This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# Contrastive Learning for Label Efficient Semantic Segmentation

Xiangyun Zhao[1,2*] zhaoxiangyun915@gmail.com
Raviteja Vemulapalli[2] ravitejavemu@google.com
Philip Andrew Mansfield[2] memes@google.com
Boqing Gong[2] bgong@google.com
Bradley Green[2] brg@google.com
Lior Shapira[2] liorshap@google.com
Ying Wu[1] yingwu@northwestern.edu
[1]Northwestern University [2]Google Research

## Abstract

*Collecting labeled data for the task of semantic segmentation is expensive and time-consuming, as it requires dense pixel-level annotations. While recent Convolutional Neural Network (CNN) based semantic segmentation approaches have achieved impressive results by using large amounts of labeled training data, their performance drops significantly as the amount of labeled data decreases. This happens because deep CNNs trained with the de facto cross-entropy loss can easily overfit to small amounts of labeled data. To address this issue, we propose a simple and effective contrastive learning-based training strategy in which we first pretrain the network using a pixel-wise, label-based contrastive loss, and then fine-tune it using the cross-entropy loss. This approach increases intra-class compactness and inter-class separability, thereby resulting in a better pixel classifier. We demonstrate the effectiveness of the proposed training strategy using the Cityscapes and PASCAL VOC 2012 segmentation datasets. Our results show that pretraining with the proposed contrastive loss results in large performance gains (more than 20% absolute improvement in some settings) when the amount of labeled data is limited. In many settings, the proposed contrastive pretraining strategy, which does not use any additional data, is able to match or outperform the widely-used ImageNet pretraining strategy that uses more than a million additional labeled images.*

## 1. Introduction

In the recent past, various approaches based on Convolutional Neural Networks (CNNs) [6, 8, 64] have reported excellent results on several semantic segmentation datasets by first pretraining their models on the large-scale ImageNet [13] classification dataset and then fine-tuning them with large amounts of pixel-level annotations. This training strategy has several disadvantages: First, collecting a large, pixel-level annotated dataset is time-consuming and expensive. For example, the average time taken to label a single image in the Cityscapes dataset is 90 minutes [11]. Second, the ImageNet dataset can only be used for non-commercial research, making the ImageNet pretraining strategy unsuitable for building real-world products. Collecting a proprietary large-scale classification dataset similar to ImageNet would be expensive and time-consuming. Third, ImageNet pretraining does not necessarily help segmentation of non-natural images, such as medical images [44].

To reduce the need for large amounts of dense pixel-level annotations and the additional large-scale labeled ImageNet dataset, this work focuses on training semantic segmentation models using only a limited number of pixel-level annotated images (no ImageNet dataset). This is challenging since CNN models can easily overfit to limited training data.

Typical semantic segmentation models consist of a deep CNN feature extractor followed by a pixel-wise softmax classifier, and are trained using a pixel-wise cross-entropy loss. While these models perform well when trained with a large number of pixel-level annotated images, their performance drops significantly as the number of labeled training images decreases (see Fig. 1). This happens because CNNs trained with the cross-entropy loss can easily overfit to small amounts of labeled data, as the cross-entropy loss focuses on creating class decision boundaries and does not explicitly encourage intra-class compactness or large margins between classes [15, 37, 49].

To address this issue, we propose to first pretrain the feature extractor using a pixel-wise, label-based contrastive loss (referred to as *contrastive pretraining*), and then fine-tune the entire network including the pixel-wise softmax classifier using the cross-entropy loss (referred to as *softmax fine-tuning*). This approach increases both intra-class compactness and inter-class separability as the label-based contrastive loss [32] encourages the features of pixels from the same class to be close to each other and the features of

*This work was done when Xiangyun Zhao was interning at Google.

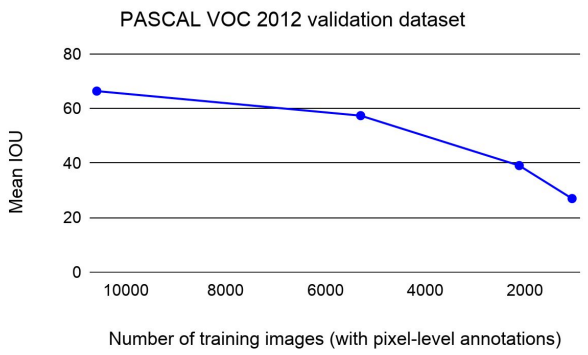


Figure 1. When trained with pixel-wise cross-entropy loss, the performance of a semantic segmentation model drops significantly as the number of labeled training images decreases. Here, we use a DeepLabV3+ [8] model with the ResNet50-based encoder of [7].

pixels from different classes to be far away. The increased intra-class compactness and inter-class separability naturally lead to a better pixel classifier in the fine-tuning stage. Figures 2 and 3 show the distributions of various classes in the softmax input feature spaces of models trained with the cross-entropy loss and the proposed strategy, respectively, using 2118 labeled images from the PASCAL VOC 2012 dataset. The mean IOU values of the corresponding models on the PASCAL VOC 2012 validation dataset are 39.1 and 62.7, respectively. The class support regions are more compact and separated when trained with the proposed strategy, leading to a better performance. We use t-SNE [52] for generating the visualizations.

Various existing semi-supervised and weakly-supervised semantic segmentation approaches also focus on reducing the need for pixel-level annotations by leveraging additional unlabeled images [5, 20, 27, 41] or weaker forms of annotations such as bounding boxes [12, 31, 42, 47] and image-level labels [1, 34, 42]. In contrast to these approaches, the proposed contrastive pretraining strategy does not use any additional data, and is complimentary to them.

