International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02083-x

Imbalance-Aware Discriminative Clustering for Unsupervised Semantic Segmentation

Mingyuan Liu[1,2] · **Jicong Zhang**[1] · **Wei Tang**[2]

Received: 22 August 2023 / Accepted: 15 April 2024

Abstract

Unsupervised semantic segmentation (USS) aims at partitioning an image into semantically meaningful segments by learning from a collection of unlabeled images. The effectiveness of current approaches is plagued by difficulties in coordinating representation learning and pixel clustering, modeling the varying feature distributions of different classes, handling outliers and noise, and addressing the pixel class imbalance problem. This paper introduces a novel approach, termed Imbalance-Aware Dense Discriminative Clustering (IDDC), for USS, which addresses all these difficulties in a unified framework. Different from existing approaches, which learn USS in two stages (i.e., generating and updating pseudo masks, or refining and clustering embeddings), IDDC learns pixel-wise feature representation and dense discriminative clustering in an end-to-end and self-supervised manner, through a novel objective function that transfers the manifold structure of pixels in the embedding space of a vision Transformer (ViT) to the label space while tolerating the noise in pixel affinities. During inference, the trained model directly outputs the classification probability of each pixel conditioned on the image. In addition, this paper proposes a new regularizer, based on the Weibull function, to handle pixel class imbalance and cluster degeneration in a single shot. Experimental results demonstrate that IDDC significantly outperforms all previous USS methods on three real-world datasets, COCO-Stuff-27, COCO-Stuff-171, and Cityscapes. Extensive ablation studies validate the effectiveness of each design. Our code is available at https://github.com/MY-LIU100101/IDDC.

Keywords Unsupervised semantic segmentation · Imbalance-Aware Dense Discriminative Clustering · End-to-end training · Deep clustering

1 Introduction

Semantic segmentation aims at assigning a categorical label to each pixel in an image. It facilitates many applications like autonomous driving (Hou et al., 2022; Hu et al., 2020; Qi et al., 2017), street scene understanding (Cheng et al., 2021; Gu et al., 2022; Liu et al., 2021), and medical image analysis (Ji et al., 2021; Peng et al., 2022; Zhou et al., 2022). In the past decade, supervised deep learning methods (Ghiasi et al., 2021; Liu et al., 2022; Vahdat et al., 2021; Wang et al., 2022; Wortsman et al., 2022) have significantly pushed forward the state-of-the-art performance of this task. However, they require a large amount of densely annotated images for training, which is not only laborious and expensive but also limits their broad use. To overcome this limitation, semi-supervised methods (Alonso et al., 2021; Kalluri et al., 2019; Ke et al., 2020; Kwon & Kwak, 2022; Lai et al., 2021; Mendel et al., 2020; Mittal et al., 2019) require only a small portion of images to be annotated; weakly-supervised methods (Ahn & Kwak, 2018; Chang et al., 2020; Lee et al., 2021; Wang et al., 2020; Wei et al., 2018; Wang et al., 2022; Zhang et al., 2021) leverage weaker forms of annotations, such as bounding boxes, image-level labels, and scribbles. In this paper, we focus on a completely label-free, but more challenging setting: unsupervised semantic segmentation (USS).

USS aims at learning semantic segmentation from a collection of unlabeled images. Without relying on any form of human annotation, USS removes the substantial labeling cost, and it can be easily applied to new data or new applications. In addition, USS can discover novel visual patterns

Communicated by Ziyue Xu.

✉ Wei Tang
tangw@uic.edu

[1] Department of Biological Science and Medical Engineering, Beihang University, Beijing 100191, China

[2] Department of Computer Science, University of Illinois, Chicago, IL 60607, USA

Published online: 14 May 2024

and structures that are not known *a priori* (Cho et al., 2021), which is particularly useful for analyzing images in novel domains. Despite its importance, USS is a very challenging problem because of lacking semantic supervision, large intra-class variation, and severe class imbalance.

Existing methods commonly tackle USS in a two-stage learning framework. They can be divided into two categories. The first category of methods generates pseudo labels by clustering pixel embeddings from pretrained models via K-means, and then iteratively refines the segmentation (Caron et al., 2018; Cho et al., 2021; Gao et al., 2022; Yin et al., 2022). However, their performances could be strongly affected by initially generated pseudo masks. Moreover, it is difficult to determine how frequently the two processes should be alternated: constantly changing pseudo labels can confuse the feature learning process and produce unstable results (Zhan et al., 2020), while updating pseudo labels slowly can make feature learning overfit to the initial label guesses.

The second category of methods, represented STEGO (Hamilton et al., 2022), learns low-dimensional and clustering-friendly pixel embeddings, and then groups them into clusters for USS (Li et al., 2023; Melas-Kyriazi et al., 2022; Pang et al., 2022; Seong et al., 2023; Van Gansbeke et al., 2021; Zadaianchuk et al., 2022; Ziegler & Asano, 2022). There are, however, several limitations. First, the processes of feature learning and pixel clustering are separated, rather than end-to-end (i.e., all modules are simultaneously trained). Since the representation learning process is completely unaware of the subsequent clustering task, including the number of clusters and the clustering objective, the model optimization could be suboptimal for clustering. Second, the clustering process, most commonly K-means, is generative and makes strong assumptions about the shape of clusters. It is not good at handling high-dimensional features and is susceptible to outliers, compared to its discriminative counterpart (Ng & Jordan, 2001). Third, the severe pixel class imbalance problem is largely neglected in USS. Directly integrating representation learning and clustering often leads to degenerate solutions: a few clusters are empty (Caron et al., 2018; Ji et al., 2019). A straightforward remedy (Ji et al., 2019; Krause et al., 2010; Van Gansbeke et al., 2020) is to enforce that class labels are evenly distributed, which, however, contradicts the highly skewed class distribution of pixels in practice.

This paper introduces a novel approach, termed Imbalance-Aware Dense Discriminative Clustering (IDDC), for USS. It addresses all aforementioned limitations in a unified framework.

First, IDDC is distinguished by its discriminative design, which directly outputs pixel-wise classification probabilities conditioned on the input image. Different from the widely used K-means, IDDC does not assume the generative distribution or shape of a cluster. Thus, it is more flexible to handle the varying feature distributions of different classes and datasets, more robust to outliers, and better categorizes pixels in the high-dimensional feature space.

Second, IDDC learns pixel-wise feature representation and dense discriminative clustering in an end-to-end and self-supervised manner, through a novel objective function that transfers the manifold structure of pixels in the embedding space of a vision Transformer (ViT) to the label space. IDDC coordinates representation learning and pixel clustering seamlessly, bypassing intermediate learning objectives such as generating and updating pseudo labels, or learning dimensionally reduced and clustering-friendly embeddings.

Last but not least, IDDC includes a new regularizer, based on the Weibull function (Murthy et al., 2004), to handle pixel class imbalance and cluster degeneration in a single shot. Entropy has been widely used in existing deep clustering methods (Barber & Agakov, 2005; Bridle et al., 1991; Krause et al., 2010; Van Gansbeke et al., 2020) to avoid empty clusters, by promoting that pixels are evenly assigned to each cluster. However, the class distribution of pixels in the real world is highly skewed, caused by the different sizes and occurring frequencies of each object and stuff. We will show that our new regularizer can simultaneously model the skewed distribution and avoid any empty cluster.

Our contributions are summarized as follows:

- We propose a novel approach, named Imbalance-Aware Dense Discriminative Clustering (IDDC), for USS. It directly predicts pixel-wise classification probabilities from an image and jointly learns dense feature representations and pixel labeling in a discriminative, end-to-end, and self-supervised manner. Compared with the existing USS methods, which commonly adopt a two-stage learning framework and generative clustering, IDDC addresses their difficulties in coordinating representation learning and pixel clustering, modeling the imbalanced class distribution, and handling outliers.
- We design a novel objective function that effectively learns IDDC by transferring the manifold structure of pixels in the ViT embedding space to the label space. We have an in-depth investigation of handling noisy training signals, caused by pixels that belong to different classes but are similar in the ViT embedding space. This is critical to self-supervised dense representation learning and pixel labeling.
- We introduce a novel regularizer, based on the Weibull function. It not only avoids clustering degeneration, i.e., empty clusters, but also addresses the pixel class imbalance problem in USS. The latter has been ignored by prior work.
- Experiments on three large-scale real-world datasets COCO-Stuff-27, COCO-Stuff-171, and Cityscapes show that IDDC outperforms the state-of-the-art methods by a large margin. It serves as a new baseline for the nascent

but challenging task of USS. We validate the effectiveness of each individual design of IDDC through a large number of ablation studies.

2 Related Works

2.1 Unsupervised Semantic Segmentation (USS)

USS aims at learning to label every pixel in an image without any form of human annotation. Early works model USS as a patch-level grouping problem, where different pixel orderings of a patch (Ouali et al., 2020), various augmented views of a patch (Mirsadeghi et al., 2021), or spatially adjacent patches (Ji et al., 2019) are clustered together by maximizing their mutual information. However, they tend to produce inaccurate segmentation and are mostly applied for segmenting stuff, e.g., sky and road, due to their negligence of the correlation between patches from different images and the imbalanced nature of the class distribution.

