Exploring Cross-Image Pixel Contrast for Semantic Segmentation

Wenguan Wang[1*], Tianfei Zhou[1*], Fisher Yu[1], Jifeng Dai[2], Ender Konukoglu[1], Luc Van Gool[1]
[1] Computer Vision Lab, ETH Zurich [2] SenseTime Research

https://github.com/tfzhou/ContrastiveSeg

Abstract

Current semantic segmentation methods focus only on mining "local" context, i.e.*, dependencies between pixels within individual images, by context-aggregation modules (*e.g.*, dilated convolution, neural attention) or structure-aware optimization criteria (*e.g.*, IoU-like loss). However, they ignore "global" context of the training data,* i.e.*, rich semantic relations between pixels across different images. Inspired by recent advance in unsupervised contrastive representation learning, we propose a pixel-wise contrastive algorithm for semantic segmentation in the fully supervised setting. The core idea is to enforce pixel embeddings belonging to a same semantic class to be more similar than embeddings from different classes. It raises a pixel-wise metric learning paradigm for semantic segmentation, by explicitly exploring the structures of labeled pixels, which were rarely explored before. Our method can be effortlessly incorporated into existing segmentation frameworks without extra overhead during testing. We experimentally show that, with famous segmentation models (*i.e.*, DeepLabV3, HRNet, OCR) and backbones (*i.e.*, ResNet, HRNet), our method brings performance improvements across diverse datasets (*i.e.*, Cityscapes, PASCAL-Context, COCO-Stuff, CamVid). We expect this work will encourage our community to rethink the current de facto training paradigm in semantic segmentation.*

1. Introduction

Semantic segmentation, which aims to infer semantic labels for all pixels in an image, is a fundamental problem in computer vision. In the last decade, semantic segmentation has achieved remarkable progress, driven by the availability of large-scale datasets (*e.g.*, Cityscapes [15]) and rapid evolution of convolutional networks (*e.g.*, VGG [63], ResNet [32]) as well as segmentation models (*e.g.*, fully convolutional network (FCN) [51]). In particular, FCN [51] is the cornerstone of modern deep learning techniques for segmentation, due to its unique advantage in end-to-end pixel-wise representation learning. However, its spatial invariance nature hinders the ability of modeling useful context among pixels (within images). Thus a main stream of subsequent effort delves into network designs for effective context aggregation, *e.g.*, dilated convolution [80, 8, 9], spatial pyramid pooling [84], multi-layer feature fusion [58, 47] and neural attention [35, 24]. In addition, as the widely adopted pixel-wise cross entropy loss fundamentally lacks the spatial discrimination power, some alternative optimization criteria are proposed to explicitly address object structures during segmentation network training [40, 2, 86].

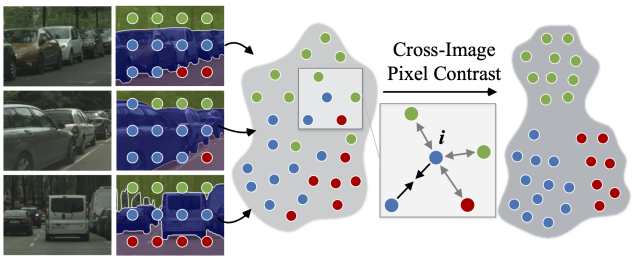


Figure 1: **Main idea.** Current segmentation models learn to map pixels (b) to an embedding space (c), yet ignoring intrinsic structures of labeled data (*i.e.*, inter-image relations among pixels from a same class, noted with same color in (b)). Pixel-wise contrastive learning is introduced to foster a new training paradigm (d), by *explicitly* addressing intra-class compactness and inter-class dispersion. Each pixel (embedding) i is pulled closer (→←) to pixels (●) of the same class, but pushed far (←→) from pixels (● ●) from other classes. Thus a better-structured embedding space (e) is derived, eventually boosting the performance of segmentation models.

Basically, these segmentation models (excluding [37]) utilize deep architectures to project image pixels into a highly non-linear embedding space (Fig. 1(c)). However, they typically learn the embedding space that only makes use of "local" context around pixel samples (*i.e.*, pixel dependencies *within* individual images), but ignores "global" context of the whole dataset (*i.e.*, pixel semantic relations *across* images). Hence, an essential issue has been long ignored in the field: *what should a good segmentation embedding space look like*? Ideally, it should not only **1)** address the categorization ability of individual pixel embeddings, but also **2)** be well structured to address intra-class compactness and inter-class dispersion. With regard to **2)**, pixels

*The first two authors contribute equally to this work.

from a same class should be closer than those from different classes, in the embedding space. Prior studies [49, 60] in representation learning also suggested that encoding intrinsic structures of training data (*i.e.*, **2**)) would facilitate feature discriminativeness (*i.e.*, **1**)). So we speculate that, although existing algorithms have achieved impressive performance, it is possible to learn a better structured pixel embedding space by considering both **1)** and **2)**.

Recent advance in unsupervised representation learning [12, 31] can be ascribed to the resurgence of contrastive learning – an essential branch of deep metric learning [39]. The core idea is "learn to compare": given an *anchor* point, distinguish a similar (or *positive*) sample from a set of dissimilar (or *negative*) samples, in a projected embedding space. Especially, in the field of computer vision, the contrast is evaluated based on image feature vectors; the augmented version of an anchor image is viewed as a positive, while all the other images in the dataset act as negatives.

The great success of unsupervised contrastive learning and our aforementioned speculation together motivate us to rethink the current de facto training paradigm in semantic segmentation. Basically, the power of unsupervised contrastive learning roots from the structured comparison loss, which takes the advantage of the context within the training data. With this insight, we propose a pixel-wise contrastive algorithm for more effective dense representation learning in the **fully supervised setting**. Specifically, in addition to adopting the pixel-wise cross entropy loss to address class discrimination (*i.e.*, property **1**)), we utilize a pixel-wise contrastive loss to further shape the pixel embedding space, through exploring the structural information of labeled pixel samples (*i.e.*, property **2**)). The idea of the pixel-wise contrastive loss is to compute *pixel-to-pixel contrast*: enforce embeddings to be similar for positive pixels, and dissimilar for negative ones. As the pixel-level categorical information is given during training, the positive samples are the pixels belonging to a same class, and the negatives are the pixels from different classes (Fig. 1(d)). In this way, the global property of the embedding space can be captured (Fig. 1(e)) for better reflecting intrinsic structures of training data and enabling more accurate segmentation predictions.