Pixel-wise cross-entropy loss ignores the relationships between pixels. To address this issue, region-based loss functions such as region mutual information loss [65] and affinity field loss [30] have been proposed. Different from these loss functions which model pixel relationships in the label space, the proposed contrastive loss models pixel relationships in the feature space. Also, while these loss functions only model relationships between pixels within a local neighborhood, the proposed loss encourages the features of same class pixels to be similar and features of different class pixels to be dissimilar irrespective of their image locations.

Some recent works such as [4, 43, 55, 58, 59] also used pixel-wise contrastive loss for the task of semantic segmentation. However, these works focus on leveraging unlabeled data through self-supervised contrastive learning, and make

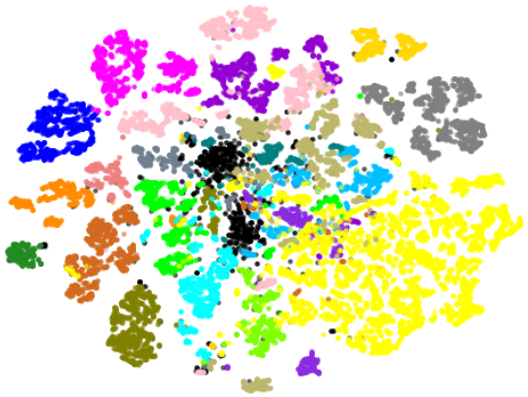

Figure 2. Distribution of various classes in the softmax input feature space of a model trained using only cross-entropy loss on 2118 labeled images from the PASCAL VOC 2012 dataset.

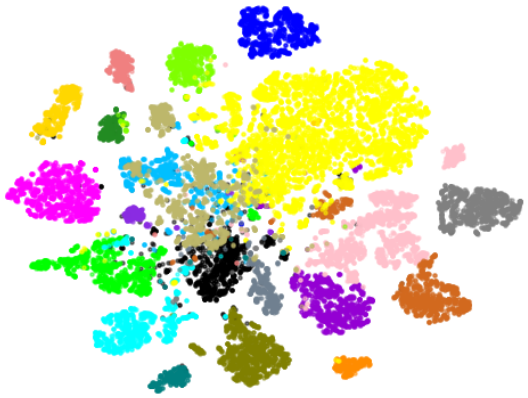

Figure 3. Distribution of various classes in the softmax input feature space of a model trained using the proposed training strategy on 2118 labeled images from the PASCAL VOC 2012 dataset.

use of the labels only in the fine-tuning stage. In contrast, we focus on supervised contrastive learning, and make use of the labels in both pretraining and fine-tuning stages.

We perform experiments on two widely-used semantic segmentation benchmark datasets, namely, Cityscapes and PASCAL VOC 2012, and show that pixel-wise, label-based contrastive pretraining results in large performance gains when the amount of labeled data is limited.

Our main contributions are as follows:

- **New loss functions:** We extend supervised contrastive learning [32] to the task of semantic segmentation. We propose and evaluate three variants of pixel-wise, label-based contrastive loss.
- **Simple training approach:** We propose a simple contrastive learning-based pretraining strategy for improving the performance of semantic segmentation models. We consider the simplicity of our pretraining strategy as its main strength since it can be easily adopted by existing and future semantic segmentation approaches.

- **Strong results:** We show that label-based contrastive pretraining results in large performance gains on two widely-used semantic segmentation datasets when the amount of labeled data is limited. We also show that, in most settings, the proposed contrastive pretraining which does not use any additional data, outperforms the widely-used ImageNet pretraining which uses more than a million additional labeled images.
- **Detailed analyses:** We show visualizations of class distributions in the feature spaces of trained models to provide insights into why the proposed training strategy works better (Fig. 2 and 3). We also present ablation studies that justify our two-stage training strategy.

## 2. Related works

**Self-supervised contrastive learning** These approaches learn representations in a discriminative fashion by contrasting positive pairs against negative pairs. Recently, several approaches based on contrastive loss [22] have been proposed for self-supervised visual representation learning [9, 10, 14, 24, 35, 57, 60]. These approaches treat each instance as a class and use contrastive loss-based instance discrimination for representation learning. Specifically, they use augmented version of an instance to form the positive pair and other randomly sampled instances to form negative pairs for the contrastive loss. Some recent works [29, 45] also explored hard negative mining strategies for contrastive learning. Noting that using a large number of negatives is crucial for the success of contrastive loss-based representation learning, various recent approaches use memory banks to store the representations [24, 51, 57]. Inspired by the effectiveness of self-supervised contrastive learning for image-level recognition tasks, various recent approaches extended it to pixel-level prediction tasks [4, 43, 55, 58, 59].

**Supervised contrastive learning** Recently, [32] proposed supervised contrastive loss for the task of image classification. This loss can be seen as a generalization of the widely-used metric learning losses such as N-pairs [46] and triplet [56] losses to the scenario of multiple positives and negatives generated using class labels. Different from [32], this work focuses on much tougher pixel-level semantic segmentation task, and proposes three variants of pixel-wise, label-based contrastive loss. Since collecting labeled data for the task of semantic segmentation is difficult, we focus on the limited labeled data setting, and show that label-based contrastive learning is highly effective. Concurrent to this work, (still unpublished) [54] also introduced a pixel-wise, label-based contrastive loss for semantic segmentation. However, [54] trains segmentation models using both cross-entropy and contrastive losses simultaneously, which is different from our contrastive pretraining followed by softmax fine-tuning strategy. Our experiments show that the proposed two-stage training is more effective than joint training. Also, [54] does not demonstrate the effectiveness of contrastive learning in the limited labeled data setting.