Afterward, pixel-level solutions with cross-image learning are proposed. They follow two-stage learning strategies: generating pseudo masks and then refining them, or learning pixel-wise embeddings and then clustering them. The former line of work iteratively generates and refines pseudo masks as supervision. PiCIE (Cho et al., 2021), inspired by deep cluster (Caron et al., 2018), contrastively learns descriptive pixel embeddings and iteratively updates a K-means cluster for generating pseudo masks. PASS (Gao et al., 2022) learns a self-supervised model and achieves pseudo masks with the assistance of pixel-attention maps. Yin et al. (2022) clusters pretrained pixel embeddings using K-means as pseudo labels, and refines them in a bootstrapping manner. The later line of work, represented by STEGO (Hamilton et al., 2022), attempts to refine embeddings from pretrained models and group them using the downstream K-means clustering. Hamilton et al. (2022) distills dense correspondences obtained from a pretrained network to learn low-dimensional pixel-wise features and then clusters them using K-means. Pang et al. (2022) extracts invariance from video frames to learn more descriptive features. Seong et al. (2023) extracts positive pixel pairs for contrastive learning, based on pretrained embeddings and local adjacency. Li et al. (2023) over-segments each image into several regions and clusters regional representations using K-means for USS.

IDDC differs with STEGO (Hamilton et al., 2022) significantly in modeling, learning, and addressing class imbalance. (1) *Modeling.* STEGO first extracts low-dimensional features through distillation and subsequently utilizes the K-means algorithm for pixel clustering, which assumes spherical distributions of input features, equal variance across clusters, and uniform cluster sizes. In contrast, IDDC is discriminative and directly outputs the classification probability of each pixel conditioned on the input image; it does not impose any of these assumptions. Therefore, IDDC is more flexible in handling varying feature distributions across different classes and datasets, more robust to outliers, and more effectively categorizes pixels in the high-dimensional feature space. (2) *Learning.* STEGO learns distillation and K-means separately in two stages. The two stages are optimized under different learning objectives, which is suboptimal. For example, the distillation stage is completely unaware of the clustering objective or cluster number and therefore cannot adjust itself based on the need of the clustering stage. In contrast, IDDC can be trained end-to-end under a unified learning objective. As a result, it allows representation learning and pixel clustering to seamlessly coordinate with each other for more effective learning. (3) *Addressing class imbalance.* STEGO clusters distilled features using K-means, which often generates balanced clustering results (Lu et al., 2019; Xiong et al., 2006) that are inconsistent with the imbalanced class distribution. In contrast, IDDC proposes a novel regularization term based on the Weibull function to address pixel class imbalance and cluster degradation in a single shot.

2.2 Unsupervised Object-Centric Segmentation

Some works (Van Gansbeke et al., 2021, 2022; Zadaianchuk et al., 2022; Melas-Kyriazi et al., 2022; Ziegler & Asano, 2022) focus on segmenting objects in an image. They adopt a two-stage self-supervised learning framework. The first stage extracts the foreground region or object parts, and learns their visual representations. The second stage clusters these regions into different object categories via K-means. Concretely, Van Gansbeke et al. (2021) extracts foreground regions via a supervised salient detection network and learns their features via contrastive learning. Then, K-means clusters the feature vectors average-pooled from each salient object region to determine their categories. COMUS (Zadaianchuk et al., 2022) and MaskDistill (Van Gansbeke et al., 2022) extract foreground regions using unsupervised saliency detection, pixel embeddings, and the attention mechanism, respectively. The extracted regions are then clustered as pseudo labels for training segmentation networks that segment multiple objects. Leopart (Ziegler & Asano, 2022) extracts foreground regions by clustering self-supervised dense features and then divides them into multiple objects through community detection. DSM (Melas-Kyriazi et al., 2022) selects top-k eigenvectors using a spectral segmentation method to identify object parts, and the feature vectors average-pooled from each part are clustered by K-means to obtain the semantic labels.

These works separately learn foreground regions, features, and clustering, and they only label objects in an image. In contrast, IDDC learns dense features and pixel clustering in an end-to-end and discriminative manner, and it labels every

pixel, including both objects and stuff. In addition, IDDC explicitly handles noise in self-supervision and pixel class imbalance, which were not studied in these works.

2.3 Discriminative Clustering

Discriminative clustering aims at learning a discriminative classifier, i.e., separation boundaries between clusters, from unlabeled data. Traditional methods optimize mutual information (Barber & Agakov, 2005; Bridle et al., 1991; Krause et al., 2010) and margin maximization (Xu et al., 2004; Zhao et al., 2008). They aim at balancing class distributions or finding maximum margin hyperplanes from data. Deep learning methods (Van Gansbeke et al., 2020; Ji et al., 2019; Ouali et al., 2020; Schmarje et al., 2021; Ghasedi Dizaji et al., 2017; Chang et al., 2017) follow the idea of connectivity-based clustering, where pairs of visually similar images are grouped together. To avoid clustering degeneration, they assume category labels are distributed evenly across the dataset.

Differently, IDDC focuses on USS, a dense pixel labeling task. Cho et al. (2021) show that directly applying an image clustering method to pixels does not perform well. In addition, it is nontrivial to handle the computational overhead brought by dense labeling. Moreover, we conducted an in-depth investigation on handling the pixel imbalance problem and noisy training signals caused by pixels that belong to different classes but are similar in the embedding space, which are neglected before.

2.4 Self-Supervised Learning

Self-supervised learning aims to dig meaningful visual representations from unlabeled images. Early works achieve this goal by solving pretext tasks, such as image inpainting, solving jigsaw puzzles, and predicting rotations (Alexey et al., 2015; Bojanowski & Joulin, 2017; Doersch et al., 2015; He et al., 2020; Komodakis & Gidaris, 2018; Noroozi & Favaro, 2016). Recent works are primarily based on contrastive learning (Chen et al., 2020; Chen & He, 2021; Grill et al., 2020; He et al., 2020). They pull together different views of the same image, while pushing away different images. After training, the learned image representation can be used for downstream tasks.

As global representations are insufficient for dense prediction (He et al., 2019; Purushwalkam & Gupta, 2020), some research extends image-level methods to pixel-level ones (Hung et al., 2019; Li et al., 2021; Roh et al., 2021; Wang et al., 2021) by establishing dense correspondences across views. Another group of methods exploits the ability of vision transformer (ViT) (Touvron et al., 2021) that patch-level similarities are better preserved than convolutional neural networks since they could model long-range feature interactions (Caron et al., 2021; Hamilton et al., 2022; Zhou et al., 2021). Our approach uses the pretrained dense features from a self-supervised ViT as pixel embeddings.

3 Method

3.1 Overview

Unsupervised semantic segmentation (USS) aims at learning a semantic segmentation model from a collection of unlabeled images. Given a new image, it will be able to assign a categorical label to each pixel.

Prior USS methods commonly adopt a two-stage sequential or iterative learning framework. The first stage is a neural network that maps an image to its dense features. The second stage is a conventional clustering algorithm, mostly typically K-means, that clusters pixels in the feature space. As discussed in Sect. 1, this framework is plagued by several difficulties: it is difficult to coordinate representation learning and pixel clustering because these two processes are separated, rather than end-to-end; it is difficult to model the variability of objects and stuff and handle outliers because K-means assumes a spherical shape of each cluster; it is difficult to deal with the class imbalance of pixels in the real world, which is completely ignored.

We propose a novel approach, termed Imbalance-Aware Dense Discriminative Clustering (IDDC), for USS. It addresses all these difficulties in a unified framework. An overview of IDDC is illustrated in Fig. 1. IDDC directly models the classification probability of each pixel conditioned on the image, without any assumption about the generative distribution or shape of a cluster. Its network consists of a backbone (i.e., a ViT pretrained on unlabeled images) that extracts dense features from an image and a segmentation head that predicts the pixel-wise class probabilities (through Softmax). Both components are learned in an end-to-end, discriminative, and self-supervised manner, by transferring the manifold structure of pixels in the ViT embedding space to the label space, alleviating the adverse impact of noise in pixel affinities, and handling cluster degeneration and pixel class imbalance. Concretely, we take advantage of the recent advancement of the self-supervised ViT: its dense embeddings of images well preserve the semantic similarity between pixels (Sect. 3.2). We leverage this manifold structure of pixels to supervise the learning of IDDC (Sect. 3.3). However, the pairwise pixel affinities are very noisy: pixels close in the ViT embedding space can belong to different classes. We have an in-depth investigation of how to handle the noise (Sect. 3.3). Finally, we introduce a new regularizer based on the Weibull function to model the skewed class distribution of pixels and avoid empty clusters (Sect. 3.4).

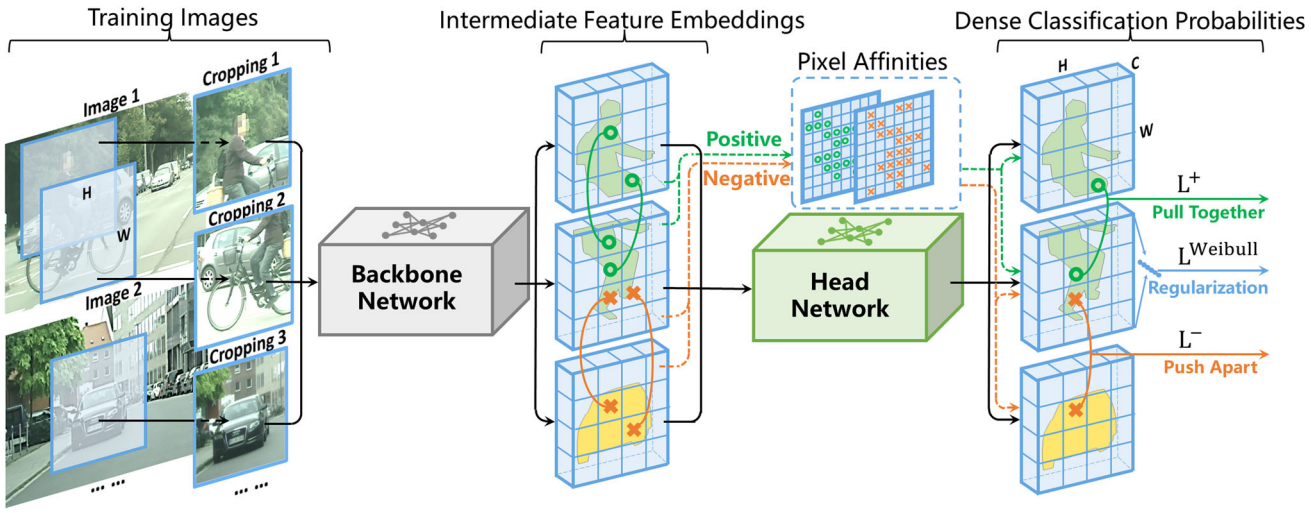

Fig. 1 Overview of Imbalance-Aware Dense Discriminative Clustering (IDDC) for unsupervised semantic segmentation (USS). IDDC directly models the classification probability of each pixel conditioned on the image. Its network consists of a backbone (i.e., a ViT pretrained on unlabeled images) that extracts dense features from an image and a head that predicts the pixel-wise class probabilities. Both components are learned in an end-to-end, discriminative, and self-supervised manner, through a novel objective function that consists of three loss terms. L^+ and L^- provide effective training signals for learning semantic segmentation by transferring the manifold structure of pixels in the ViT embedding space to the label space and tolerating noise in pixel affinities. L^{Weibull} is a new regularizer, based on the Weibull function, that avoids cluster degeneration and addresses pixel class imbalance