With our supervised pixel-wise contrastive algorithm, two novel techniques are developed. **First**, we propose a *region* memory bank to better address the nature of semantic segmentation. Faced with huge amounts of highly-structured pixel training samples, we let the memory store pooled features of semantic regions (*i.e.*, pixels with a same semantic label from a same image), instead of pixel-wise embeddings only. This leads to *pixel-to-region contrast*, as a complementary for the pixel-to-pixel contrast strategy. Such memory design allows us to access more representative data samples during each training step and fully explore structural relations between pixels and semantic-level segments, *i.e.*, pixels and segments belonging to a same class should be close in the embedding space. **Second**, we propose different sampling strategies to make better use of informative samples and let the segmentation model pay more attention to those segmentation-hard pixels. Previous works have confirmed that hard negatives are crucial for metric learning [39, 60, 62], and our study further reveals the importance of mining both informative negatives/positives and anchors in this supervised, dense image prediction task.

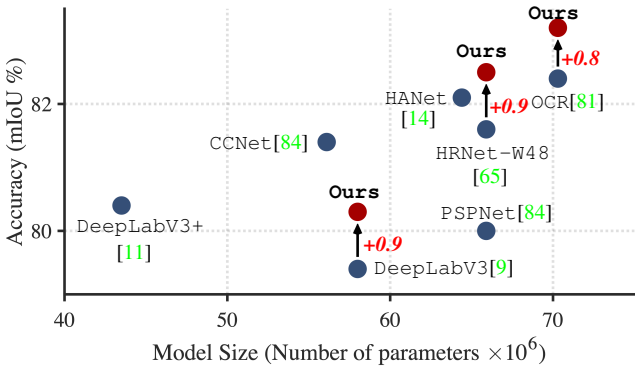


Figure 2: **Accuracy *vs.* model size** on Cityscapes `test` [15]. Our contrastive enables consistent performance improvements over state-of-the-arts, *i.e.*, DeepLabV3[9], HRNet[65], OCR[81], without bringing any change to base networks during inference.

In a nutshell, our **contributions** are three-fold:

- We propose a supervised, pixel-wise contrastive learning method for semantic segmentation. It lifts current image-wise training strategy to an inter-image, pixel-to-pixel paradigm. It essentially learns a *well structured* pixel semantic embedding space, by making full use of the global semantic similarities among labeled pixels.
- We develop a region memory to better explore the large visual data space and support to further calculate pixel-to-region contrast. Integrated with pixel-to-pixel contrast computation, our method exploits semantic correlations among pixels, and between pixels and semantic regions.
- We demonstrate that more powerful segmentation models with better example and anchor sampling strategies could be delivered instead of selecting random pixel samples.

Our method can be seamlessly incorporated into existing segmentation networks without any changes to the base model and without extra inference burden during testing (Fig. 2). Hence, our method shows consistently improved intersection-over-union segmentation scores over challenging datasets (*i.e.*, Cityscapes [15], PASCAL-Context [53], COCO-Stuff[5] and CamVid[3]), using state-of-the-art segmentation architectures (*i.e.*, DeepLabV3 [9], HRNet [65] and OCR [81]) and standard backbones (*i.e.*, ResNet [32], HRNet [65]). The impressive results shed light on the promises of metric learning in dense image prediction tasks. We expect this work to provide insights into the critical role of global pixel relationships in segmentation network training, and foster research on the open issues raised.

2. Related Work

Our work draws on existing literature in semantic segmentation, contrastive learning and deep metric learning. For brevity, only the most relevant works are discussed.

Semantic Segmentation. FCN [51] greatly promotes the advance of semantic segmentation. It is good at end-to-end dense feature learning, however, only perceiving limited visual context with local receptive fields. As strong dependencies exist among pixels in an image and these dependencies are informative about the structures of the objects [70], how to capture such dependencies becomes a vital issue for further improving FCN. A main group of follow-up effort attempts to aggregate multiple pixels to explicitly model context, for example, utilizing different sizes of convolutional/pooling kernels or dilation rates to gather multi-scale visual cues [80, 84, 8, 9], building image pyramids to extract context from multi-resolution inputs, adopting the Encoder-Decoder architecture to merge multi-layer features [58, 47, 66], applying CRF to recover detailed structures [50, 87], and employing neural attention [67, 29] to directly exchange context between paired pixels [10, 35, 36, 24]. Apart from investigating context-aggregation network modules, another line of work turns to designing context-aware optimization objectives [40, 2, 86], *i.e*., directly verify segmentation structures during training, to replace the pixel-wise cross entropy loss.

Though impressive, these methods only address pixel dependencies within individual images, neglecting the global context of the labeled data, *i.e*., pixel semantic correlations across different training images. Through a pixel-wise contrastive learning formulation, we map pixels in different categories to more distinctive features. The learned pixel features are not only discriminative for semantic classification within images, but also, more crucially, across images.

Contrastive Learning. Recently, the most compelling methods for learning representations without labels have been unsupervised contrastive learning [55, 34, 73, 13, 12], which significantly outperformed other pretext task-based alternatives [43, 26, 18, 54]. With a similar idea to exemplar learning [19], contrastive methods learn representations in a discriminative manner by contrasting similar (positive) data pairs against dissimilar (negative) pairs. A major branch of subsequent studies focuses on how to select the positive and negative pairs. For image data, the standard positive pair sampling strategy is to apply strong perturbations to create multiple views of each image data [73, 12, 31, 34, 6]. Negative pairs are usually randomly sampled, but some hard negative example mining strategies [41, 57, 38] were recently proposed. In addition, to store more negative samples during contrast computation, fixed [73] or momentum updated [52, 31] memories are adopted. Some latest studies [41, 33, 71] also confirm label information can assist contrastive learning based image-level pattern pre-training.

We raise a *pixel-to-pixel* contrastive learning method for semantic segmentation in the fully supervised setting. It yields a new training protocol that explores global pixel relations in labeled data for regularizing segmentation embedding space. Though a few concurrent works also address contrastive learning in dense image prediction [75, 7, 69], the ideas are significantly different. First, they typically consider contrastive learning as a *pre-training* step for dense image embedding. Second, they simply use the local context within individual images, *i.e*., only compute the contrast among pixels from augmented versions of a *same* image. Third, they do not notice the critical role of metric learning in complementing current well-established pixel-wise cross-entropy loss based training regime (*cf.* §3.2).

Deep Metric Learning. The goal of metric learning is to quantify the similarity among samples using an optimal distance metric. Contrastive loss [28] and triplet loss [60] are two basic types of loss functions for deep metric learning. With a similar spirit of increasing and decreasing the distance between similar and dissimilar data samples, respectively, the former one takes pairs of sample as input while the latter is composed of triplets. Deep metric learning [22] has proven effective in a wide variety of computer vision tasks, such as image retrieval [64] and face recognition [60].