**Semantic segmentation** Since CNNs have been introduced to solve the semantic segmentation problem [17, 38], several deep CNN-based approaches have been proposed that gradually improved the performance [6, 7, 8, 21, 53, 61, 62, 63, 64] using large amounts of pixel-level annotations. However, collecting dense pixel-level annotations is difficult and costly.

To address this issue, several existing works use the large-scale ImageNet classification dataset for pretraining their models, and also leverage additional weaker forms of supervision such as image-level labels [1, 26, 34, 42], bounding boxes [12, 31, 42, 47], scribbles [36, 50] and points [2], or unlabeled images [5, 19, 39, 40, 48, 67]. In contrast to these approaches, the proposed training strategy does not require any additional data or annotations.

Another relevant line of work includes approaches that use region-based loss functions [30, 65] to model pixel relationships. While [30] uses a pairwise affinity loss based on KL divergence between predicted class probabilities of two pixels, [65] uses a region Mutual Information (MI) loss that maximizes the MI between predicted and groundtruth distributions of patch labels. While these losses model pixel relationships in the label space, the proposed contrastive loss models pixel relationships in the feature space.

A few existing works [3, 18, 23, 33] use metric learning based on independent pairwise similarity and dissimilarity losses for the tasks of semantic and instance segmentation. However, these works only model relationships between pixels within a local image neighborhood or an object instance. Different from these works, the proposed contrastive loss models relationships between pixels irrespective of their image locations, and contrasts a similar pair with a large number of dissimilar pairs. Also, these works do not demonstrate the effectiveness of contrastive pretraining in the limited labeled data setting.

Recently, [28] proposed to train the feature extractor of a semantic segmentation model by maximizing the log likelihood of extracted pixel features under a mixture of vMF distributions model. During inference, they first segment the pixel features using spherical K-Means clustering, and then perform k-nearest neighbor search for each segment to retrieve labels from segments in the training set. While this approach is shown to improve the performance when compared to the widely-used pixel-wise softmax training, it is very complicated as it uses a two-stage expectation-maximization algorithm for training. In comparison, the proposed approach is simple, and can be easily adopted by existing and future semantic segmentation approaches.

## 3. Proposed approach

### 3.1. Pixel-wise label-based contrastive loss

In this work, we extend supervised contrastive learning to pixel-level tasks such as semantic segmentation, and propose three pixel-wise, label-based contrastive losses to pretrain a semantic segmentation model.

Let $I$ denote an image and $\hat{I}$ its distorted version (e.g., color jittering). Let $y_p^I$ denote the class label of pixel $p$ in $I$, $N_c^I$ denote the number of pixels in $I$ with class label $c$, and $N^I$ denote the total number of pixels in $I$. Let $f_p^I$ be a $d$-dimensional, unit-normalized feature extracted from $I$ at pixel $p$, Let $\mathbb{1}_{pk}^{AB} = \mathbb{1}\left[y_p^A = y_k^B\right]$ and $e_{pk}^{AB} = exp\left(f_p^A \cdot f_k^B / \tau\right)$, where $\tau$ is a temperature parameter.

**Within-image loss** Our within-image, pixel-wise, label-based contrastive loss which encourages features of pixels in an image to cluster according to their labels is given by

$$-\frac{1}{N^I}\sum_{p=1}^{N^I}\frac{1}{N_{y_p^I}^{\hat{I}}}\sum_{q=1}^{N^{\hat{I}}}\mathbb{1}_{pq}^{I\hat{I}}\,\mathbf{log}\left(\frac{e_{pq}^{I\hat{I}}}{\sum\limits_{k=1}^{N^{\hat{I}}}e_{pk}^{I\hat{I}}}\right), \tag{1}$$

In our experiments, image $\hat{I}$ is generated from $I$ by applying distortions with probability $p = 0.8$. Hence, for some samples in a minibatch the contrastive loss is between features of original and distorted pixels, and for other samples, the loss is between features of original pixels. We compute this contrastive loss separately for each image, and then average it across the images in a minibatch.

**Cross-image loss** Our cross-image, pixel-wise, label-based contrastive loss extends the within-image loss (1) by using additional positives from another image $J$. Positive pixels from $J$ can be interpreted as harder positives since they come from a different image. We do not use additional negatives from $J$ since negatives from a different image can be interpreted as easier negatives [1]. The cross-image loss for an image pair $I$ and $J$ is given by

$$-\frac{1}{N^I}\sum_{p=1}^{N^I}\sum_{q=1}^{N^{\hat{I}}}\frac{\mathbb{1}_{pq}^{I\hat{I}}}{N_{y_p^I}^{\hat{I}}+N_{y_p^I}^{\hat{J}}}\,\mathbf{log}\left(\frac{e_{pq}^{I\hat{I}}}{\sum\limits_{k=1}^{N^{\hat{I}}}e_{pk}^{I\hat{I}}+\sum\limits_{k=1}^{N^{\hat{J}}}\mathbb{1}_{pk}^{I\hat{J}}e_{pk}^{I\hat{J}}}\right)$$
$$-\frac{1}{N^I}\sum_{p=1}^{N^I}\sum_{q=1}^{N^{\hat{J}}}\frac{\mathbb{1}_{pq}^{I\hat{J}}}{N_{y_p^I}^{\hat{I}}+N_{y_p^I}^{\hat{J}}}\,\mathbf{log}\left(\frac{e_{pq}^{I\hat{J}}}{\sum\limits_{k=1}^{N^{\hat{I}}}e_{pk}^{I\hat{I}}+\sum\limits_{k=1}^{N^{\hat{J}}}\mathbb{1}_{pk}^{I\hat{J}}e_{pk}^{I\hat{J}}}\right). \tag{2}$$

[1] When we experimented with adding negatives from another image $J$, we observed some performance drop in our initial experiments.

