current input. Compared to generative learning treating each pixel equally, contrast learning treats a portion of the input data as a more important representation and weights it when encoding[7], This makes contrast learning learned features facilitate downstream classification tasks including target detection, image segmentation and more[8]。

In contrast learning, the design of agent tasks[9]And the amplification of the negative sample library[10]Is the key to determining the performance of this model. Early contrast learning is mainly end-to-end, where

InstDisc[11]The negative sample library is expanded by establishing the form of memory bank, while the later Moco series chooses to establish a momentum update queue to ensure higher computing resource utilization. Other work has explored the comparative learning methods that compare the model performance to the batch size solution corner, such as BYOL[12]， SimSiam[13]; And CMC that improve performance by increasing mutual information[14]Other methods such as.

2.2 Image segmentation and pre-training

Pre-training is a very important step in computer vision. Usually, when we train a classification, segmentation, or detection model, we will load a pre-trained backbone network, which can speed up the training of downstream tasks. The pre-trained agent task is usually the image classification task, and transferring the trained backbone network directly to a pixel-level task such as segmentation may lead to inter-task migration difficulties. Pre-training usually retains only the backbone network, and the decoder of the image segmentation model is randomly initialized, which may also lead to the instability of the network structure.

With the development of deep learning in the field of computer vision, a variety of pre-training methods for semantic segmentation have been proposed. These methods can be divided from the following aspects, from the pre-trained network structure can be divided into pre-trained encoder and decoder or only encoder. From the pre-trained agent tasks can be divided into image-level global tasks and pixel-level local tasks. The pre-training method can be divided into supervised pre-training and unsupervised pre-training. The specific method is like for the slices at different positions of the same image, to restrict the consistency of their original position characteristics[15]. You can also randomly crop two kinds of an image, paste the cropped image as a foreground on two different background pictures, and train the network to produce consistency on the foreground features of different synthetic images[16]. Such pre-training methods are all restricting the invariance of the model on the pixel-level features of the same position. Pre-trained datasets are pretrained using large publicly available datasets such as Imagenet or only some dedicated image semantic segmentation datasets[17]Then the image segmentation task.

3 Methods

$I\hat{I}(I,\hat{I})h^I = f(I)h^{\hat{I}} = f(\hat{I})$ This paper proposes a pre-training method for image segmentation based on multi-scale contrast learning. Given the original image and its color enhanced image, the image pair input, and then uses the feature extraction network to feature extract the image, i. e., the feature extraction network here can be ResNet or any other mainstream feature extractor. The acquired image feature pairs are mapped to higher order representations by projection heads at different scales, and contrast learning on three global, local and pixel scales.$(h^I, h^{\hat{I}})$

Figure 1 shows the overall network framework of this paper. The method proposed in this paper combines three scales of contrast loss to pre-train the network: (1) global contrast loss, (2) local contrast loss, and (3) pixel contrast loss.$\mathcal{L}_g\mathcal{L}_l\mathcal{L}_p$