3.2 Dense Embedding as Segmentation Hint

We demonstrate that the dense embeddings of images learned by a self-supervised vision Transformer (ViT) are semantically meaningful. Although there is noise, two pixels closer in the ViT embedding space are more likely to belong to the same class.

Given an input image, a ViT-based backbone produces a dense feature map. Let $\boldsymbol{u}_i$ denote the ℓ_2-normalized embedding vector of the i-th pixel. We calculate the affinity between two pixels i and j through the cosine similarity of their embedding vectors: $\boldsymbol{u}_i \cdot \boldsymbol{u}_j$. Figure 2 shows the distributions of pairwise pixel affinities calculated on the Cityscapes training set (Cordts et al., 2016). We show the distributions corresponding to pixel pairs of the same class and pixel pairs of different classes in red and blue, respectively. The backbone is a ViT-small model with a patch size 16, pretrained by Zhou et al. (2021) on ImageNet (Deng et al., 2009) without labels.

We can make two observations from Fig. 2. First, pixel pairs of the same class are more likely to have high affinities than pixel pairs of different classes. If the affinity of two pixels is larger than 0.2 (the crossing point of the two lines in Fig. 2), they are more likely to be from the same class. This motivates us to leverage the affinity between two pixels in the ViT embedding space as a hint to identify their relationship in the label space, which can be used for learning semantic segmentation. Second, pixels of different classes can still

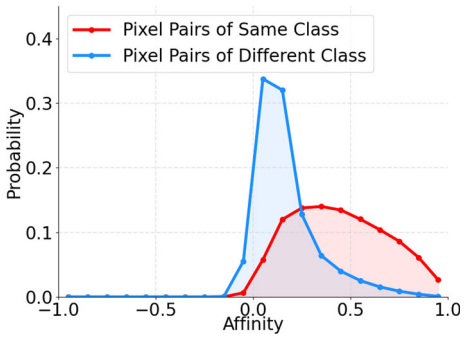

Fig. 2 Distributions of pairwise affinities between pixels of the same class (red) and of different classes (blue). The affinity between two pixels is calculated as their cosine similarity in the ViT embedding space

have high affinities, making the aforementioned hint noisy. It is critical to handle the noise for effective self-supervised learning.

3.3 Noise-tolerant Learning from Pixel Manifold

We introduce a basic objective function that learns USS by transferring the manifold structure of pixels in the ViT embedding space to the label space (Sect. 3.3.1) and alleviating the adverse impact of noise in pixel affinities (Sect. 3.3.2).

3.3.1 Learning by Transferring Pixel Manifold

The pairwise pixel affinities calculated in Sect. 3.2 characterize the manifold structure of pixels in the ViT embedding space. Our basic objective function learns USS by preserving this manifold structure in the label space. That is, two pixels close or distant in the ViT embedding space should also be close or distant in the label space. Concretely, our basic objective function consists of a positive term and a negative term. The former pulls the classification probabilities of every positive pixel pair together; the latter pushes the classification probabilities of every negative pixel pair away. Here, a positive pixel pair means two pixels whose affinity is larger than a threshold t^+; a negative pixel pair means two pixels whose affinity is smaller than a threshold t^-.

To obtain a positive sample pair, we first sample two partially overlapped croppings in the same image and then sample one pixel from each cropping to form a positive sample pair. The overlapping region between the two croppings guarantees there exist pixel pairs of the same class. The random geometric transformation perturbs the pixel embedding and augments data, which promotes robust learning and avoids overfitting. Let $\boldsymbol{u}_i$ and $\boldsymbol{u}_j^+$ denote the ℓ_2-normalized embedding vectors of two pixels from the two croppings, respectively, and their corresponding classification probabilities are $\boldsymbol{v}_i$ and $\boldsymbol{v}_j^+$. The positive loss term is formulated as:

$$L^+ = \frac{1}{N^+} \sum_{\forall i,j} w^+(\boldsymbol{u}_i \cdot \boldsymbol{u}_j^+; t^+) h^+(\boldsymbol{v}_i \cdot \boldsymbol{v}_j^+) \tag{1}$$

where $w^+(\boldsymbol{u}_i \cdot \boldsymbol{u}_j^+; t^+)$ selects positive pixel pairs by returning $\boldsymbol{u}_i \cdot \boldsymbol{u}_j^+$ if $\boldsymbol{u}_i \cdot \boldsymbol{u}_j^+$ is greater than t^+ and returning zero otherwise, N^+ counts the total number of positive pixel pairs, and h^+ is a monotonically decreasing function (as the objective function will be minimized). $\boldsymbol{v}_i \cdot \boldsymbol{v}_j^+$ can be interpreted as the probability that the two pixels belong to the same class if they are independent. In addition to rejecting pixel pairs that are not positive, $w^+(\boldsymbol{u}_i \cdot \boldsymbol{u}_j^+; t^+)$ uses the affinities of positive pixel pairs in the ViT embedding space to weigh their similarities in the label space. This accounts for the observation that a pixel pair with a larger affinity is more likely to belong to the same class. The design of h^+ is critical to handling noise and will be detailed in the following section.

To obtain a negative sample pair, we first sample two croppings respectively from two different images and then sample one pixel from each cropping to form a negative sample pair. The negative loss term is formulated as:

$$L^- = \frac{1}{N^-} \sum_{\forall i,j} w^-(\boldsymbol{u}_i \cdot \boldsymbol{u}_j^-; t^-) h^-(\boldsymbol{v}_i \cdot \boldsymbol{v}_j^-) \tag{2}$$

where $w^-(\boldsymbol{u}_i \cdot \boldsymbol{u}_j^-; t^-)$ selects negative pixel pairs by returning $1 - \boldsymbol{u}_i \cdot \boldsymbol{u}_j^-$ if $\boldsymbol{u}_i \cdot \boldsymbol{u}_j^-$ is smaller than t^- and returning zero otherwise, N^- counts the total number of negative pixel pairs, and h^- is a monotonically increasing function. Minimizing L^- will push two pixels further away in the label space if they are more distant in the ViT embedding space.

3.3.2 Tolerating Noise

The pairwise pixel affinities are noisy because two pixels of different classes can have a high affinity, as illustrated in Fig. 2. We explore different formulations of $h^+(\cdot)$ and $h^-(\cdot)$, and analyze their capabilities to tolerate the noise.

The logarithmic function is widely used in probability-related loss functions, such as the cross-entropy loss. We consider:

$$\begin{aligned} h^+(\boldsymbol{v}_i \cdot \boldsymbol{v}_j^+) &= -\log(\boldsymbol{v}_i \cdot \boldsymbol{v}_j^+) \\ h^-(\boldsymbol{v}_i \cdot \boldsymbol{v}_j^-) &= -\log(1 - \boldsymbol{v}_i \cdot \boldsymbol{v}_j^-) \end{aligned} \tag{3}$$

The logarithmic function is suitable for supervised learning on well-labeled data because it imposes an asymptotic infinitely large penalty on incorrect classifications. In USS, however, noise is unavoidable. Pixel pairs of the same class can have small affinities, and pixel pairs of different classes can have large affinities. Using a logarithmic function in these scenarios will lead to large gradients but in the wrong direction. It will not allow the network to ignore the noise (because of the large penalty and gradient), thus severely disturbing its learning.

The analysis above motivates us to look for a function whose range is finite in the domain [0, 1]. The linear function is the simplest one that satisfies this requirement. We consider:

$$\begin{aligned} h^+(\boldsymbol{v}_i \cdot \boldsymbol{v}_j^+) &= 1 - \boldsymbol{v}_i \cdot \boldsymbol{v}_j^+ \\ h^-(\boldsymbol{v}_i \cdot \boldsymbol{v}_j^-) &= \boldsymbol{v}_i \cdot \boldsymbol{v}_j^- \end{aligned} \tag{4}$$

The linear function has a constant slope and will result in the same gradient regardless of the loss value, i.e., the inconsistency between two pixels' affinity and their classes. Similar to the ℓ_1-norm used in robust regression (Xu et al., 2008), the linear function can tolerate noise as long as the noise is not too much. However, a potential issue is that the learning process could focus on the easy cases and ignore the hard ones, because the loss function treats them equally and optimizing the former is easier. This hinders effective learning.