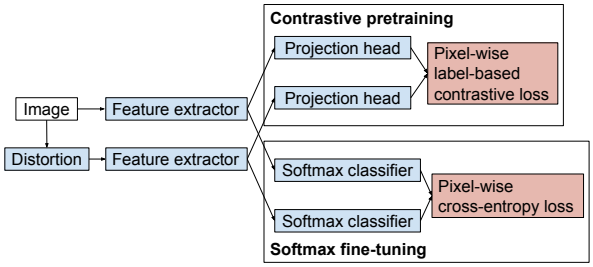


Figure 4. Proposed training strategy: Pixel-wise supervised contrastive pretraining followed by softmax fine-tuning.

For each image in a minibatch, we compute the cross-image contrastive loss by pairing the image with another random image from the minibatch, and average the loss across all the images.

**Batch loss** We also considered a batch variant that treats all the pixels in a minibatch as a single bag of pixels for computing the contrastive loss (1). While one would expect this batch variant to outperform within-image and cross-image variants (due to interactions across multiple images), our experimental results indicate the opposite. Please see Section 4.4 for further details.

### 3.2. Proposed training strategy

Typical semantic segmentation models consist of a deep CNN feature extractor followed by a pixel-wise softmax classifier. We first pretrain the CNN feature extractor from scratch with a pixel-wise, class label-based contrastive loss. Following [9, 10, 32] we use a projection head while training with contrastive loss, i.e., the features $f_p^I$ used in loss (1) and (2) are the outputs of a projection head rather than the original feature extractor (see Fig. 4). After contrastive pretraining, we discard the projection head, add a pixel-wise softmax classifier on top of the feature extractor, and fine-tune the entire network with pixel-wise cross-entropy loss.

Note that there is no interaction between pixels from different images in the within-image contrastive loss. So, it is crucial to train the entire network in the fine-tuning stage. While within-image loss-based contrastive pretraining encourages pixels within an image to cluster according to their labels, softmax fine-tuning rearranges these clusters so that they fall on the correct side of the decision boundary.

## 4. Experiments

### 4.1. Datasets and metrics

**PASCAL VOC 2012** [16]: This dataset consists of 10,582 training, 1,449 validation, and 456 test images with annotations for one background and 20 foreground object classes. The performance is measured in terms of pixel Intersection-Over-Union (IOU) averaged across the 21 classes.

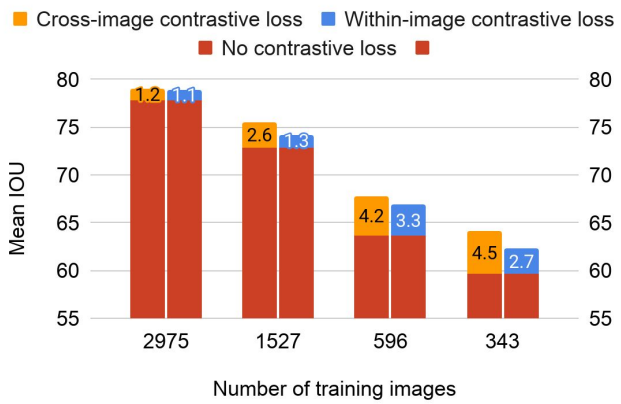


Figure 5. Improvement on the Cityscapes dataset due to contrastive pretraining.

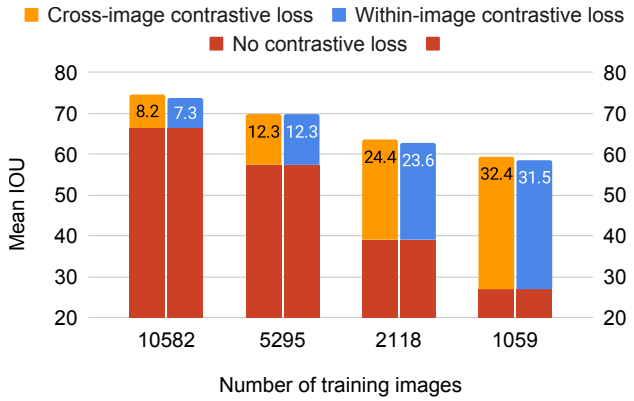


Figure 6. Improvement on the PASCAL VOC 2012 dataset due to contrastive pretraining.

**Cityscapes** [11]: This dataset consists of 2975 training, 500 validation, and 1525 test images. Following the evaluation protocol of [6], 19 semantic labels are used for evaluation, and the void label is ignored. The performance is measured in terms of pixel IOU averaged across the 19 classes.

All the results reported in this section correspond to the validation splits of these datasets. Please refer to the supplementary material for results on the test splits.

### 4.2. Model architecture

Our feature extractor follows DeepLabV3+ [8] encoder-decoder architecture with the ResNet50-based encoder of DeepLabV3 [7]. The output spatial resolution of the feature extractor is four times lower than the input resolution. Our projection head consists of three $1 \times 1$ convolution layers with 256 channels followed by a unit-normalization layer. The first two layers in the projection head use the ReLU activation function.