Although a few prior methods address the idea of metric learning in semantic segmentation, they only account for the local content from objects [29] or instances [16, 1, 22, 42]. It is worth noting [37] also explores cross-image information of training data, *i.e*., leverage perceptual pixel groups for non-parametric pixel classification. Due to its clustering based metric learning strategy, [37] needs to retrieve extra labeled data for inference. Differently, our core idea, *i.e*., exploit inter-image pixel-to-pixel similarity to enforce global constraints on the embedding space, is conceptually novel and rarely explored before. It is executed by a compact training paradigm, which enjoys the complementary advantages of unary, pixel-wise cross-entropy loss and pair-wise, pixel-to-pixel contrast loss, without bringing any extra inference cost or modification to the base network during deployment.

3. Methodology

Before detailing our supervised pixel-wise contrastive algorithm for semantic segmentation (§3.2), we first introduce the contrastive formulation in unsupervised visual representation learning and the notion of memory bank (§3.1).

3.1. Preliminaries

Unsupervised Contrastive Learning. Unsupervised visual representation learning aims to learn a CNN encoder f_{CNN} that transforms each training image I to a feature vector $\boldsymbol{v} = f_{\text{CNN}}(I) \in \mathbb{R}^D$, such that $\boldsymbol{v}$ best describes I. To achieve this goal, contrastive approaches conduct training by distinguishing a *positive* (an augmented version of *an-*

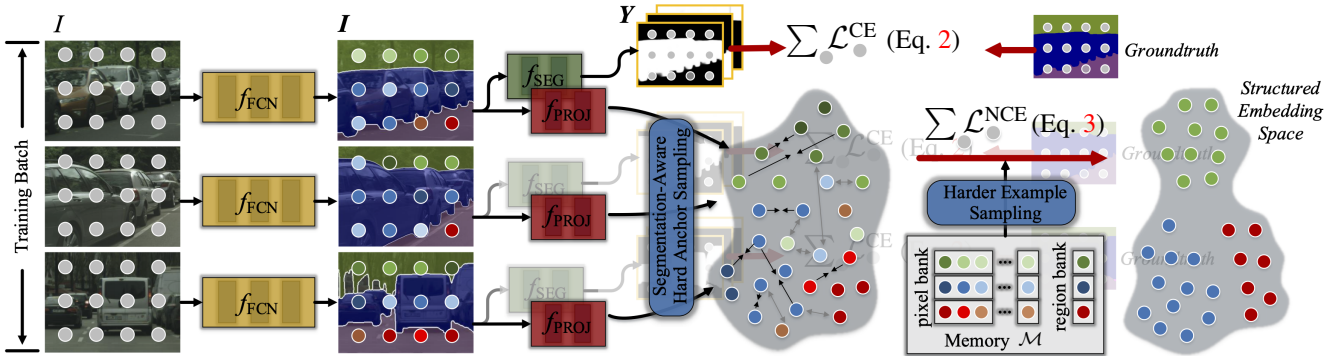


Figure 3: **Detailed illustration** of our pixel-wise contrastive learning based semantic segmentation network architecture.

chor I) from several *negatives* (images randomly drawn from the training set excluding I), based on the principle of similarity between samples. A popular loss function for contrastive learning, called InfoNCE [27, 55], takes the following form:

$$\mathcal{L}_I^{\text{NCE}} = -\log \frac{\exp(\boldsymbol{v}\cdot\boldsymbol{v}^+/\tau)}{\exp(\boldsymbol{v}\cdot\boldsymbol{v}^+/\tau)+\sum_{\boldsymbol{v}^-\in\mathcal{N}_I}\exp(\boldsymbol{v}\cdot\boldsymbol{v}^-/\tau)}, \quad (1)$$

where $\boldsymbol{v}^+$ is an embedding of a positive for I, $\mathcal{N}_I$ contains embeddings of negatives, '$\cdot$' denotes the inner (dot) product, and $\tau>0$ is a temperature hyper-parameter. Note that all the embeddings in the loss function are ℓ_2-normalized.

Memory Bank. As revealed by recent studies [73, 13, 31], a large set of negatives (*i.e.*, $|\mathcal{N}_I|$) is critical in unsupervised contrastive representation learning. As the number of negatives is limited by the mini-batch size, recent contrastive methods utilize large, external memories as a bank to store more navigate samples. Specifically, some methods [73] directly store the embeddings of all the training samples in the memory, however, easily suffering from asynchronous update. Some others choose to keep a queue of the last few batches [68, 13, 31] as memory. In [13, 31], the stored embeddings are even updated on-the-fly through a momentum-updated version of the encoder network f_{CNN}.

3.2. Supervised Contrastive Segmentation

Pixel-Wise Cross-Entropy Loss. In the context of semantic segmentation, each pixel i of an image I has to be classified into a semantic class $c\in\mathcal{C}$. Current approaches typically cast this task as a pixel-wise classification problem. Specifically, let f_{FCN} be an FCN encoder (*e.g.*, ResNet [32]), that produces a dense feature $\boldsymbol{I}\in\mathbb{R}^{H\times W\times D}$ for I, from which the pixel embedding $\boldsymbol{i}\in\mathbb{R}^D$ of i can be derived (*i.e.*, $\boldsymbol{i}\in\boldsymbol{I}$). Then a segmentation head f_{SEG} maps $\boldsymbol{I}$ into a categorical score map $\boldsymbol{Y}=f_{\text{SEG}}(\boldsymbol{I})\in\mathbb{R}^{H\times W\times|\mathcal{C}|}$. Further let $\boldsymbol{y}=[y_1,\cdots,y_{|\mathcal{C}|}]\in\mathbb{R}^{|\mathcal{C}|}$ be the *unnormalized* score vector (termed as *logit*) for pixel i, derived from $\boldsymbol{Y}$, *i.e.*, $\boldsymbol{y}\in\boldsymbol{Y}$. Given $\boldsymbol{y}$ for pixel i w.r.t its groundtruth label $\bar{c}\in\mathcal{C}$, the cross-entropy loss is optimized with $\texttt{softmax}$ (*cf.* Fig. 3):

$$\mathcal{L}_i^{\text{CE}} = -\mathbf{1}_{\bar{c}}^\top \log(\texttt{softmax}(\boldsymbol{y})), \quad (2)$$

where $\mathbf{1}_{\bar{c}}$ denotes the one-hot encoding of $\bar{c}$, the logarithm is defined as element-wise, and $\texttt{softmax}(y_c)=\frac{\exp(y_c)}{\sum_{c'=1}^{|\mathcal{C}|}\exp(y_{c'})}$. Such training objective design mainly suffers from two limitations. **1)** It penalizes pixel-wise predictions independently but ignores relationship between pixels [86]. **2)** Due to the use of $\texttt{softmax}$, the loss only depends on the relative relation among logits and cannot directly supervise on the learned representations [56]. These two issues were rarely noticed; only a few structure-aware losses are designed to address **1)**, by considering pixel affinity [40], optimizing intersection-over-union measurement [2], or maximizing the mutual information between the groundtruth and prediction map [86]. Nevertheless, these alternative losses only consider the dependencies between pixels within an image (*i.e.*, local context), regardless of the semantic correlations between pixels across images (*i.e.*, global structure).