Finally, we consider an exponential function:

$$\begin{aligned} h^+(\boldsymbol{v}_i \cdot \boldsymbol{v}_j^+) &= \exp(1 - \boldsymbol{v}_i \cdot \boldsymbol{v}_j^+) \\ h^-(\boldsymbol{v}_i \cdot \boldsymbol{v}_j^-) &= \exp(\boldsymbol{v}_i \cdot \boldsymbol{v}_j^-) \end{aligned} \tag{5}$$

Like a linear function, an exponential function has a finite range of output values and gradients in the domain [0, 1]. But unlike a linear function, an exponential function will have larger gradients when the loss values are larger and thus will push harder on hard cases.

In sum, both the linear function and the exponential function can better tolerate noise than the logarithmic function because of their limited range of output values and gradients in the domain [0, 1]. Compared with the linear function, the exponential function puts more focus on the hard cases than the easy ones. In USS, it is impossible to distinguish between hard cases and noise. But our experiments indicate that the exponential function works better than the linear function, and both of them outperform logarithmic function by a large margin. The reason could be that both tolerating noise and handling hard cases are important to effective learning; the exponential function can achieve a good balance between them.

3.4 Addressing Cluster Degeneration and Pixel Class Imbalance

While the model introduced till now can be used for learning USS, two critical issues remain. The first issue is cluster degeneration: there often exist empty clusters. The second issue is pixel class imbalance: the class distribution of pixels in the real world is highly skewed, caused by the different sizes and occurring frequencies of each object and stuff.

Previous discriminative clustering methods (Ji et al., 2019; Krause et al., 2010; Ouali et al., 2020; Van Gansbeke et al., 2020) address the cluster degeneration problem via an entropy-based regularizer:

$$L^{\text{entropy}} = \sum_c p_c \log(p_c), \quad \text{where} \quad p_c = \frac{1}{N} \sum_{\forall i} v_{i,c} \tag{6}$$

where i and c index a sample and a class, respectively, $v_{i,c}$ is the predicted probability of the c-th class on the i-th sample, N is the total number of samples, and p_c is the frequency of the c-th class occurring in the training data. It avoids empty clusters by enforcing a uniform distribution of classes.

However, the class distribution of pixels in the real world is highly skewed rather than uniform. For example, in the Cityscapes dataset, the road class dominates and its number of pixels is more than one thousand times larger than those of rare classes. This class imbalance problem has been ignored by prior work in USS.

To address class degeneration and imbalanced class distribution, we introduce a regularization term to regularize the learning process. This term serves two purposes. First, it imposes a significant penalty on the presence of empty clusters. Second, it tolerates a skewed pixel class distribution.

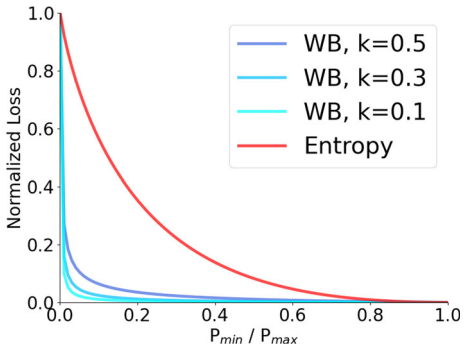

Fig. 3 Comparison of two regularizers respectively based on the Weibull (WB) function and entropy. The x-axis is the ratio of the number of pixels of the rarest class to that of the commonest class. The ratio ranges from 0.001 to 1. Both regularizers are normalized to have a unit upper bound

As shown in Fig. 3, given the x-axis represents the ratio of pixels distributed to two categories, an L-shaped function ideally aligns with the aforementioned two purposes. On one hand, the vertical part of "L" generates a large loss value, effectively penalizing the occurrence of empty clusters. On the other hand, the horizontal part of "L" produces near-zero loss values on an imbalanced class distribution, enabling pixels to be clustered based on their distance in the embedding space. We model the regularization term as a Weibull function (Murthy et al., 2004; Weibull, 1951) because its shape can be flexibly adjusted by its shape parameter $k \in (0, 1)$ and it is differentiable. The smaller the value of k, the closer the Weibull distribution is to an L shape. Concretely, the proposed regularizer is formulated as:

$$L^{\text{Weibull}} = \sum_c k p_c^{k-1} \exp(-p_c^k) \tag{7}$$

It could be observed from Fig. 3 that, similar to the widely used entropy regularizer in clustering (Van Gansbeke et al., 2020), the Weibull regularizer will cause a large penalty in case of any empty clusters, i.e., when the ratio is close to zero. But different from the entropy regularizer, the Weibull regularizer will cause a much smaller penalty than the entropy regularizer when the classes are imbalanced, i.e., when the ratio is much lower than one.

3.5 Overall Objective Function

The overall objective function of IDDC combines the basic objective function in Sect. 3.3 and the Weibull regularizer in Sect. 3.4:

$$L = L^{+} + \lambda_1 L^{-} + \lambda_2 L^{\text{Weibull}} \tag{8}$$

where λ_1 and λ_2 are trade-off hyper-parameters. L^+ and L^- provide effective training signals for learning semantic segmentation by transferring the manifold structure of pixels in the ViT embedding space to the label space and tolerating noise in pixel affinities. L^{Weibull} avoids cluster degeneration and addresses pixel class imbalance.

4 Experiments

4.1 Datasets and Evaluation Metrics

We validate the effectiveness of IDDC on three large-scale real-world datasets. Cityscapes (Cordts et al., 2016) is annotated for street view understanding, including 27 classes. It has 2,957 and 500 finely annotated images for training and validation, respectively. The resolution of each image is 1024×2048. COCO-Stuff-171 (Caesar et al., 2018) is a large-scale scene-centric dataset. There are 117,266 training images and 5,000 validation images, including 171 categories (80 things and 91 stuff). COCO-Stuff-27 (Caesar et al., 2018) is a simplified version of the COCO-Stuff-171 dataset, and is widely used for evaluating USS. Following previous arts (Hamilton et al., 2022; Ji et al., 2019), 49,629 images are used for training and 2,175 images for validation. The 80 things and 91 stuff categories are merged into 27 categories (12 objects and 15 stuff) for evaluating USS.

Following previous works (Hamilton et al., 2022; Cho et al., 2021; Ji et al., 2019), the Hungarian matching algorithm (Kuhn, 1955) is used to align the discovered clusters to the annotated classes for evaluation and visualization. The performance is evaluated via two metrics: mean Intersection over Union (mIoU) and Accuracy (ACC). IoU measures the number of pixels common between the ground truth and prediction segments of a class divided by the total number of pixels present across both segments. mIoU is the average IoU over all classes. ACC is the number of correctly classified pixels divided by the number of all pixels.

4.2 Implementation Details

4.2.1 Network Architecture

The backbone networks on all datasets are ViT-based models (Touvron et al., 2021). They are pretrained by unsupervised methods including DINO Caron et al. (2021) and iBoT Zhou et al. (2021) on the ImageNet dataset Deng et al. (2009) without labels. The segmentation head consists of two convolutional layers activated by ReLU and one convolutional layer terminated by Softmax.

4.2.2 Training

Our approach is implemented in Pytorch. We train the model with the Adam optimizer and a batch size of 64. The initial learning rate is 5e−4 for the head network and 5e−7 for the backbone. We use a polynomial learning rate policy: the initial learning rate is multiplied by $(1 - iter/max\_iter)^{power}$ and $power$ equals 0.9. The numbers of training epochs are 5, 20, and 50 for COCO-Stuff-27, COCO-Stuff-171, and Cityscapes, respectively. The batch size is 64. In non-end-to-end training, the backbone network is fixed, and the head is trainable. In our end-to-end setting, the backbone network is fixed during the initial two-fifths and the final one-fifth of training epochs. The former avoids the disturbance of the pretrained backbone caused by the back propagation from the randomly initialized head network. The latter guarantees a fixed supervisory signal for stable learning. Input images are randomly resized with a ratio between 0.8 and 1.2, randomly flipped, and randomly cropped to 224×224. Following previous works (Van Gansbeke et al., 2021; Hamilton et al., 2022; Ji et al., 2019; Cho et al., 2021), we set the number of clusters as the number of ground truth classes. The training takes less than two hours on a single NVIDIA V100 GPU card, so IDDC is quite efficient.

4.2.3 Hyper-parameters

For COCO-Stuff-27, in the ViT-S/8 experiment, λ_1 and λ_2 in the overall objective function Eq. (8) are set to 1.4 and 0.25, respectively. The two thresholds t^+ and t^- used for selecting positive and negative pixel pairs are set to 0.2 and 0.12, respectively. In the ViT-S/16 experiment, λ_1, λ_2, t^+, and t^- are set to 1.4, 0.4, 0.2, and 0.12, respectively. For COCO-Stuff-171, λ_1, λ_2, t^+, and t^- are 9.0, 0.61, 0.15, and 0.15, respectively. For Cityscapes, they are 0.38, 0.22, 0.2, and 0.28 in ViT-B/8, and they are 0.25, 0.3, 0.15, and 0.25 in ViT-S/8. We have conducted extensive ablation studies (Sect. 4.4) on the impact of these hyper-parameters.