### 4.3. Training and inference

Following [7, 8], we use $513 \times 513$ random crops extracted from preprocessed (random left-right flipping and scaling) input images for training. All the models are trained from scratch using stochastic gradient descent on 8 replicas with minibatches of size 16, momentum of 0.9, weight decay of $4e^{-5}$, and cosine learning rate decay. When we use softmax training without contrastive pretraining, we use an initial learning rate of 0.03 and 600K [2] training steps when the number of labeled images is above 2500 in the case of PASCAL VOC 2012 dataset and above 1000 in the case of Cityscapes dataset, and 300K training steps in other settings[3]. For contrastive pretraining, we use an initial learning rate of 0.1 and 300K training steps. For softmax fine-tuning after contrastive pretraining, we use an initial learning rate of 0.007 and 300K training steps except when the number of labeled images is below 2500 in the case of PASCAL VOC 2012. In this case, we use 50K training steps[3]. The temperature $\tau$ of contrastive loss is set to 0.07. We use color distortions from [9] for contrastive pretraining, and random brightness and contrast adjustments for softmax fine-tuning [4].

For $513 \times 513$ input, our feature extractor produces a $129 \times 129$ feature map. Since the memory complexity of our contrastive loss is quadratic in the number of pixels, to avoid GPU memory issues, we resize the feature map to $65 \times 65$ using bilinear resizing before computing the contrastive loss. The corresponding low-resolution label map is obtained from the original label map using nearest neighbor downsampling. For softmax training, we follow [8] and upsample the logits from $129 \times 129$ to $513 \times 513$ using bilinear resizing before computing the pixel-wise cross entropy loss.

Since the model is fully-convolutional, during inference, we directly run it on an input image and upsample the output logits to input resolution using bilinear resizing.

### 4.4. Performance gain by contrastive pretraining

Figures 5 and 6 show the performance improvements on the validation splits of Cityscapes and PASCAL VOC 2012 datasets, respectively, obtained by contrastive pretraining. Both within-image and cross-image contrastive loss-based pretraining consistently improve the performance on both the datasets for different amounts of training data. On the Cityscapes dataset, we see large gains (more than 4 points) when the number of labeled training images is less than 600, and a decent gain (1.2 points) even when using the entire training set of 2975 images. On the PASCAL VOC 2012 dataset, we see huge gains (up to about 30 points) for all label counts, and we are able to reduce the labeling re-

[2] 600K steps were chosen after trying 300K, 600K and 1M with different learning rates.

[3] We observed overfitting with longer training when the number of labeled images is low.

[4] Using hue and saturation adjustments from [9] while training the softmax classifier resulted in a slight drop in the performance.

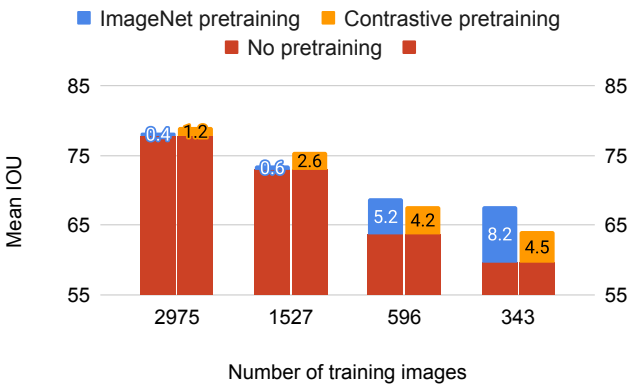


Figure 7. Improvement on the Cityscapes dataset due to different pretraining strategies.

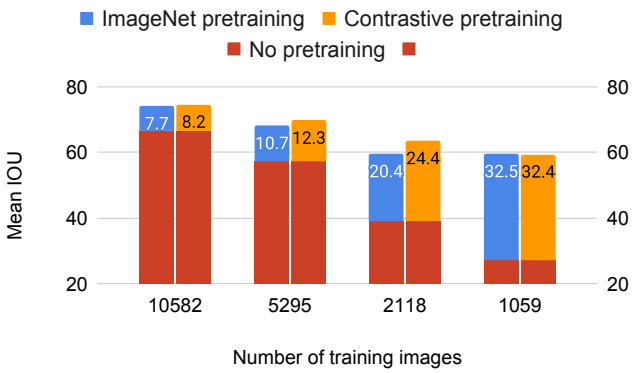


Figure 8. Improvement on the PASCAL VOC 2012 dataset due to different pretraining strategies.

quirements by $2\times$ while improving the performance. Notably, by using 1059 images, we are able to outperform the model trained using only cross-entropy loss on $5\times$ more data (5295 images). These results clearly demonstrate the effectiveness of the proposed contrastive pretraining.

Cross-image contrastive loss outperforms within-image contrastive loss since it makes use of positives from other images which can be seen as harder positives when compared to within-image positives. In most of the settings, the cross-image loss outperforms the within-image loss by 0.8 or more points. Specifically, on the Cityscapes dataset, the cross-image loss outperforms the within-image loss by 1.8 points when only 343 labeled images are available.

We also conducted experiments with the batch variant of our contrastive loss which considers all the pixels in a minibatch as a single bag of pixels for computing the loss (1). Note that the memory complexity of loss (1) is quadratic in the number of pixels. Hence, to avoid GPU memory issues, we randomly sample 10K pixels from the entire minibatch for computing the batch contrastive loss. Table 1 compares the performance of the batch variant with the other two variants. While one would expect the batch variant to perform better because of the pixel interactions across multiple images, the results indicate the opposite. Note that, while we have contrastive loss terms for every pixel in the within-image and cross-image variants, only a subset of pixels contribute to the loss in the batch variant. We believe this to be the reason for the poor performance of the batch variant. In the near future, we plan to explore hybrid variants that will have loss terms for as many pixels as possible while still forming pixel pairs across multiple images.