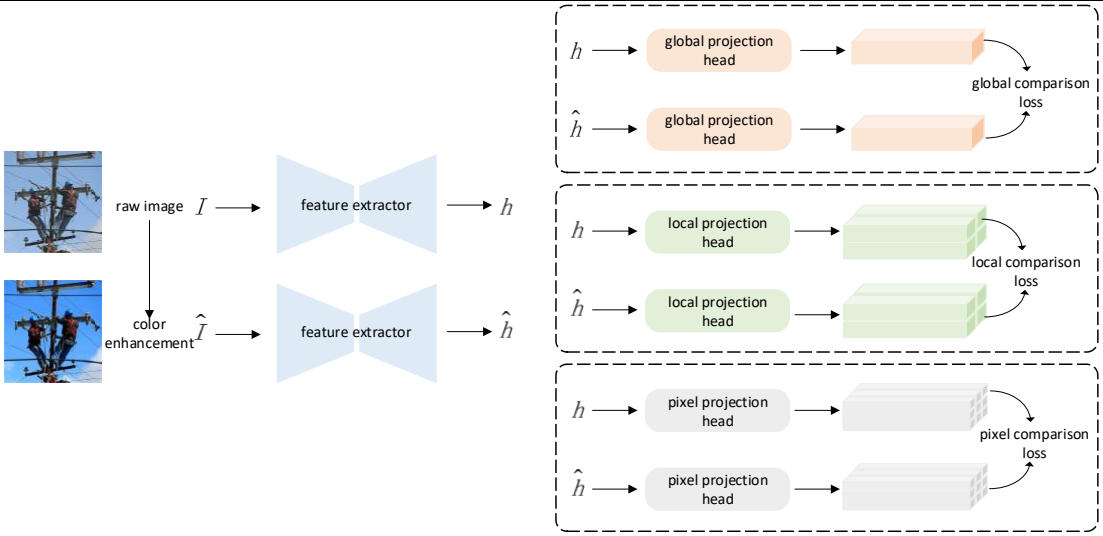

Figure 1. Overall network structure diagram

.1 3 for global-and local-based contrast learning

For general contrast self-supervised algorithms, it is based on global feature contrast learning. Specifically, color enhancement of the image gives a pair[17], Then the features are extracted by the feature extraction network, and then the global contrast loss is calculated through the global projection head. The global projection head here consists of a global pooling layer and two fully connected layers, and the global contrast loss is expressed as:

$$\mathcal{L}_g = -\log \frac{\exp\left(z^I \cdot z_+^{\hat{I}}/\tau\right)}{\exp\left(z^I \cdot z_+^{\hat{I}}\right) + \sum_{z_-^{\hat{I}}} \exp\left(z^I \cdot z_-^{\hat{I}}/\tau\right)}$$

$z^I I z_+^{\hat{I}} z_-^{\hat{I}}$Represents features of the current image, positive sample pair image features of the current image, and negative sample pair image features of the current image. Using the contrast loss function will close the positive key while pushing it away from other negative keys$z^I z_+^{\hat{I}}$[18]。

$h^I \in R^{H\times W\times K} h^{\hat{I}} \in R^{H\times W\times K}$ However, the contrast learning based on local features improves the above framework, and the main difference is that the algorithm calculates the contrast loss between the local feature vectors output by the local projection head at the local feature level. Specifically, the features are extracted from the feature extraction network, and the image features are denoted as,. Instead of pooling globally, it is directly spatially. Then each image block passes through a local projection head. The projection head is a 1 * 1 convolution layer[19]$z^I \in R^{S\times S\times E} z^{\hat{I}} \in R^{S\times S\times E}$, The output features of the local projection head are marked as,. Finally, the contrast loss of the local features is calculated:

$$\mathcal{L}_l = \frac{1}{S^2}\sum_s -\log \frac{\exp\left(z_s^I \cdot z_{s_+}^{\hat{I}}/\tau\right)}{\exp\left(z_s^I \cdot z_{s_+}^{\hat{I}}\right) + \sum_{z_{s_-}^{\hat{I}}} \exp\left(z_s^I \cdot z_{s_-}^{\hat{I}}/\tau\right)}$$

$z_s^I IS z_{s_+}^{\hat{I}} z_s^I z_{s_}^{\hat{I}} z_s^I z_s^I z_s^{\hat{I}}$Where, representing the feature vector at the first spatial position in the current image, namely query, representing the matching feature vector, that is, the positive feature, representing the mismatching feature vector, and the global loss function, and is no longer represents a whole image, but represents a block of image, that is, the local feature of the view. In addition, we stipulate that the matching mode is to correspond to the local feature vector of the image pair one to one.$z^I z^{\hat{I}}$

.2 3. Pixel-based contrast learning

Contrast learning based on pixel-level features is further improved on contrast learning based on local features, since image segmentation is pixel-level classification, considering the predictive power of pixel-level features. Specifically, the feature extraction network directly uses three 1 * 1 convolution layers without block processing$h^I h^{\hat{I}[20]} z^I \in R^{H\times W\times K} z^{\hat{I}} \in R^{H\times W\times K}$Get higher-order features of the image, and finally calculate the contrast loss of pixel-level features based on this feature. The contrast loss of pixel-level features is the pixels with the same labels in the original image and its enhanced image as positive pairs, and the pixels with different labels as negative pairs. By pulling in the distance between the positive pairs in the feature space and pushing the distance between the far negative pairs, the feature extraction network can extract the image features more accurately. This article defines the number of pixels in the image, represents the feature vector of the pixels, the image, the label of the number of pixels, is the symbol function, if the label of the pixels and the pixels in the image is 1, or 0, so the pixel contrast loss can be expressed as:$N^I I z_p^A A p y_p^A A p N_{y_p^I}^{\hat{I}} \hat{I} y_p^I E_{pk}^{AB} A p B k$

$$\mathrm{L}_p = -\frac{1}{N^I}\sum_{p=1}^{N^I}\sum_{q=1}^{N^{\hat{I}}}\frac{E_{pq}^{I\hat{I}}}{N_{y_p^I}^{\hat{I}}}\log\left(\frac{exp\left(z_p^I \cdot z_q^{\hat{I}} / \tau\right)}{\sum_{k=1}^{N^{\hat{I}}} exp\left(z_p^I \cdot z_k^{\hat{I}} / \tau\right)}\right)$$

Therefore, our total loss is:

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_l + \mathcal{L}_p$$

4 Experiment

This section presents the experimental setup, comparison and other methods, and ablation experiments.