4.3 Comparison with State-of-the-Art Methods

We compare our method with the state-of-the-art methods on COCO-Stuff-27, COCO-Stuff-171, and Cityscapes. For fair and comprehensive comparisons, experiments are conducted on different ViT models pretrained by DINO (Caron et al., 2021) on the ImageNet dataset without labels. The results on COCO-Stuff-27 are reported in Table 1. IDDC outperforms all prior methods by a large margin. Performance on COCO-Stuff-171 is demonstrated in Table 2. Compared to the other two datasets with 27 target categories, COCO-Stuff-171 has 171 categories. IDDC outperforms other methods, highlighting its superior scalability to datasets with more categories. For Cityscapes, previous methods (Hamilton et al.,

Table 1 Comparison with the state-of-the-art methods on the COCO-Stuff-27 validation dataset

Methods (Pub'Year)	Backbone	Acc%	mIoU%
Modified DC (ECCV'18) (Caron et al., 2018)	R18+FPN	32.2	9.8
IIC (ICCV'19) (Ji et al., 2019)	R18+FPN	21.8	6.7
PiCIE (CVPR'21) (Cho et al., 2021)	R18+FPN	48.1	13.8
PiCIE+H (CVPR'21) (Cho et al., 2021)	R18+FPN	50.0	14.4
STEGO (ICLR'22) (Hamilton et al., 2022)	ViT-S/16	52.5	23.7
HP (CVPR'23) (Seong et al., 2023)	ViT-S/16	54.5	24.3
IDDC	ViT-S/16	**59.9**	**25.8**
ACSeg (CVPR'23) (Seong et al., 2023)	ViT-S/8	–	16.4
TransFGU (ECCV'22) (Yin et al., 2022)	ViT-S/8	52.7	17.5
HP (CVPR'23) (Seong et al., 2023)	ViT-S/8	57.2	24.6
STEGO (ICLR'22) (Hamilton et al., 2022)	ViT-S/8	47.7	24.0
STEGO+CRF (ICLR'22) (Hamilton et al., 2022)	ViT-S/8	48.3	24.5
IDDC	ViT-S/8	58.3	25.5
IDDC+CRF	ViT-S/8	**58.8**	**25.8**

ViT backbones are pretrained using DINO without labels
Best results are highlighted in bold

Table 2 Comparison with state-of-the-art methods on the COCO-Stuff-171 dataset

Methods (Pub'Year)	Backbone	Acc%	mIoU%
PiCIE (CVPR'21) (Cho et al., 2021)	ViT-S/8	18.5	3.0
TransFGU (ECCV'22) (Yin et al., 2022)	ViT-S/8	**34.3**	11.9
IDDC	ViT-S/8	**34.3**	**12.2**

ViT backbones are pretrained using DINO without labels
Best results are highlighted in bold

2022; Cho et al., 2021) are evaluated on the center cropping of the validation images. Specifically, only the 1024×1024 center region (resized to 320×320) from the entire image, whose size is 1024×2048, is used. In this work, we evaluate our method on both the center cropping and the entire image in Table 3 and Table 4, respectively. In all settings, IDDC outperforms existing state-of-the-art methods.

4.4 Ablation Studies

We conduct extensive controlled experiments to validate the effectiveness of designs in IDDC. In Sec 4.4.1, we prove the existence of the cluster degeneration phenomenon and the pixel class imbalance problem. We show how the proposed regularizer solves both problems in single shot. In Sec 4.4.2, we compare different formulations of positive and negative terms for noise-tolerant learning. In Sec 4.4.3, we validate the necessity of all three terms in the overall objective function and the impact of their trade-off hyper-parameters λ_1 and λ_2. In Sect. 4.4.4, we examine different settings when selecting the positive and negative pixel pairs. In Sect. 4.4.5, we explore the impact of the number of training epochs. The generalization capability of IDDC to different backbones is validated in Sect. 4.4.6. Finally, the effectiveness of the end-to-end training is demonstrated in Sect. 4.4.7

We conduct the ablation experiments on Cityscapes full images. To conduct a large number of experiments more efficiently, we adopt a lighter backbone (ViT-small with a patch size 16 pretrained by iBoT (Zhou et al., 2021)) and a linear head. We keep the backbone fixed and only train the head before Sect. 4.4.7. Our overall objective function includes three terms; changing or removing one term can lead to a sub-optimal combination of the trade-off hyper-parameters λ_1 and λ_2. Thus, we examine the combination of different values of λ_1 and λ_2 for each design to seek optimal performance. They are demonstrated in the form of line graphs, such as Fig. 5, where each point represents an experiment.

4.4.1 Regularization Term

Figure 4 validates the effectiveness of the regularization term. It shows that a) cluster degeneration happens without a regularization term, b) the Weibull function performs better than the widely used entropy, and c) the proposed regularizer is more suitable for learning from class imbalanced data.

Figure 4a shows that cluster degeneration happens when learning with positive and negative terms only. The blue line with star marks indicates that the number of discovered classes is smaller than expected, even though the true

Table 3 Comparison with state-of-the-art methods on the Cityscapes validation dataset (27 classes)

Methods (Pub'Year)	Backbone	Acc%	mIoU%
Modified DC (ECCV'18) (Cho et al., 2021)	R18+FPN	40.7	7.1
IIC (ICCV'19) (Ji et al., 2019)	R18+FPN	47.9	6.4
PiCIE (CVPR'21) (Cho et al., 2021)	R18+FPN	65.6	12.3
HP (CVPR'23) (Seong et al., 2023)	ViT-B/8	**79.5**	18.4
STEGO (ICLR'22) (Hamilton et al., 2022)	ViT-B/8	66.4	19.6
STEGO+CRF (ICLR'22) (Hamilton et al., 2022)	ViT-B/8	73.2	21.0
IDDC	ViT-B/8	78.0	21.6
IDDC+CRF	ViT-B/8	78.6	**21.7**
TransFGU (ECCV'22) (Yin et al., 2022)	ViT-S/8	77.9	16.8
HP (CVPR'23) (Seong et al., 2023)	ViT-S/8	**80.1**	18.4
IDDC	ViT-S/8	79.9	**22.0**

All methods are evaluated on the center croppings of the original images. ViT backbones are pretrained using DINO without labels
Best results are highlighted in bold

Table 4 Comparison with state-of-the-art methods on the Cityscapes validation dataset (27 classes)

Methods (Pub'Year)	Backbone	Acc%	mIoU%
STEGO (ICLR'22) (Hamilton et al., 2022)	ViT-B/8	64.0	19.8
STEGO (ICLR'22)+CRF (Hamilton et al., 2022)	ViT-B/8	68.9	20.9
IDDC	ViT-B/8	76.3	22.2
IDDC+CRF	ViT-B/8	**77.0**	**22.4**

All methods are evaluated on the full-size original images. ViT backbones are pretrained using DINO without labels
Best results are highlighted in bold

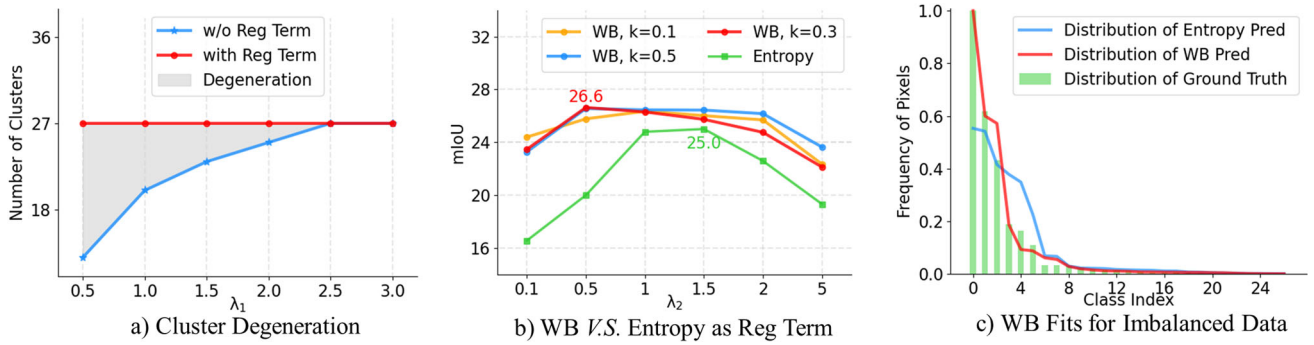

Fig. 4 Ablation study on the regularization term. **a** shows the cluster degeneration happens when training without our regularization term, but it is avoided when training with our regularization term. **b** demonstrates the superiority of the Weibull (WB) function over the entropy as the regularization term. **c** compares the class distribution of pixels obtained by taking the Weibull function or entropy as the regularizer with the ground truth distribution

number of classes is set in the objective function. In contrast, the red line with round marks shows that degeneration is avoided after including the regularization term. Although using a large weight for the negative term could also avoid degeneration in the absence of the regularization term, seeking it is time-consuming in practice. Furthermore, the best performance without the regularization term is 22.7% mIoU, which is 3.9% lower than that with the regularization term.

Having verified the necessity of the regularization term in avoiding clustering degeneration, we now explore an optimal form of it. Previous methods (Krause et al., 2010; Van Gansbeke et al., 2020) take entropy as regularization for the classification task, based on the assumption that labels spread uniformly across categories. This works well on balanced image classification datasets such as ImageNet and CIFAR-10/100. However, labels are highly imbalanced in the semantic segmentation task, as shown by the green bars in Fig. 4c. Thus, we propose a regularizer based on the Weibull function. It is an "L" shape function that imposes a large penalty if there exist any empty clusters but only causes

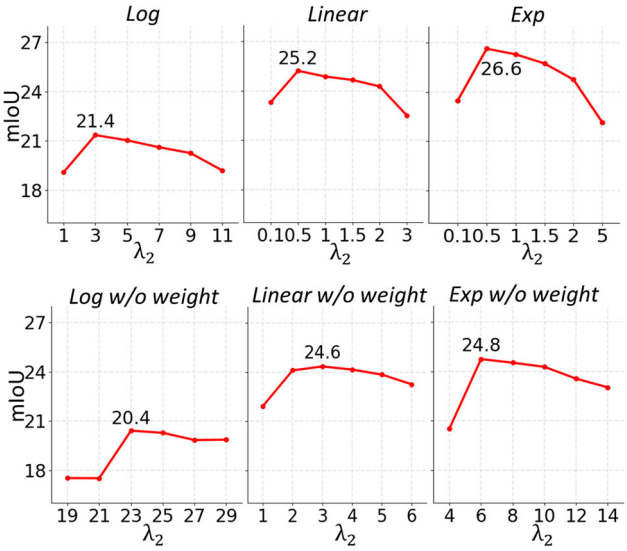

Fig. 5 Ablation study on different formulations of h^+ and h^- and the effectiveness of their weights w^+ and w^- on the Cityscapes validation dataset

small loss values if the numbers of pixels are highly imbalanced in different clusters. The line graph in Fig. 4b shows the performance of different Weibull function settings and their superiority over the entropy-based regularizer. The red line and blue line in Fig. 4c respectively show the ranked numbers of pixels in each discovered class of our proposed regularizer and the entropy-based regularizer. The results indicate that the model trained with the Weibull regularizer fits well for the imbalanced class distribution. Thus, it is more suitable for IDDC to learn from imbalanced data.