The performance improvements seen on the PASCAL VOC 2012 dataset are much higher than those seen on the Cityscapes dataset. We conjecture that this is because of the presence of an additional background category in the PASCAL VOC 2012 dataset. This category is comprised of diverse visual content from a wide variety of object classes

Table 1. Performance of different contrastive loss variants.

| Dataset | Cityscapes | PASCAL VOC |
|---|---|---|
| Training images | 596 images | 2118 images |
| No contrastive loss | 63.6 | 39.1 |
| Batch loss | (↑ 2.4) 66.0 | (↑ 22.6) 61.7 |
| Within-image loss | (↑ 3.3) 66.9 | (↑ 23.6) 62.7 |
| Cross-image loss | (↑ **4.2**) **67.8** | (↑ **24.4**) **63.5** |

(everything other than the 20 foreground object classes). Hence, explicit enforcement of intra-class compactness and inter-class separability by contrastive pretraining is helping more in the case of PASCAL VOC 2012 dataset. To verify our conjecture experimentally, we trained the segmentation model on PASCAL VOC 2012 dataset (2118 labeled images) ignoring the background category. When evaluated on the foreground categories, within-image loss-based contrastive pretraining improved the mean IOU by 3.6 points, which is much lower than the 23.6 points gain achieved in the presence of background category. This suggests that the presence of additional background category is contributing to the huge gains on the PASCAL VOC 2012 dataset.

### 4.5. Comparison with ImageNet pretraining

Most of the existing semantic segmentation approaches pretrain their models on the large-scale ImageNet classification dataset [13] to achieve state-of-the-art results. Even works such as [25, 66] which show that ImageNet pretraining can be omitted for the task of object detection on some datasets, acknowledge that ImageNet pretraining is important for semantic segmentation. While it can lead to significant performance gains, ImageNet pretraining may not be used for building commercial products. Collecting such a large-scale proprietary dataset is also time-consuming and expensive. In contrast, the proposed strategy achieves performance gains without using any additional data. Figures 7 and 8 compare the performances of ImageNet-pretrained

Table 2. Comparison of different training strategies.

| Strategy | Cityscapes (596 images) | PASCAL VOC (2118 images) |
|---|---|---|
| Only cross-entropy | 63.6 | 39.1 |
| Joint training | (↑ 1.0) 64.6 | (↑ 0.7) 39.8 |
| Proposed | (↑ **3.3**) **66.9** | (↑ **23.6**) **62.7** |

Table 3. Effect of distortions for contrastive pretraining.

| Distortions | Cityscapes (596 images) | PASCAL VOC (2118 images) |
|---|---|---|
| ✗ | 66.5 | 62.9 |
| ✓ | (1.0 ↑) 67.5 | 62.7 |

and contrastive-pretrained models. Contrastive pretraining matches or outperforms ImageNet pretraining in most of the cases except when the number of labeled images is below 600 for the Cityscapes dataset. This is a significant result given that the proposed contrastive pretraining does not use any additional data and ImageNet pretraining uses more than a million additional labeled images.

## 4.6. Ablation studies

In this section, we perform ablation studies on the Cityscapes (596 training images) and PASCAL VOC 2012 (2118 training images) datasets using the within-image contrastive loss.

### 4.6.1 Joint training

The proposed approach first pretrains the feature extractor using a label-based contrastive loss, and then fine-tunes the entire network using the cross-entropy loss. An alternative training strategy is to train with both losses at the same time. Table. 2 compares these two training strategies. While joint training with both losses performs slightly better than training with only cross-entropy loss, it performs significantly worse when compared to the proposed approach [5].

### 4.6.2 Importance of distortions for contrastive loss

In the case of contrastive loss-based self-supervised learning, distortions are necessary to generate positive pairs [9]. But, in the case of label-based contrastive learning, positive pairs can be generated using labels, and hence, it is unclear how important distortions are. In this work, we use the color distortions from a recent self-supervised learning method [9] that worked well for the downstream task of image classification. Table 3 shows the effect of using these distortions in the contrastive pretraining stage. We can see a small performance gain on the Cityscapes dataset and no gain on the PASCAL VOC 2012 dataset [6]. These results suggest that distortions that work well for image recognition may not work for semantic segmentation. This also warrants a careful study of various distortions to find the ones that are most suitable for the task of semantic segmentation.

[5] Contrastive loss weight for joint training was chosen using grid search.

Table 4. Contrastive pretraining with OCR [61] approach.

| Contrastive pretraining | Cityscapes (343 images) | Cityscapes (596 images) |
|---|---|---|
| ✗ | 57.1 | 62.2 |
| ✓ | (↑ 6.3) 63.4 | (↑ 3.4) 65.6 |

## 4.7. Additional results

### 4.7.1 Contrastive pretraining with OCR [61] approach

While we use DeepLabV3+ [8] as the baseline model in all our experiments, the proposed contrastive pretraining strategy can be easily adopted by other existing or future semantic segmentation models. To demonstrate this, we combine (within-image) contrastive loss-based pretraining strategy with the recent OCR [61] approach using the code provided by its authors [7]. Table 4 shows the corresponding results on the Cityscapes dataset. Contrastive pretraining leads to significant performance gains demonstrating that it can be effective with multiple segmentation models.

Note that the OCR [61] approach performs worse than our DeepLab V3+ baseline (62.2 vs 63.6 for 596 training images and 57.1 vs 59.6 for 343 training images). This may be because the OCR model has more learnable parameters, and is more prone to overfitting when the number of training images is low.