4.1 Experimental setup

Verify the proposed method on two publicly available image segmentation datasets: PASCAL-VOC 2012[21]And Cityscapes[22]data set. The PASCAL-VOC2012 dataset consists of 10582 training images, 1449 validation images, and 456 test images with annotations of one background object class and 20 foreground object classes. Cityscapes Data set contains street records from 50 different cities of various stereo video sequence, the data set of fine and coarse two sets of evaluation standard, are generally using the fine evaluation standard, namely by 5000 fine annotation sample set for training and evaluation, sample set including 2975 training map, 500 validation map and 1525 test map, each image size is 1024x2048.The parallel ratio (MIoU) is used to measure the effectiveness of this method.

4.2 Comparison of image segmentation performance based on different numbers of training samples

To verify the effectiveness of the proposed method, this paper compares the image segmentation

performance without pre-trained and multi-trained scale contrast learning based on PASCAL-VOC 2012 and Cityscapes datasets based on different numbers of training samples. On the Cityscapes data set, even when all training maps (2975) are used, the pre-trained image segmentation can be improved by about 10%, while when the training sample is less than 600, the pre-trained based image segmentation performance MIoU can be significantly improved, with an increase of about 12%. On PASCAL-VOC2012 data set, after pre-training model performance has greatly improved, including training samples to one tenth of the original, the pre-training image segmentation performance improvement points more than 34%, training samples reduced to one fifth of the original, the pre-training image segmentation performance improvement points also significantly increased by about 31%. The above results verify the effectiveness of the proposed pre-training method based on multi-scale contrast learning, which has a good segmentation performance in the case of insufficient training samples.

Table 1 Image segmentation performance based on different numbers of training samples

data set	PASCAL-VOC 2012 MIoU				C ityscapes MIoU			
Number of training samples	1059	2118	5295	10582	343	596	1527	2975
random initialization	23.6	30.1	36.4	40.7	48.2	54.6	60.8	63.5
O urs	57.8	61.9	64.2	66.7	64.3	68.4	72.8	73.5

4.3 Ablation experiments

In this paper, three different scales of contrast loss: global contrast loss, local contrast loss, and pixel contrast loss.$\mathcal{L}_g\mathcal{L}_l\mathcal{L}_p$ Table 2 shows the elimination of different contrast loss functions on the performance of downstream image segmentation tasks on the PASCAL-VOC 2012 and Cityscapes datasets. Ablation experiments demonstrate the effectiveness of the proposed multiscale contrast loss function.

Table 2 Performance of different contrasting loss variables

data set Number of training samples	Cityscapes 1527	PASCAL-VOC 2012 5295
No contrast loss	60.8	36.4
$\mathcal{L}_g$	71.1	57.5
$\mathcal{L}_l$	71.3	58.9
$\mathcal{L}_p$	71.6	59.6
$\mathcal{L}_g + \mathcal{L}_l$	71.9	60.7
$\mathcal{L}_g + \mathcal{L}_p$	72.1	61.5
$\mathcal{L}_l + \mathcal{L}_p$	72.3	62.9
$\mathcal{L}_g + \mathcal{L}_l + \mathcal{L}_p$	72.8	64.2

4.4 Method Comparison

To verify the effectiveness of the proposed method, this paper and the existing methods are compared, including the AF of the region-based loss function[23]And RMI[24], And the SimCLR based on global contrast learning[17], All of these methods used ImageNet for pre-training. Table 3 shows the segmentation results of all methods using 2118 training images and 5295 training images on the PASCAL-VOC 2012 dataset, respectively.

The experimental results show that the performance of the proposed method on the PASCAL-VOC 2012 dataset exceeds the previous methods and verifying the effectiveness of the proposed method.