Pixel-to-Pixel Contrast. In this work, we develop a pixel-wise contrastive learning method that addresses both **1)** and **2)**, through regularizing the embedding space and exploring the global structures of training data. We first extend Eq. (1) to our supervised, dense image prediction setting. Basically, the data samples in our contrastive loss computation are training image pixels. In addition, for a pixel i with its groundtruth semantic label $\bar{c}$, the positive samples are other pixels also belonging to the class $\bar{c}$, while the negatives are the pixels belonging to the other classes $\mathcal{C}\backslash\bar{c}$. Our supervised, pixel-wise contrastive loss is defined as:

$$\mathcal{L}_i^{\text{NCE}} = \frac{1}{|\mathcal{P}_i|}\sum_{\boldsymbol{i}^+\in\mathcal{P}_i} -\log \frac{\exp(\boldsymbol{i}\cdot\boldsymbol{i}^+/\tau)}{\exp(\boldsymbol{i}\cdot\boldsymbol{i}^+/\tau)+\sum_{\boldsymbol{i}^-\in\mathcal{N}_i}\exp(\boldsymbol{i}\cdot\boldsymbol{i}^-/\tau)}, \quad (3)$$

where $\mathcal{P}_i$ and $\mathcal{N}_i$ denote pixel embedding collections of the positive and negative samples, respectively, for pixel i. Note that the positive/negative samples and the anchor i are not restricted to being from a same image. As Eq. (3) shows, the purpose of such *pixel-to-pixel contrast* based loss design is to learn an embedding space, by pulling the same class pixel samples close and by pushing different class samples apart.

The pixel-wise cross-entropy loss in Eq. (2) and our contrastive loss in Eq. (3) are complementary to each other; the former lets segmentation networks learn discriminative

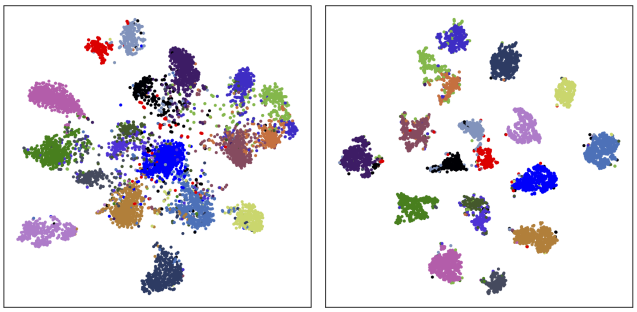


Figure 4: Visualization of features learned with (left) the pixel-wise entropy loss (*i.e.*, $\mathcal{L}^{\text{CE}}$ in Eq. (2)) and (right) our pixel contrast based optimization objective (*i.e.*, $\mathcal{L}^{\text{SEG}}$ in Eq. (4)) on Cityscapes `val` [15]. Features are colored according to class labels. As seen, the proposed $\mathcal{L}^{\text{SEG}}$ begets a well-structured semantic feature space.

pixel features that are meaningful for classification, while the latter helps to regularize the embedding space with improved intra-class compactness and inter-class separability through *explicitly* exploring global semantic relationships between pixel samples. Thus the overall training target is:

$$\mathcal{L}^{\text{SEG}} = \sum\nolimits_i \left(\mathcal{L}_i^{\text{CE}} + \lambda \mathcal{L}_i^{\text{NCE}}\right), \tag{4}$$

where $\lambda > 0$ is the coefficient. As shown in Fig. 4, the learned pixel embeddings by $\mathcal{L}^{\text{SEG}}$ become more compact and well separated. This suggests that, by enjoying the advantage of unary cross-entropy loss and pair-wise metric loss, segmentation network can generate more discriminative features, hence producing more promising results. Quantitative analyses are later provided in §4.2 and §4.3.

Pixel-to-Region Contrast. As stated in §3.1, memory is a critical technique that helps contrastive learning to make use of massive data to learn good representations. However, since there are vast numbers of pixel samples in our dense prediction setting and most of them are redundant (*i.e.*, sampled from harmonious object regions), directly storing all the training pixel samples, like traditional memory [12], will greatly slow down the learning process. Maintaining several last batches in a queue, like [68, 13, 31], is also not a good choice, as recent batches only contain a limited number of images, reducing the diversity of pixel samples. Thus we choose to maintain a pixel queue per category. For each category, only a small number, *i.e.*, V, of pixels are randomly selected from each image in the latest mini-batch, and pulled into the queue, with a size of $T \gg V$. In practice we find this strategy is very efficient and effective, but the under-sampled pixel embeddings are too sparse to fully capture image content. Therefore, we further build a region memory bank that stores more representative embeddings absorbed from image segments (*i.e.*, semantic regions).

Specifically, for a segmentation dataset with a total of N training images and $|\mathcal{C}|$ semantic classes, our region memory is built with size $|\mathcal{C}| \times N \times D$, where D is the dimension of pixel embeddings. The $(\bar{c}, n)$-th element in the region memory is a D-dimensional feature vector obtained by average pooling all the embeddings of pixels labeled as $\bar{c}$ category in the n-th image. The region memory brings two advantages: 1) store more representative "pixel" samples with low memory consumption; and 2) allow our pixel-wise contrastive loss (*cf.* Eq. (3)) to further explore pixel-to-region relations. With regard to 2), when computing Eq. (3) for an anchor pixel i belonging to $\bar{c}$ category, stored region embeddings with the same class $\bar{c}$ are viewed as positives, while the region embeddings with other classes $\mathcal{C} \backslash \bar{c}$ are negatives. For the pixel memory, the size is $|\mathcal{C}| \times T \times D$. Therefore, for the whole memory (denoted as $\mathcal{M}$), the total size is $|\mathcal{C}| \times (N+T) \times D$. We examine the design of $\mathcal{M}$ in §4.2. In the following sections, we will not distinguish pixel and region embeddings in $\mathcal{M}$, unless otherwise specified.