### 4.7.2 Comparison with region-based loss functions

As pixel-wise cross-entropy loss ignores the relationships between pixels, region-based loss functions [30, 65] have been proposed which model pixel relationships in the label space. Different from these loss functions, the proposed contrastive loss models pixel relationships in the feature space. Table 5 compares the proposed approach with [30, 65]. For fair comparison, we train [30, 65] with ResNet50-based DeepLabV3+ model using the code provided by authors of [65] [8]. The proposed training approach (which does not use any additional data) clearly outperforms the region-based loss functions [30, 65] even when they use ImageNet pretraining. This suggests that modeling pixel relationships with a loss in the feature space is more effective than modeling with losses in the label space.

[6] Differences lower than 0.5 are too small to draw any conclusion.

[7] https://github.com/openseg-group/openseg.pytorch?v=2

[8] https://github.com/ZJULearning/RMI

Table 5. Comparison with region loss-based approaches. Here, IN and CT refer to ImageNet and contrastive pretraining, respectively.

| Approach | Pretraining | | PASCAL VOC | |
|---|---|---|---|---|
| | IN | CT | (1059 images) | (2118 images) |
| AF [30] | ✗ | ✗ | 27.8 | 43.0 |
| | ✓ | ✗ | 57.4 | 60.4 |
| RMI [65] | ✗ | ✗ | 27.9 | 37.5 |
| | ✓ | ✗ | 58.0 | 61.9 |
| Proposed | ✗ | ✓ | 59.4 | 63.5 |

#### 4.7.3 Comparison with self-supervised learning

Recently, [55, 58, 59] explored pixel-wise self-supervised contrastive learning for the task of semantic segmentation. They first pretrain their networks on the ImageNet dataset using pixel-wise self-supervised contrastive loss, and then fine-tune them end-to-end on the target semantic segmentation dataset. Using the full labeled dataset, [55] reported a mean IOU of 69.4 for the PASCAL VOC 2012 dataset, and [55], [58] and [59] reported mean IOUs of 75.7, 76.5 and 77.2, respectively, for the Cityscapes dataset. In comparison, using the full labeled dataset, we achieve 79.0 and 74.6 on the Cityscapes and PASCAL VOC 2012 datasets, respectively [9]. In fact, we match the results of [55] by using 50% fewer labeled images (69.7 for PASCAL using 5.9K images and 75.5 for Cityscapes using 1.5K images).

#### 4.7.4 Semi-supervised setting

While we mainly focus on the supervised setting, the proposed training strategy can be easily extended to the semi-supervised setting where we have access to additional unlabeled images. We demonstrate this using a simple pseudo labeling strategy. In this setting, we first train a model with labeled images using the proposed approach. Then, we generate pseudo labels for the unlabeled images by running the trained model on them and thresholding the output predictions. Specifically, we assign a pixel to a class if that class receives the highest score for that pixel and that score is above a threshold T [10]. If the model does not produce a score above T for any of the classes for a pixel, then that pixel is ignored. Once we generate pseudo labels for the unlabeled images, we retrain the model on both labeled and pseudo-labeled images using the proposed approach.

Table 6 compares the proposed approach with the recent semi-supervised CCT [41] approach which has been shown to outperform several existing semi-supervised and weakly-supervised approaches. The proposed approach with contrastive pretraining outperforms CCT by a huge margin (25-35 points). When we train the proposed pseudo label-based semi-supervised approach only with cross-entropy loss, i.e., no contrastive pretraining, its performance is similar to CCT verifying that the gap between CCT and the proposed approach is mainly due to contrastive pretraining despite some architectural differences between the networks [11]. Also, the proposed approach performs competitively when compared to ImageNet-pretrained CCT, reaffirming the effectiveness of supervised contrastive pretraining.

Please refer to the supplementary material for additional results in the semi-supervised setting.

[9] This is not necessarily a fair comparison since the network architectures used by [55, 58, 59] are different from ours, and these approaches use additional unlabeled ImageNet dataset.

[10] For our experiments, we use a threshold of 0.8 for all the foreground classes of the PASCAL VOC 2012 and Cityscapes datasets, and a threshold of 0.97 for the background class of the PASCAL VOC 2012 dataset.

Table 6. Performance in the semi-supervised setting. Here, IN and CT refer to ImageNet and contrastive pretraining, respectively.

| Approach | Pretraining | | PASCAL VOC | |
|---|---|---|---|---|
| | IN | CT | (1059 images) | (2118 images) |
| CCT [41] | ✗ | ✗ | 27.1 | 40.3 |
| Proposed | ✗ | ✗ | 28.6 | 41.4 |
| Proposed | ✗ | ✓ | 60.4 | 65.2 |
| CCT [41] | ✓ | ✗ | 62.9 | 65.1 |

## 5. Conclusions and future work

Deep CNN-based semantic segmentation models trained with cross-entropy loss perform poorly when trained with limited labeled data. To address this issue, we proposed a simple and effective contrastive learning-based training strategy in which we first pretrain the feature extractor of the model using a pixel-wise label-based contrastive loss and then fine-tune the entire network including the softmax classifier using the cross-entropy loss. This training approach increases both intra-class compactness and inter-class separability, thereby enabling a better pixel classifier. We performed experiments on PASCAL VOC 2012 and Cityscapes datasets, and achieved large performance gains by using contrastive pretraining. Specifically, in many settings, the proposed contrastive pretraining strategy which does not use any additional data, matches or outperforms the widely-used supervised ImageNet pretraining strategy.