4.4.2 Forms of Positive and Negative Terms

Our positive and negative terms in the overall objective function are designed to learn USS by transferring the pixel manifold structure in the ViT embedding space to the label space and tolerating noise in pixel affinities. We examine their effectiveness in this ablation study. The results are reported in Fig. 5. The log, linear, and exp in the figure respectively correspond to the three formulations of h^+ and h^- introduced in Sec 3.3.2. We also report results obtained after removing the weights w^+ and w^- from h^+ and h^-. Table 5 summarizes their best performances.

We could observe that the exponential function performs the best, and removing weights will degrade the performance. Both the linear function and the exponential function perform much better than the logarithmic function because of their capability to tolerate noise, as discussed in Sect. 3.3.2. The exponential function performs better than the linear function because the former has a larger penalty on hard cases. The weights w^+ and w^- benefit the performance because pixel pairs closer in the embedding space are more likely to be from the same class.

Table 5 Ablation study on different formulations of h^+ and h^- (denoted as h) and the effectiveness of their weights w^+ and w^- on the Cityscapes validation dataset

h	Weights	Acc%	mIoU%
Log	–	68.8	20.4
Linear	–	77.2	24.4
Exp	–	79.5	24.8
Log	✓	76.8	21.4
Linear	✓	78.7	25.2
Exp	✓	**80.2**	**26.6**

Best results are highlighted in bold

4.4.3 Necessity of Each Term and Impact of Trade-off Hyper-Parameters

We conduct ablation experiments to demonstrate the necessity of each term in the overall objective function and the impact of their trade-off hyper-parameters λ_1 and λ_2, by training without the negative term, without the regularization term, and with different values of λ_1 and λ_2. The results are shown in Fig. 6. Their optimal performances together with a K-means baseline are summarized in Table 6.

We could observe that directly clustering the ViT features using K-means achieves only 12.4% mIoU and 50.1% Acc. The performance is improved to 26.6 % mIoU and 80.2% Acc using our method, which demonstrates the effectiveness of the proposed IDDC. Table 6 shows removing the regularization term or the negative term leads to a performance drop to 22.7% mIoU and 23.8% mIoU, respectively. This indicates that the two terms are complementary to each other and could work harmoniously for better performance.

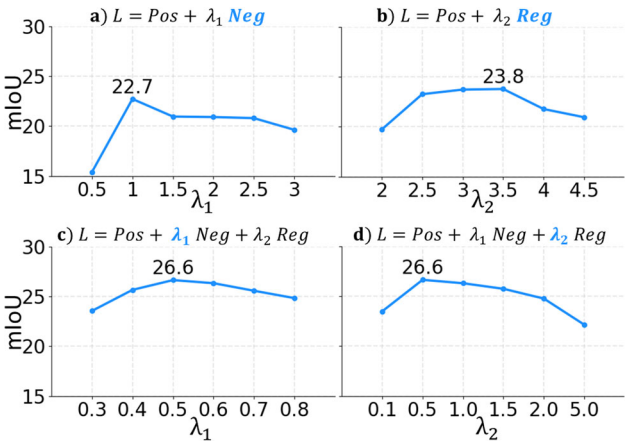

Fig. 6 Ablation study on the necessity of the three terms in the overall objective function and the impact of trade-off hyper-parameters λ_1 and λ_2 on the Cityscapes validation dataset

Table 6 Ablation study on the necessity of the positive, negative, and regularization terms in the overall objective function on the Cityscapes validation dataset

Pos	Neg	Reg	Acc%	mIoU%
–	–	–	50.1	12.4
✓	✓	–	70.2	22.7
✓	–	✓	77.5	23.8
✓	✓	✓	**80.2**	**26.6**

The baseline (first row) clusters the ViT embeddings of each pixel through K-means

4.4.4 Pixel Pair Selection

We use two thresholds t^+ and t^- to select positive and negative pixel pairs. Figure 7a demonstrates the impact of different values of t^+ and t^- on the performance. We can observe that the performance is the best when both thresholds are set to 0.2. It is worth mentioning that the experimental results are consistent with the statistical results mentioned earlier in Fig. 2. Both experiments are conducted on an iBot-pretrained ViT-small model with a patch size 16. The intersection of the two lines in Fig. 2 shows that the value around 0.2 is a good threshold, meaning when the affinity of two pixels is larger than about 0.2, the distribution of pixel pairs of the same class on the affinity axis is more concentrated than that of pixel pairs of different classes, and vice versa. When adapting our method to new datasets, we can use the statistical results on a small number of labeled images to roughly determine the hyperparameters t^+ and t^-.

4.4.5 Number of Training Epochs

We validate whether the number of training epochs will affect the performance. It is difficult to select an intermediate optimal model in unsupervised applications without the guidance of human labels. Thus, if there is a large performance fluctuation during training, the utility of the model will be degraded. Figure 7b shows the validation performance of the model after each training epoch. We could observe that there is no obvious performance fluctuation, which means IDDC is stable.

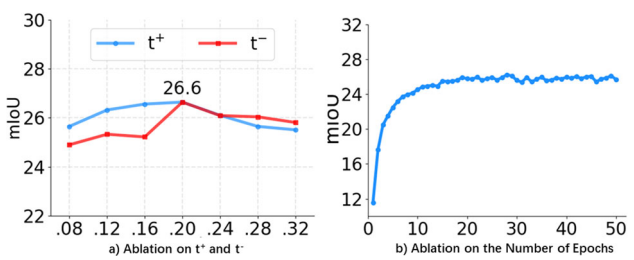

Fig. 7 Ablation study of **a** the thresholds t^+ and t^- on the Cityscapes validation dataset, **b** and the performance fluctuation at different numbers of training epochs on the Cityscapes validation dataset

4.4.6 Generalization to Different Backbones

We explore whether the proposed method is effective on other backbones. The results are shown in Table 7. ViT (Touvron et al., 2021) and XCiT (Ali et al., 2021) are different vision Transformer models that could produce dense feature embeddings. XCiT follows a new self-attention mechanism that operates across feature channels rather than tokens. iBoT (Zhou et al., 2021) and DINO (Caron et al., 2021) are two self-supervised learning strategies. iBoT takes advantage of its online tokenizer, which simplifies the training step.

We could draw three conclusions from the results. First, IDDC could be generalized to other backbones pretrained by different self-supervised learning strategies. Actually, it outperforms previous state-of-the-art methods in all settings. Second, a backbone pretrained on a larger dataset may lead to better performance. ImageNet 1k (IN-1K) with 1.2M images is a subset of ImageNet 22K (IN-22K) with 14 M images (Deng et al., 2009). We can see that models trained on IN-22K perform better. Third, measured by mIoU, models with smaller patch sizes are likely to perform better. The patch size changes the resolution of the intermediate visual representations. For example, given an input image size 224×224, the size of the feature map is 14×14 if the patch size is 16 and is 28×28 if the patch size is 8. In the segmentation task, a smaller patch size offers more detailed spatial information and hence benefits the performance.

4.4.7 Effectiveness of End-to-End Training

The end-to-end mechanism allows simultaneous training of all modules so that they could coordinate well with each other and learn toward the same target. Results using trainable and fixed backbone networks (i.e. end-to-end and w/o

Table 7 Ablation study on different backbone networks, different self-supervised training methods, different unlabeled datasets for pretraining, and different patch sizes on the Cityscapes validation dataset

Method	Data	Arch	Patch	Acc%	mIoU%
iBoT	IN-1K	ViT-Base	16	73.6	26.4
iBoT	IN-22K	ViT-Base	16	80.5	26.3
DINO	IN-1K	ViT-Small	16	76.8	24.5
DINO	IN-1K	ViT-Small	8	78.9	23.1
DINO	IN-1K	ViT-Base	16	77.4	19.8
DINO	IN-1K	ViT-Base	8	76.3	22.2
DINO	IN-1K	XCiT-Small	16	77.5	22.7
DINO	IN-1K	XCiT-Small	8	76.6	23.1

Table 8 Ablation study of the effectiveness of the end-to-end training

Training strategy	Acc%	mIoU%
w/o end-to-end	80.5	26.3
end-to-end	80.2	27.7

end-to-end) are demonstrated in Table 8 respectively. We could observe that using the end-to-end training achieves comparable ACC and superior mIoU.

4.5 Visualization

Figures 8 and 9 visualize the segmentation results of our proposed IDDC and the current state-of-the-art method STEGO (Hamilton et al., 2022) on the Cityscapes and the COCO-Stuff datasets respectively. STEGO addresses USS in a two-stage learning framework. Fig. 8 shows qualitative results obtained on Cityscapes. We can observe that IDDC can rectify two deficiencies of STEGO. The first deficiency is the local segmentation chaos in regions containing multiple objects and stuff, e.g., the crowded street in d), e), and f). The second deficiency is the wrongly classified segments, e.g., the sidewalk and buildings in b). IDDC can largely address these deficiencies because its end-to-end and discriminative learning facilitates better class separation and seamlessly coordinates representation learning and pixel clustering.