Hard Example Sampling. Prior research [60, 39, 41, 57, 38] found that, in addition to loss designs and the amount of training samples, the discriminating power of the training samples is crucial for metric learning. Considering our case, the gradient of the pixel-wise contrastive loss (*cf.* Eq. (3)) w.r.t. the anchor embedding $\boldsymbol{i}$ can be given as:

$$\frac{\partial \mathcal{L}_i^{\text{NCE}}}{\partial \boldsymbol{i}} = -\frac{1}{\tau |\mathcal{P}_i|} \sum\nolimits_{\boldsymbol{i}^+ \in \mathcal{P}_i} \Big((1 - p_{i^+}) \cdot \boldsymbol{i}^+ - \sum\nolimits_{\boldsymbol{i}^- \in \mathcal{N}_i} p_{i^-} \cdot \boldsymbol{i}^- \Big), \tag{5}$$

where $p_{i^{+/-}} \in [0, 1]$ denotes the matching probability between a positive/negative $i^{+/-}$ and the anchor i, *i.e.*, $p_{i^{+/-}} = \frac{\exp(\boldsymbol{i} \cdot \boldsymbol{i}^{+/-}/\tau)}{\sum_{\boldsymbol{i}' \in \mathcal{P}_i \cup \mathcal{N}_i} \exp(\boldsymbol{i} \cdot \boldsymbol{i}'/\tau)}$. We view the negatives with dot products (*i.e.*, $\boldsymbol{i} \cdot \boldsymbol{i}^-$) closer to 1 to be *harder*, *i.e.*, negatives which are similar to the anchor i. Similarly, the positives with dot products (*i.e.*, $\boldsymbol{i} \cdot \boldsymbol{i}^+$) closer to -1 are considered as *harder*, *i.e.*, positives which are dissimilar to i. We can find that, harder negatives bring more gradient contributions, *i.e.*, p_{i^-}, than easier negatives. This principle also holds true for positives, whose gradient contributions are $1 - p_{i^+}$. Kalantidis *et al.* [38] further indicate that, as training progresses, more and more negatives become too simple to provide significant contributions to the unsupervised contrastive loss (*cf.* Eq. (1)). This also happens in our supervised setting (*cf.* Eq. (3)), for both negatives and positives. To remedy this problem, we propose the following sampling strategies:

- *Hardest Example Sampling.* Inspired by hardest negative mining in metric learning [4], we first design a "hardest example sampling" strategy: for each anchor pixel embedding $\boldsymbol{i}$, only sampling top-K hardest negatives and positives from the memory bank $\mathcal{M}$, for the computation of the pixel-wise contrastive loss (*i.e.*, $\mathcal{L}^{\text{NCE}}$ in Eq. (3)).
- *Semi-Hard Example Sampling.* Some studies propose to make use of harder negatives, as optimizing with the hardest negatives for metric learning likely leads to bad local minima [60, 74, 23]. Thus we further design a "semi-hard example sampling" strategy: for each anchor embedding $\boldsymbol{i}$, we first collect top 10% nearest negatives (resp. top

10% farthest positives) from the memory bank $\mathcal{M}$, from which we randomly then sample K negatives (resp. K positives) for our contrastive loss computation.

- *Segmentation-Aware Hard Anchor Sampling.* Rather than mining informative positive and negative examples, we develop an anchor sampling strategy. We treat the categorization ability of an anchor embedding as its importance during contrastive learning. This leads to "segmentation-aware hard anchor sampling": the pixels with incorrect predictions, *i.e.*, $c \neq \bar{c}$, are treated as *hard anchors*. For the contrastive loss computation (*cf.* Eq. (3)), half of the anchors are randomly sampled and half are the hard ones. This anchor sampling strategy enables our contrastive learning to focus more on the pixels hard for classification, delivering more segmentation-aware embeddings.

In practice, we find "semi-hard example sampling" strategy performs better than "hardest example sampling". In addition, after employing "segmentation-aware hard anchor sampling" strategy, the segmentation performance can be further improved. See §4.2 for related experiments.

3.3. Detailed Network Architecture

Our algorithm has five major components (*cf.* Fig. 3):

- *FCN Encoder*, f_{FCN}, which maps each input image I into dense embeddings $\boldsymbol{I} = f_{\text{FCN}}(I) \in \mathbb{R}^{H \times W \times D}$. In our algorithm, any FCN backbones can be used to implement f_{FCN} and we test two commonly used ones, *i.e.*, ResNet [32] and HRNet [65], in our experiments.
- *Segmentation Head*, f_{SEG}, that projects $\boldsymbol{I}$ into a score map $\boldsymbol{Y} = f_{\text{SEG}}(\boldsymbol{I}) \in \mathbb{R}^{H \times W \times |\mathcal{C}|}$. We conduct evaluations using different segmentation heads in mainstream methods (*i.e.*, DeepLabV3 [9], HRNet [65], and OCR [81]).
- *Project Head*, f_{PROJ}, which maps each high-dimensional pixel embedding $\boldsymbol{i} \in \boldsymbol{I}$ into a 256-d ℓ_2-normalized feature vector [12], for the computation of the contrastive loss $\mathcal{L}_{\text{NCE}}$. f_{PROJ} is implemented as two 1×1 convolutional layers with `ReLU`. Note that the project head is only applied during training and is removed at inference time. Thus it does not introduce any changes to the segmentation network or extra computational cost in deployment.
- *Memory Bank*, $\mathcal{M}$, which consists of two parts that store pixel and region embeddings, respectively. For each training image, we sample $V = 10$ pixels per class. For each class, we set the size of the pixel queue as $T = 10N$. The memory bank is also discarded after training.
- *Joint Loss*, $\mathcal{L}_{\text{SEG}}$ (*cf.* Eq. (4)), that takes the power of representation learning (*i.e.*, $\mathcal{L}_{\text{CE}}$ in Eq. (2)) and metric learning (*i.e.*, $\mathcal{L}_{\text{NCE}}$ in Eq. (3)) for more distinct segmentation feature learning. In practice, we find our method is not sensitive to the coefficient λ (*e.g.*, when $\lambda \in [0.1, 1]$) and empirically set λ as 1. For $\mathcal{L}_{\text{NCE}}$ in Eq. (3), we set the temperature τ as 0.1. For sampling, we find "semi-hard example sampling" + "segmentation-aware hard anchor sampling" performs the best and set the numbers of sampled instances (*i.e.*, K) as 1,024 and 2,048 for positive and negative, respectively. For each mini-batch, 50 anchors are sampled per category (half are randomly sampled and the other half are segmentation-hard ones).

4. Experiment

4.1. Experimental Setup

Datasets. Our experiments are conducted on four datasets:

- **Cityscapes** [15] has 5,000 finely annotated urban scene images, with 2,975/500/1,524 for `train`/`val`/`test`. The segmentation performance is reported on 19 challenging categories, such as person, sky, car, and building.
- **PASCAL-Context** [53] contains 4,998 and 5,105 images in `train` and `test` splits, respectively, with precise annotations of 59 semantic categories.
- **COCO-Stuff** [5] consists of 10,000 images gathered from COCO [48]. It is split into 9,000 and 1,000 images for `train` and `test`. It provides rich annotations for 80 object classes and 91 stuff classes.
- **CamVid** [3] has 367/101/233 images for `train`/`val`/`test`, with 11 semantic labels in total.