In this work, we used a pseudo labeling-based approach to leverage unlabeled images. In the future, we plan to explore the proposed contrastive loss in conjunction with consistency-based loss functions [20, 40, 41] which are commonly-used for semi-supervised learning.

**Acknowledgements** We thank Yukun Zhu and Liang-Chieh Chen from Google Research for their support with the DeepLab codebase.

[11] While CCT uses PSPNet [64], we use DeeplabV3+ model [8]

## References


[1] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 3

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *Eur. Conf. Comput. Vis.*, 2016. 3

[3] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *CoRR*, abs/1708.02551, 2017. 3

[4] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Adv. Neural Inform. Process. Syst.*, 2020. 2, 3

[5] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *Eur. Conf. Comput. Vis.*, 2020. 2, 3

[6] 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 Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 1, 3, 5

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 2, 3, 5

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Eur. Conf. Comput. Vis.*, 2018. 1, 2, 3, 5, 7, 8

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. 3, 4, 5, 7

[10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. *CoRR*, abs/2006.10029, 2020. 3, 4

[11] 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. 1, 5

[12] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Int. Conf. Comput. Vis.*, 2015. 2, 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009. 1, 6

[14] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Adv. Neural Inform. Process. Syst.*, 2014. 3

[15] Gamaleldin F. Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In *Adv. Neural Inform. Process. Syst.*, 2018. 1

[16] 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. 4

[17] Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1915–1929, 2013. 3

[18] Alireza Fathi, Zbigniew Wojna, Vivek Rathod, Peng Wang, Hyun Oh Song, Sergio Guadarrama, and Kevin P. Murphy. Semantic instance segmentation via deep metric learning. *CoRR*, abs/1703.10277, 2017. 3

[19] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum. *CoRR*, abs/2004.08514, 2020. 3

[20] Geoffrey French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham D. Finlayson. Consistency regularization and cutmix for semi-supervised semantic segmentation. *CoRR*, abs/1906.01916, 2019. 2, 8

[21] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3

[22] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2006. 3

[23] Adam W. Harley, Konstantinos G. Derpanis, and Iasonas Kokkinos. Segmentation-aware convolutional networks using local attention masks. In *Int. Conf. Comput. Vis.*, 2017. 3

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 3

[25] Kaiming He, Ross B. Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Int. Conf. Comput. Vis.*, 2019. 6

[26] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3

[27] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *Brit. Mach. Vis. Conf.*, 2018. 2

[28] Jyh-Jing Hwang, Stella X. Yu, Jianbo Shi, Maxwell D. Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *Int. Conf. Comput. Vis.*, 2019. 3

[29] Yannis Kalantidis, Mert Bülent Sariyildiz, Noé Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Adv. Neural Inform. Process. Syst.*, 2020. 3

[30] 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. 2, 3, 7, 8

[31] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2, 3

[32] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *CoRR*, abs/2004.11362, 2020. 1, 2, 3, 4

[33] Shu Kong and Charless C. Fowlkes. Recurrent pixel embedding for instance grouping. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3

[34] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2, 3

[35] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. *CoRR*, abs/2005.04966, 2020. 3

[36] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3

[37] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Int. Conf. Mach. Learn.*, 2016. 1

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 3

[39] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high- and low-level consistency. *CoRR*, abs/1908.05724, 2019. 3

[40] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. *CoRR*, abs/2007.07936, 2020. 3, 8

[41] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 8

[42] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Int. Conf. Comput. Vis.*, 2015. 2, 3

[43] Pedro O. Pinheiro, Amjad Almahairi, Ryan Y. Benmalek, Florian Golemo, and Aaron C. Courville. Unsupervised learning of dense visual representations. In *Adv. Neural Inform. Process. Syst.*, 2020. 2, 3

[44] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *arXiv preprint arXiv:1902.07208*, 2019. 1

[45] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *CoRR*, abs/2010.04592, 2020. 3

[46] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Adv. Neural Inform. Process. Syst.*, 2016. 3

[47] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3

[48] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Int. Conf. Comput. Vis.*, 2017. 3

[49] Shizhao Sun, Wei Chen, Liwei Wang, and Tie-Yan Liu. Large margin deep neural networks: Theory and algorithms. *CoRR*, abs/1506.05232, 2015. 1

[50] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised CNN segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3

[51] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *CoRR*, abs/1906.05849, 2019. 3

[52] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9(86):2579–2605, 2008. 2

[53] Raviteja Vemulapalli, Oncel Tuzel, Ming-Yu Liu, and Rama Chellappa. Gaussian conditional random field network for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3

[54] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *CoRR*, abs/2101.11939, 2021. 3

[55] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. *CoRR*, abs/2011.09157, 2020. 2, 3, 8

[56] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, 2009. 3

[57] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3

[58] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. *CoRR*, abs/2102.04803, 2021. 2, 3, 8

[59] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *CoRR*, abs/2011.10043, 2020. 2, 3, 8

[60] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3

[61] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Eur. Conf. Comput. Vis.*, 2020. 3, 7

[62] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *CoRR*, abs/1809.00916, 2018. 3

[63] Hang Zhang, Han Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3

[64] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1, 3, 8

[65] Shuai Zhao, Yang Wang, Zheng Yang, and Deng Cai. Region mutual information loss for semantic segmentation. *arXiv preprint arXiv:1910.12037*, 2019. 2, 3, 7, 8

[66] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. In *Adv. Neural Inform. Process. Syst.*, 2020. 6

[67] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *CoRR*, abs/2010.09713, 2020. 3