Figuure 9 shows qualitative results obtained on COCO-Stuff. The aforementioned two problems still exist in the results of STEGO and could be addressed by our IDDC. Examples of the local segmentation chaos could be seen in a) b) c) and e), where noisy segmentation happens at the boundary region of objects and stuff. Representative example of wrongly classified segments could be seen in b), where the wave of water is regarded as a category different from the ocean and river.

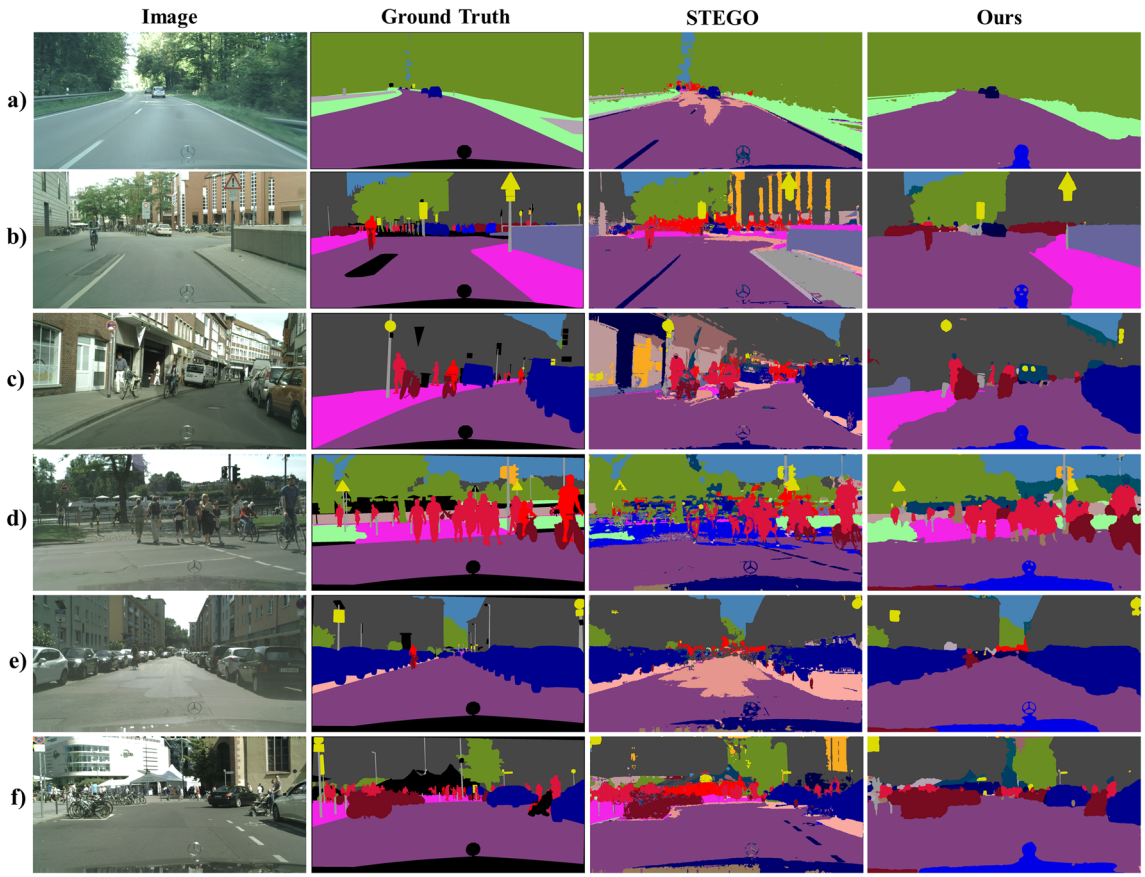

Fig. 8 Qualitative results on the Cityscapes dataset

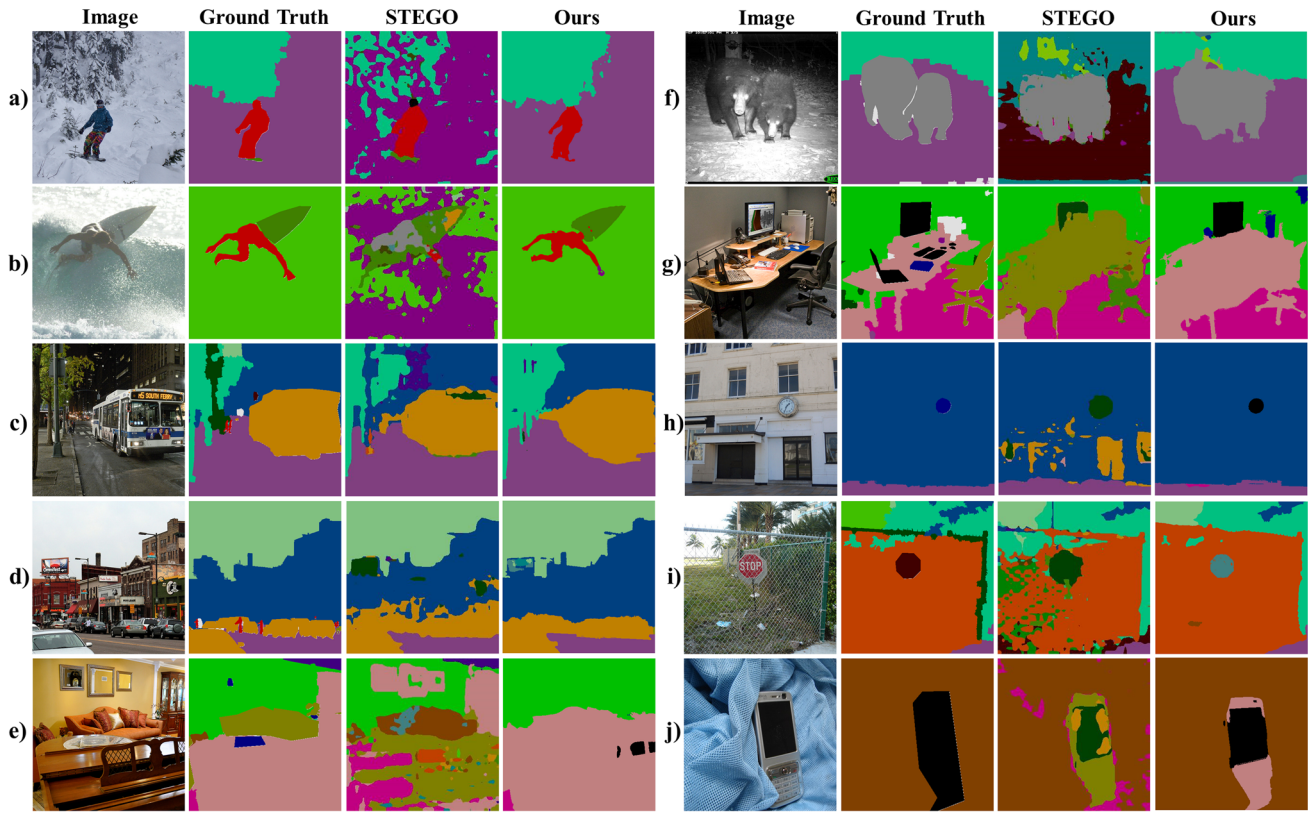

Fig. 9 Qualitative results on the COCO-Stuff dataset

5 Conclusion

This paper introduces a novel approach, termed Imbalance-Aware Dense Discriminative Clustering (IDDC), for unsupervised semantic segmentation (USS). IDDC directly models the classification probability of each pixel conditioned on the image and learns pixel-wise feature representation and dense discriminative clustering in an end-to-end and self-supervised manner. We propose a novel objective function that learns IDDC by transferring the manifold structure of pixels in the ViT embedding space to the label space, tolerating the noise in pixel affinities, and addressing pixel class imbalance via a new Weibull regularizer. IDDC overcomes the difficulties of previous methods in coordinating representation learning and pixel clustering, handling outliers and noise, and modeling the skewed class distribution. IDDC significantly outperforms all previous state-of-the-art methods on two large-scale real-world datasets. Extensive ablation studies demonstrate the effectiveness of each individual design.

Acknowledgements This work was supported in part by Wei Tang's startup funds from the University of Illinois Chicago and the National Science Foundation (NSF) award CNS-1828265.

Data Availibility The datasets generated during and/or analysed during the current study are respectively available in the GitHub repository at https://github.com/nightrome/cocostuff, and in the Cityscapes repository at https://www.cityscapes-dataset.com/.

References

Ahn, J., & Kwak, S. (2018). Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4981–4990.

Alexey, D., Fischer, P., Tobias, J., Springenberg, M.R., & Brox, T. (2015). Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(9), 1734–1747

Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al. (2021). Xcit: Cross-covariance image transformers. *Advances in Neural Information Processing Systems, 34*, 20014–20027.

Alonso, I., Sabater, A., Ferstl, D., Montesano, L., & Murillo, A.C.(2021). Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8219–8228.

Barber, D., & Agakov, F.(2005). Kernelized infomax clustering. In *Advances in neural information processing systems*, pp. 17–24.

Bojanowski, P., & Joulin, A.(2017). Unsupervised learning by predicting noise. In *International conference on machine learning*, pp. 517–526 . PMLR.

Bridle, J., Heading, A., & MacKay, D. (1991). Unsupervised classifiers, mutual information and'phantom targets. In *Advances in neural information processing systems*, pp. 1537–1544.

Caesar, H., Uijlings, J., & Ferrari, V.(2018). Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218.