Training. As mentioned in §3.3, various backbones (*i.e.*, ResNet [32] and HRNet [65]) and segmentation networks (*i.e.*, DeepLabV3 [9], HRNet [65], and OCR [81]) are exploited in our experiments to thoroughly validate the proposed algorithm. We follow conventions [65, 81, 14, 76] for training hyper-parameters. For fairness, we initialize all backbones using corresponding weights pretrained on ImageNet [59], with the remaining layers being randomly initialized. For data augmentation, we use color jitter, horizontal flipping and random scaling with a factor in $[0.5, 2]$. We use SGD as our optimizer, with a momentum 0.9 and weight decay 0.0005. We adopt the polynomial annealing policy [9] to schedule the learning rate, which is multiplied by $(1 - \frac{iter}{total\_iter})^{power}$ with $power = 0.9$. Moreover, for Cityscapes, we use a mini-batch size of 8, and an initial learning rate of 0.01. All the training images are augmented by random cropping from 1024×2048 to 512×1024. For the experiments on `test`, we follow [65] to train the model for 100K iterations. Note that we do not use any extra training data (*e.g.*, Cityscapes coarse [15]). For PASCAL-Context and COCO-Stuff, we opt a mini-batch size of 16, an initial learning rate of 0.001, and crop size of 520×520. We train for 60K iterations over their `train` sets. For CamVid, we train the model for 6K iterations, with batch size 16, learning rate 0.02 and original image size.

Testing. Following general protocol [65, 81, 61], we average the segmentation results over multiple scales with flipping, *i.e.*, the scaling factor is 0.75 to 2.0 (with intervals of 0.25) times of the original image size. Note that, during testing, there is no any change or extra inference step intro-

Pixel Contrast	Backbone	mIoU (%)
Baseline (*w/o* contrast)	HRNetV2-W48	78.1
Intra-Image Contrast	HRNetV2-W48	78.9 (+0.8)
Inter-Image Contrast	HRNetV2-W48	**81.0** (**+2.9**)

Table 1: **Comparison of different contrastive mechanisms** on Cityscapes `val` [15]. See §4.2 for more details.

Memory	Backbone	mIoU (%)
Baseline (*w/o* contrast)	HRNetV2-W48	78.1
Mini-Batch (*w/o* memory)	HRNetV2-W48	79.8 (+1.7)
Pixel Memory	HRNetV2-W48	80.5 (+2.6)
Region Memory	HRNetV2-W48	80.2 (+2.1)
Pixel + Region Memory	HRNetV2-W48	**81.0** (**+2.9**)

Table 2: **Comparison of different memory bank designs** on Cityscapes `val` [15]. See §4.2 for more details.

duced to the base segmentation models, *i.e.*, the projection head, f_{PROJ}, and memory bank, $\mathcal{M}$, are directly discarded.

Evaluation Metric. Following the standard setting, mean intersection-over-union (mIoU) is used for evaluation.

Reproducibility. Our model is implemented in PyTorch and trained on four NVIDIA Tesla V100 GPUs with a 32GB memory per-card. Testing is conducted on the same machine. Our implementations are available at `https://github.com/tfzhou/ContrastiveSeg`.

4.2. Diagnostic Experiment

We first study the efficacy of our core ideas and essential model designs, over Cityscapes `val` [15]. We adopt HRNet [65] as our base segmentation network (denoted as "Baseline (*w/o* contrast)" in Tables 1-3). To perform extensive ablation experiments, we train each model for 40K iterations while keeping other hyper-parameters unchanged.

Inter-Image *vs.* Intra-Image Pixel Contrast. We first investigate the effectiveness of our core idea of inter-image pixel contrast. As shown in Table 1, additionally considering cross-image pixel semantic relations (*i.e.*, "Inter-Image Contrast") in segmentation network learning leads to a substantial performance gain (*i.e.*, **2.9**%), compared with "Baseline (*w/o* contrast)". In addition, we develop another baseline, "Intra-Image Contrast", which only samples pixels from same images during the contrastive loss (*i.e.*, $\mathcal{L}^{\text{NCE}}$ in Eq. (5)) computation. The results in Table 1 suggest that, although "Intra-Image Contrast" also boosts the performance over "Baseline (*w/o* contrast)" (*i.e.*, 78.1%→78.9%), "Inter-Image Contrast" is more favored.

Memory Bank. We next validate the design of our memory bank. The results are summarized in Table 2. Based on "Baseline (*w/o* contrast)", we first derive a variant, "Mini-Batch *w/o memory*": only compute pixel contrast within each mini-batch, without outside memory. It gets 79.8% mIoU. We then provision this variant with pixel and region memories separately, and observe consistent performance gains (79.8% → 80.5% for pixel memory and 79.8% → 80.2% for region memory). This verifies that **i)** leveraging more pixel samples during contrastive learning leads to better pixel embeddings; and **ii)** both pixel-to-pixel and pixel-to-region relations are informative cues. Finally, after using both the two memories, a higher score (*i.e.*, 81.0%) is achieved, revealing **i)** the effectiveness of our memory design; and **ii)** necessity of comprehensively considering both pixel-to-pixel contrast and pixel-to-region contrast.

Sampling: Anchor	Sampling: Pos./Neg.	Backbone	mIoU (%)
Baseline (*w/o* contrast)		HRNetV2-W48	78.1
Random	Random	HRNetV2-W48	79.3 (+1.2)
	Hardest	HRNetV2-W48	79.4 (+1.3)
	Semi-Hard	HRNetV2-W48	80.1 (+2.0)
Seg.-aware hard	Random	HRNetV2-W48	80.2 (+2.1)
	Hardest	HRNetV2-W48	80.5 (+2.4)
	Semi-Hard	HRNetV2-W48	**81.0** (**+2.9**)

Table 3: **Comparison of different hard example sampling strategies** on Cityscapes `val` [15]. See §4.2 for more details.

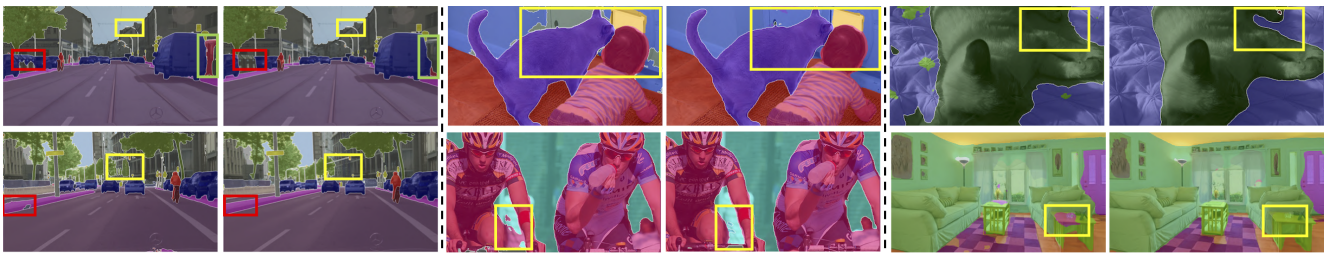
Hard Example Mining. Table 3 presents a comprehensive examination of various hard example mining strategies proposed in §3.2. Our main observations are the following: **i)** For positive/negative sampling, mining meaningful pixels (*i.e.*, "hardest" or "semi-hard" sampling), rather than "random" sampling, is indeed useful; **ii)** Hence, "semi-hard" sampling is more favored, as it improves the robustness of training by avoiding overfitting outliers in the training set. This corroborates related observations in unsupervised setting [72] and indicates that segmentation may benefit from more intelligent sample treatment; and **iii)** For anchor sampling, "seg.-aware hard" strategy further improves the performance (*i.e.*, 80.1%→81.0%) over "random" sampling only. This suggests that exploiting task-related signals in supervised metric learning may help develop better segmentation solutions, which has remained relatively untapped.