Table 3 Performance of different pre-training methods for image segmentation

method	PASCAL-VOC 2012 2118	PASCAL-VOC 2012 10582
random initialization	30.1	40.7
SimCLR	60.8	64.3
AF	60.4	63.2
RMI	61.1	64.5
O urs	61.9	66.7

5 Conclusion

This paper proposes a pre-training method for image segmentation based on multi-scale contrast learning. Data data data through limited annotation data enhancement, and then trains the feature extractor with multi-scale contrast loss function to extract the accurate feature extractor and finally obtain the image segmentation pre-training model. The pre-trained model can acquire networks that can be used for specific image segmentation tasks. Experiments on two publicly available datasets show that the proposed pre-training method can maximize the use of data labels with limited annotation data, and can also achieve good performance on small-scale image segmentation data sets, which verifies the effectiveness of this method.

Competing interests

No conflict of interest exits in this manuscript.

Consent for publication

Manuscript is approved by all authors for publication.

Availability of data and materials

The data and materials of this experiment are available.

Competing interests

No conflict of interest exits in this manuscript.

Acknowledgements

Not applicable.

Funding:

The work was supported by Key Technologies for Constructing Dynamic Semantic Knowledge Model and Decision Application(524623230003).

Reference:

[1] LeCun Y, Bottou L, Bengio Y, et al.Gradient-based learning applied to document recognition[J].Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[2] Krizhevsky A, Sutskever I, Hinton G E.Imagenet classification with deep convolutional neural networks[J].Communications of the ACM, 2017, 60(6): 84-90.

[3] Szegedy C, Liu W, Jia Y, et al.Going deeper with convolutions[C].IEEE Computer Vision and Pattern Recognition.Boston:IEEE, 2015:1-9.

[4] Simonyan K, Zisserman A.Very Deep Convolutional Networks for Large-Scale Image Recognition[J].Computer ence,2014.

[5] 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.2016: 770-778.

[6] Hadsell R, Chopra S, LeCun Y.Dimensionality reduction by learning an invariant mapping[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06).IEEE, 2006, 2: 1735-1742.

[7] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, Jie Tang.Self-supervised Learning: Generative or Contrastive.IEEE Transactions on Knowledge and Data Engineering, 2021.

[8] Xinlong Wang, Rufeng Zhang, Chunhua Shen1, Tao Kong, Lei Li.Dense Contrastive Learning for Self-Supervised Visual Pre-Training.Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[9] Nanxuan Zhao, Zhirong Wu, Rynson W.H.Lau, Stephen Lin.What Makes Instance Discrimination Good for Transfer Learning?International Conference on Learning Representations, 2022.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick.Momentum Contrast for Unsupervised Visual Representation Learning.Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[11] Zhirong Wu, Yuanjun Xiong, Stella X.Yu, Dahua Lin.Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination.IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H.Richemond, et al.Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning.Advances in Neural Information Processing Systems, 2020.

[13] Xinlei Chen, Kaiming He Exploring Simple Siamese Representation Learning.Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[14] Yonglong Tian, Dilip Krishnan, Phillip Isola.Contrastive Multiview Coding.Proceedings of the European Conference on Computer Vision, 2020

[15] Xie Z, Lin Y, Zhang Z, et al.Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2021: 16684-16693.

[16] Wang F, Wang H, Wei C, et al.CP2: Copy-Paste Contrastive Pretraining for Semantic Segmentation[J].arXiv preprint arXiv:2203.11709, 2022.

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Ge-offrey Hinton.A simple framework for contrastive learning of visual representations.In Proc.Int.Conf.Mach.Learn.,2020.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick.Momentum contrast for unsupervised visual representation learning.In Proc.IEEE Conf.Comp.Vis.Patt.Recogn., 2020.

[19] Zhao X, Vemulapalli R, Mansfield P A, et al.Contrastive learning for label efficient semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision.2021: 10623-10633.

[20] Wang X, Zhang R, Shen C, et al.Dense contrastive learning for self-supervised visual pre-training[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2021: 3024-3033.

[21] Mark Everingham, S.M.Ali Eslami, Luc Van Gool, Christopher K.I.Williams, John M.Winn, and Andrew Zisserman.The pascal visual object classes challenge: A retrospective.Int.J.Comput.Vis.,111(1):98–136, 2015.

[22] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele.The cityscapes dataset for semantic urban scene understanding.In IEEE Conf.Comput.Vis.Pattern Recog., 2016.

[23] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X Yu.Adaptive affinity fields for semantic segmentation.In Proceedings of the European Conference on Computer Vision(ECCV), pages 587–602, 2018.

[24] Shuai Zhao, YangWang, Zheng Yang, and Deng Cai.Region.mutual information loss for semantic segmentation.arXiv preprint arXiv:1910.12037, 2019.