Caron, M., Bojanowski, P., Joulin, A., & Douze, M.(2018). Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A.(2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660.

Chang, Y.-T., Wang, Q., Hung, W.-C., Piramuthu, R., Tsai, Y.-H., & Yang, M.-H.(2020). Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8991–9000.

Chang, J., Wang, L., Meng, G., Xiang, S., & Pan, C.(2017). Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pp. 5879–5887.

Chen, X., & He, K. (2021). Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607 . PMLR.

Cheng, B., Schwing, A., & Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems, 34*, 17864–17875.

Cho, J.H., Mall, U., Bala, K., & Hariharan, B (2021) Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16794–16804.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B.(2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on computer vision and pattern recognition*, pp. 248–255 . IEEE.

Doersch, C., Gupta, A., & Efros, A.A.(2015). Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430.

Gao, S., Li, Z.-Y., Yang, M.-H., Cheng, M.-M., Han, J., & Torr, P. (2023). Large-scale unsupervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(6), 7457–7476.

Ghasedi Dizaji, K., Herandi, A., Deng, C., Cai, W., & Huang, H.(2017). Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5736–5745.

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E.D., Le, Q.V., & Zoph, B.(2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2918–2928.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems, 33*, 21271–21284.

Gu, J., Kwon, H., Wang, D., Ye, W., Li, M., Chen, Y.-H., Lai, L., Chandra, V., & Pan, D.Z.(2022). Multi-scale high-resolution vision transformer for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12094–12103.

Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., & Freeman, W.T. (2022). Unsupervised semantic segmentation by distilling feature correspondences. In *International conference on learning representations*, pp. 1–26.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R.(2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738.

He, K., Girshick, R., & Dollár, P.(2019). Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4918–4927.

Hou, Y., Zhu, X., Ma, Y., Loy, C.C., & Li, Y.(2022). Point-to-voxel knowledge distillation for lidar semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8479–8488.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., & Markham, A.(2020). Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11108–11117.

Hung, W.-C., Jampani, V., Liu, S., Molchanov, P., Yang, M.-H., & Kautz, J.(2019). Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 869–878.

Ji, X., Henriques, J.F., & Vedaldi, A.(2019). Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9865–9874.

Ji, W., Yu, S., Wu, J., Ma, K., Bian, C., Bi, Q., Li, J., Liu, H., Cheng, L., & Zheng, Y.(2021). Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*(CVPR), pp. 12341–12351.

Kalluri, T., Varma, G., Chandraker, M., & Jawahar, C.(2019). Universal semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5259–5270.

Ke, Z., Qiu, D., Li, K., Yan, Q., & Lau, R.W. (2020). Guided collaborative training for pixel-wise semi-supervised learning. In *european conference on computer vision*, pp. 429–445 . Springer.

Komodakis, N., & Gidaris, S.(2018). Unsupervised representation learning by predicting image rotations. In *International conference on learning representations (ICLR)*.

Krause, A., Perona, P., & Gomes, R.(2010). Discriminative clustering by regularized information maximization. Advances in Neural Information Processing Systems **23**.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly,2*(1–2), 83–97.

Kwon, D., & Kwak, S.(2022). Semi-supervised semantic segmentation with error localization network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9957–9967.

Lai, X., Tian, Z., Jiang, L., Liu, S., Zhao, H., Wang, L., & Jia, J.(2021). Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1205–1214.

Lee, S., Lee, M., Lee, J., & Shim, H.(2021). Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5495–5505.

Li, K., Wang, Z., Cheng, Z., Yu, R., Zhao, Y., Song, G., Liu, C., Yuan, L., & Chen, J.(2023). Acseg: Adaptive conceptualization for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, pp. 7162–7172.

Li, X., Zhou, Y., Zhang, Y., Zhang, A., Wang, W., Jiang, N., Wu, H., & Wang, W.(2021). Dense semantic contrast for self-supervised visual representation learning. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 1368–1376.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., & Dong, L., *et al.* (2022). Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12009–12019.

Liu, M., Schonfeld, D., & Tang, W.(2021). Exploit visual dependency relations for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9726–9735.

Lu, Y., Cheung, Y.-M., & Tang, Y. Y. (2019). Self-adaptive multiprototype-based competitive learning approach: A k-means-type algorithm for imbalanced data clustering. *IEEE Transactions on Cybernetics, 51*(3), 1598–1612.

Melas-Kyriazi, L., Rupprecht, C., Laina, I., & Vedaldi, A.(2022). Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8364–8375.

Mendel, R., Souza, L.A.d., Rauber, D., Papa, J.P., & Palm, C.(2020). Semi-supervised segmentation based on error-correcting supervision. In *European conference on computer vision*, pp. 141–157 . Springer.

Mirsadeghi, S. E., Royat, A., & Rezatofighi, H. (2021). Unsupervised image segmentation by mutual information maximization and adversarial regularization. *IEEE Robotics and Automation Letters, 6*(4), 6931–6938.

Mittal, S., Tatarchenko, M., & Brox, T. (2019). Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence, 43*(4), 1369–1379.

Murthy, D.P., Xie, M., & Jiang, R.(2004). Weibull models. Wiley

Ng, A., & Jordan, M.(2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pp. 841–848.

Noroozi, M., & Favaro, P.(2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pp. 69–84 . Springer.

Ouali, Y., Hudelot, C., & Tami, M. (2020). Autoregressive unsupervised image segmentation. In *European conference on computer vision*, pp. 142–158 . Springer.

Pang, B., Li, Y., Zhang, Y., Peng, G., Tang, J., Zha, K., Li, J., & Lu, C.(2022). Unsupervised representation for semantic segmentation by implicit cycle-attention contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, pp. 2044–2052.

Peng, C., Myronenko, A., Hatamizadeh, A., Nath, V., Siddiquee, M.M.R., He, Y., Xu, D., Chellappa, R. (2022) Yang, D. Hypersegnas: Bridging one-shot neural architecture search with 3d medical image segmentation using hypernet. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20741–20751.

Purushwalkam, S., & Gupta, A. (2020). Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems, 33*, 3407–3418.

Qi, C.R., Su, H., Mo, K., & Guibas, L.J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660.

Roh, B., Shin, W., Kim, I., & Kim, S.(2021). Spatially consistent representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1144–1153.

Schmarje, L., Brünger, J., Santarossa, M., Schröder, S.-M., Kiko, R., & Koch, R. (2021). Fuzzy overclustering: Semi-supervised classification of fuzzy labels with overclustering and inverse cross-entropy. *Sensors, 21*(19), 6661.

Seong, H.S., Moon, W., Lee, S., & Heo, J.-P.(2023). Leveraging hidden positives for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19540–19549.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H.(2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357 . PMLR.

Vahdat, A., Kreis, K., & Kautz, J. (2021). Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems, 34*, 11287–11302.

Van Gansbeke, W., Vandenhende, S., & Van Gool, L.(2022). Discovering object masks with transformers for unsupervised semantic segmentation. arXiv preprint arXiv:2206.06363.

Van Gansbeke, W., Vandenhende, S., Georgoulis, S., & Van Gool, L.(2021). Unsupervised semantic segmentation by contrasting object mask proposals. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10052–10062.

Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., & Van Gool, L.(2020). Scan: Learning to classify images without labels. In *European conference on computer vision*, pp. 268–285 . Springer.

Wang, Z., Rao, Y., Yu, X., Zhou, J., & Lu, J.(2022). Semaffinet: Semantic-affine transformation for point cloud segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11819–11829.

Wang, W., Sun, G., & Van Gool, L. (2024). Looking beyond single images for weakly supervised semantic segmentation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(3), 1635–1649.

Wang, Y., Zhang, J., Kan, M., Shan, S., & Chen, X.(2020). Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12275–12284.

Wang, X., Zhang, R., Shen, C., Kong, T., & Li, L. (2021). Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3024–3033.

Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., & Huang, T.S.(2018). Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7268–7277.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18(3), 293–297.

Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., & Kornblith, S., *et al.*(2022). Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998 . PMLR.

Xiong, H., Wu, J., & Chen, J.(2006). K-means clustering versus validation measures: a data distribution perspective. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 779–784.

Xu, H., Caramanis, C., & Mannor, S.(2008). Robust regression and lasso. Advances in Neural Information Processing Systems **21**.

Xu, L., Neufeld, J., Larson, B., & Schuurmans, D.(2004). Maximum margin clustering. Advances in Neural Information Processing Systems **17**.

Yin, Z., Wang, P., Wang, F., Xu, X., Zhang, H., Li, H., & Jin, R.(2022). Transfgu: a top-down approach to fine-grained unsupervised semantic segmentation. In *European conference on computer vision*, pp. 73–89 . Springer.

Zadaianchuk, A., Kleindessner, M., Zhu, Y., Locatello, F., & Brox, T.(2022). Unsupervised semantic segmentation with self-supervised object-centric representations. arXiv preprint arXiv:2207.05027.

Zhan, X., Xie, J., Liu, Z., Ong, Y.-S., & Loy, C.C.(2020). Online deep clustering for unsupervised representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6688–6697.

Zhang, B., Xiao, J., Jiao, J., Wei, Y., & Zhao, Y.(2021). Affinity attention graph neural network for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(11), 8082–8096.

Zhao, B., Wang, F., & Zhang, C.(2008). Efficient multiclass maximum margin clustering. In *Proceedings of the 25th international conference on machine learning*, pp. 1248–1255.

Zhou, Z., Qi, L., Yang, X., Ni, D., & Shi, Y.(2022). Generalizable cross-modality medical image segmentation via style augmentation and dual normalization. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20856–20865.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., & Kong, T.(2021). ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832.

Ziegler, A., & Asano, Y.M.(2022). Self-supervised learning of object parts for semantic segmentation. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14502–14511.