4.3. Comparison to State-of-the-Arts

Cityscapes [15]. Table 4 lists the scores on Cityscapes `test`, under two widely used training settings [65] (trained over `train` or `train+val`). Our method brings impressive gains over 3 strong baselines (*i.e.*, DeepLabV3, HRNetV2, and OCR), and sets a new state-of-the-art.

PASCAL-Context [53]. Table 5 presents comparison results on PASCAL-Context `test`. Our approach improves the performance of base networks by solid margins (*i.e.*, 54.0→55.1 for HRNetV2, 56.2→57.2 for OCR). This is particularly impressive considering the fact that improvement on this extensively-benchmarked dataset is very hard.

COCO-Stuff [5]. Table 6 reports performance comparison of our method against seven competitors on COCO-Stuff `test`. As we find that *OCR+**Ours*** yields a mIoU of 41.0%, which leads to a promising gain of **0.5%** over its counterpart (*i.e.*, OCR with a 40.5% mIoU). Besides, *HR-*



OCR | OCR + Ours | OCR | OCR + Ours | OCR | OCR + Ours

Figure 5: **Visual comparisons** between *OCR* [81] and *OCR+Ours* (from left to right: Cityscapes, PASCAL-Context, COCO-Stuff).

Model	Backbone	mIoU (%)
Model learned on Cityscapes `train`		
PSPNet$_{17}$ [84]	D-ResNet-101	78.4
PSANet$_{18}$ [85]	D-ResNet-101	78.6
PAN$_{18}$ [44]	D-ResNet-101	78.6
AAF$_{18}$ [40]	D-ResNet-101	79.1
DeepLabV3$_{17}$ [9]	D-ResNet-101	78.1
DeepLabV3 + **Ours**	D-ResNet-101	**79.2** (+1.1)
HRNetV2$_{20}$ [65]	HRNetV2-W48	80.4
HRNetV2+ **Ours**	HRNetV2-W48	**81.4** (+1.0)
Model learned on Cityscapes `train+val`		
DFN$_{18}$ [79]	D-ResNet-101	79.3
PSANet$_{18}$ [85]	D-ResNet-101	80.1
SVCNet$_{19}$ [17]	D-ResNet-101	81.0
CPN$_{20}$ [77]	D-ResNet-101	81.3
DANet$_{19}$ [24]	D-ResNet-101	81.5
ACF$_{19}$ [82]	D-ResNet-101	81.8
DGCNet$_{19}$ [83]	D-ResNet-101	82.0
HANet$_{20}$ [14]	D-ResNet-101	82.1
ACNet$_{19}$ [25]	D-ResNet-101	82.3
DeepLabV3$_{17}$ [9]	D-ResNet-101	79.4
DeepLabV3 + **Ours**	D-ResNet-101	**80.3** (+0.9)
HRNetV2$_{20}$ [65]	HRNetV2-W48	81.6
HRNetV2+ **Ours**	HRNetV2-W48	**82.5** (+0.9)
OCR$_{20}$ [81]	HRNetV2-W48	82.4
OCR+ **Ours**	HRNetV2-W48	**83.2** (+0.8)

Table 4: **Quantitative segmentation results** on Cityscapes `test` [15]. D-ResNet-101 = Dilated ResNet-101. See §4.3.

Model	Backbone	mIoU (%)
DANet$_{19}$ [24]	D-ResNet-101	52.6
SVCNet$_{19}$ [17]	D-ResNet-101	53.2
CPN$_{20}$ [77]	D-ResNet-101	53.9
ACNet$_{19}$ [25]	D-ResNet-101	54.1
DMNet$_{19}$ [30]	D-ResNet-101	54.4
RANet$_{20}$ [61]	ResNet-101	54.9
DNL$_{20}$ [76]	HRNetV2-W48	55.3
HRNetV2$_{20}$ [65]	HRNetV2-W48	54.0
HRNetV2+ **Ours**	HRNetV2-W48	**55.1** (+1.1)
OCR$_{20}$ [81]	HRNetV2-W48	56.2
OCR+ **Ours**	HRNetV2-W48	**57.2** (+1.0)

Table 5: **Quantitative segmentation results** on PASCAL-Context `test` [53]. D-ResNet-101 = Dilated ResNet-101. See §4.3.

*NetV2+**Ours*** outperforms HRNetV2 by **0.6%**.

CamVid [3]. Table 7 shows that our method also leads to improvements over HRNetV2 and OCR on CamVid `test`.

Qualitative Results. Fig. 5 depicts qualitative comparisons of *OCR+**Ours*** against *OCR* over representative examples from three datasets (*i.e.*, Cityscapes, PASCAL-Context and COCO-Stuff). As seen, our method is capable of producing more accurate segments across various challenge scenarios.

Model	Backbone	mIoU (%)
SVCNet$_{19}$ [17]	D-ResNet-101	39.6
DANet$_{19}$ [24]	D-ResNet-101	39.7
SpyGR$_{20}$ [46]	ResNet-101	39.9
ACNet$_{19}$ [25]	ResNet-101	40.1
HRNetV2$_{20}$ [65]	HRNetV2-W48	38.7
HRNetV2+ **Ours**	HRNetV2-W48	**39.3** (+0.6)
OCR$_{20}$ [81]	HRNetV2-W48	40.5
OCR+ **Ours**	HRNetV2-W48	**41.0** (+0.5)

Table 6: **Quantitative segmentation results** on COCO-Stuff `test` [5]. D-ResNet-101 = Dilated ResNet-101. See §4.3.

Model	Backbone	mIoU (%)
DFANet$_{19}$ [45]	Xception	64.7
BiSeNet$_{18}$ [78]	D-ResNet-101	68.7
PSPNet$_{17}$ [84]	D-ResNet-101	69.1
HRNetV2$_{20}$ [65]	HRNetV2-W48	78.5
HRNetV2+ **Ours**	HRNetV2-W48	**79.0** (+0.5)
OCR$_{20}$ [81]	HRNetV2-W48	80.1
OCR+ **Ours**	HRNetV2-W48	**80.5** (+0.4)

Table 7: **Quantitative segmentation results** on CamVid `test` [3]. D-ResNet-101=Dilated ResNet-101. See §4.3.

5. Conclusion and Discussion

In this paper, we propose a new supervised learning paradigm for semantic segmentation, enjoying the complementary advantages of unary classification and structured metric learning. Through pixel-wise contrastive learning, it investigates global semantic relations between training pixels, guiding pixel embeddings towards cross-image category-discriminative representations that eventually improve the segmentation performance. Our method generates promising results and shows great potential in a variety of dense prediction tasks, such as pose estimation [89, 21] and body parsing [88, 20]. It also comes with new challenges, in particular regarding smart data sampling, metric learning loss design, class rebalancing during training, and multi-layer feature contrast. Given the massive number of technique breakthroughs over the past few years, we expect a flurry of innovation towards these promising directions.

Acknowledgment This work was supported by Zhejiang Lab's Open Fund (No. 2020AA3AB14) and CCF-Baidu Open Fund.

References

[1] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *CVPR*, 2017. 3

[2] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, 2018. 1, 3, 4

[3] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *PRL*, 30(2):88–97, 2009. 2, 6, 8

[4] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Hard negative mining for metric learning based zero-shot classification. In *ECCV*, 2016. 5

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018. 2, 6, 7, 8

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 3

[7] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *NeurIPS*, 2020. 3

[8] 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 TPAMI*, 40(4):834–848, 2017. 1, 3

[9] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 2, 3, 6, 8

[10] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016. 3

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 3, 5, 6

[13] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3, 4, 5

[14] Sungha Choi, Joanne T Kim, and Jaegul Choo. Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In *CVPR*, 2020. 2, 6, 8

[15] 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 *CVPR*, 2016. 1, 2, 5, 6, 7, 8

[16] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017. 3

[17] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *CVPR*, 2019. 8

[18] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 3

[19] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *NeurIPS*, 2014. 3

[20] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. In *CVPR*, 2018. 8

[21] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 8

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

[23] Jonathan Frankle, David J Schwab, Ari S Morcos, et al. Are all negatives created equal in contrastive instance discrimination? *arXiv preprint arXiv:2010.06682*, 2020. 5

[24] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 1, 3, 8

[25] Jun Fu, Jing Liu, Yuhang Wang, Yong Li, Yongjun Bao, Jinhui Tang, and Hanqing Lu. Adaptive context network for scene parsing. In *ICCV*, 2019. 8

[26] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 3

[27] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 4

[28] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 3

[29] Adam W Harley, Konstantinos G Derpanis, and Iasonas Kokkinos. Segmentation-aware convolutional networks using local attention masks. In *ICCV*, 2017. 3

[30] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *ICCV*, 2019. 8

[31] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 3, 4, 5

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2, 4, 6

[33] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *ICML*, 2020. 3

[34] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 3

[35] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 1, 3

[36] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 3

[37] 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 *ICCV*, 2019. 1, 3

[38] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *NeurIPS*, 2020. 3, 5

[39] Mahmut Kaya and Hasan Şakir Bilge. Deep metric learning: A survey. *Symmetry*, 11(9):1066, 2019. 2, 5

[40] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X Yu. Adaptive affinity fields for semantic segmentation. In *ECCV*, 2018. 1, 3, 4, 8

[41] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 3, 5

[42] Shu Kong and Charless Fowlkes. Recurrent pixel embedding for instance grouping. In *CVPR*, 2018. 3

[43] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *ECCV*, 2016. 3

[44] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018. 8

[45] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *CVPR*, 2019. 8

[46] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *CVPR*, 2020. 8

[47] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 1, 3

[48] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[49] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, 2016. 2

[50] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen Change Loy, and Xiaoou Tang. Deep learning markov random field for semantic segmentation. *IEEE TPAMI*, 40(8):1814–1828, 2017. 3

[51] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 3

[52] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020. 3

[53] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 2, 6, 7, 8

[54] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 3

[55] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 4

[56] Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, and Jun Zhu. Rethinking softmax cross-entropy loss for adversarial robustness. In *ICLR*, 2020. 4

[57] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *ICLR*, 2021. 3, 5

[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1, 3

[59] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 6

[60] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2, 3, 5

[61] Dingguo Shen, Yuanfeng Ji, Ping Li, Yi Wang, and Di Lin. Ranet: Region attention network for semantic segmentation. *NeurIPS*, 2020. 6, 8

[62] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *ICCV*, 2015. 2

[63] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[64] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014. 3

[65] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 2020. 2, 6, 7, 8

[66] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *CVPR*, 2019. 3

[67] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3

[68] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *CVPR*, 2020. 4, 5

[69] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021. 3

[70] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 3

[71] Longhui Wei, Lingxi Xie, Jianzhong He, Jianlong Chang, Xiaopeng Zhang, Wengang Zhou, Houqiang Li, and Qi Tian. Can semantic labels assist self-supervised visual representation learning? *arXiv preprint arXiv:2011.08621*, 2020. 3

[72] Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. On mutual information in contrastive learning for visual representations. *arXiv preprint arXiv:2005.13149*, 2020. 7

[73] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin.

Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 3, 4

[74] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Delving into inter-image invariance for unsupervised visual representations. *arXiv preprint arXiv:2008.11702*, 2020. 5

[75] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *arXiv preprint arXiv:2011.10043*, 2020. 3

[76] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *ECCV*, 2020. 6, 8

[77] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *CVPR*, 2020. 8

[78] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018. 8

[79] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *CVPR*, 2018. 8

[80] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 1, 3

[81] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 2, 6, 8

[82] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding. Acfnet: Attentional class feature network for semantic segmentation. In *ICCV*, 2019. 8

[83] Li Zhang, Xiangtai Li, Anurag Arnab, Kuiyuan Yang, Yunhai Tong, and Philip HS Torr. Dual graph convolutional network for semantic segmentation. In *BMVC*, 2019. 8

[84] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1, 2, 3, 8

[85] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018. 8

[86] Shuai Zhao, Yang Wang, Zheng Yang, and Deng Cai. Region mutual information loss for semantic segmentation. In *NeurIPS*, 2019. 1, 3, 4

[87] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 3

[88] Tianfei Zhou, Siyuan Qi, Wenguan Wang, Jianbing Shen, and Song-Chun Zhu. Cascaded parsing of human-object interaction recognition. *IEEE TPAMI*, 2021. 8

[89] Tianfei Zhou, Wenguan Wang, Si Liu, Yi Yang, and Luc Van Gool. Differentiable multi-granularity human representation learning for instance-aware human semantic parsing. In *CVPR*, 2021. 8