

Self-supervised Semantic Segmentation Grounded in Visual Concepts

Wenbin He¹, William Surmeier¹, Arvind Kumar Shekar², Liang Gou¹, Liu Ren¹

¹Robert Bosch Research and Technology Center North America

²Robert Bosch GmbH

wenbin.he2@us.bosch.com, clayton.surmeier@gmail.com, arvindkumar.shekar@de.bosch.com,
liang.gou@us.bosch.com, liu.ren@us.bosch.com

Abstract

Unsupervised semantic segmentation requires assigning a label to every pixel without any human annotations. Despite recent advances in self-supervised representation learning for individual images, unsupervised semantic segmentation with pixel-level representations is still a challenging task and remains underexplored. In this work, we propose a self-supervised pixel representation learning method for semantic segmentation by using visual concepts (i.e., groups of pixels with semantic meanings, such as parts, objects, and scenes) extracted from images. To guide self-supervised learning, we leverage three types of relationships between pixels and concepts, including the relationships between pixels and local concepts, local and global concepts, as well as the co-occurrence of concepts. We evaluate the learned pixel embeddings and visual concepts on three datasets, including PASCAL VOC 2012, COCO 2017, and DAVIS 2017. Our results show that the proposed method gains consistent and substantial improvements over recent unsupervised semantic segmentation approaches, and also demonstrate that visual concepts can reveal insights into image datasets.

1 Introduction

Semantic segmentation plays a crucial role in a broad range of applications, including autonomous driving, medical image analysis, etc. It partitions an image into semantically meaningful regions and assigns each region with a semantic label such as people, bikes, and cars. Recently, semantic segmentation models [Zhao *et al.*, 2017; Chen *et al.*, 2018a] with deep convolutional neural networks (CNNs) have shown promising results on popular benchmarks. However, these approaches rely heavily on pixel-wise annotations, which cost significant amounts of time and money to acquire. Thus computer vision community starts to pay more attention to unsupervised/self-supervised approaches.

With recent advances in self-supervised learning, visual representations can be learned from images without additional supervision. However, many self-supervised representation learning frameworks (e.g., SimCLR [Chen *et al.*,

2020], MOCO [He *et al.*, 2020]) focus on the visual representation of a whole image and require curated single-object images. Only a few recent approaches learn visual representations at the pixel level for complex scenes, including SegSort [Hwang *et al.*, 2019], Hierarchical Grouping [Zhang and Maire, 2020], and MaskContrast [Van Gansbeke *et al.*, 2021]. These pixel-level representations can be used for semantic segmentation and show promising results. The key idea of these approaches is to use contrastive learning to separate positive pixel pairs from negative pixel pairs. They use different priors to formulate positive and negative pairs, such as contour detection [Hwang *et al.*, 2019], the hierarchical grouping of image segments [Zhang and Maire, 2020], and saliency masks [Van Gansbeke *et al.*, 2021]. However, these methods largely emphasize pixels' local affinity information at individual image level but ignore the global semantics in the whole dataset, and thus tend to separate the representation of objects from different images even when these objects belong to the same semantic class.

In this work, we propose a self-supervised pixel representation learning method by leveraging some important properties of visual concepts at both local and global levels. Visual concepts can be informally defined as human-interpretable abstractions of image segments with semantic meanings (e.g., parts and objects). The idea of visual concept is inspired by recent work in the field of eXplainable AI [Bau *et al.*, 2017; Kim *et al.*, 2018], and offers us a unified framework to derive self-supervision priors at both local and global levels.

Specifically, we use three types of relationships between pixels and visual concepts to guide self-supervised learning. First, aim to leverage the relationship between pixels and local concepts (i.e., visually coherent regions in each image), we use contrastive learning to force the pixels in the same pseudo segments moving close in the embedding space, and push pixels from either different pseudo segments or different images moving far away. Secondly, we group local concepts with similar semantic meanings into global concepts with a vector quantization (VQ) method [van den Oord *et al.*, 2017]. A VQ dictionary learns discrete representations of global concepts by pushing each local concept from an image to its closest global concept vector. Lastly, we utilize the co-occurrence of different global concepts because relevant global concepts tend to appear in the same image, such as human face and body, or rider and bike. These relation-

ships based on visual concepts together regularize the self-supervised learning process.

The learned pixel embeddings can then be used for semantic segmentation by k -means clustering [Hwang *et al.*, 2019] or fine-tuning [Zhang and Maire, 2020; Van Gansbeke *et al.*, 2021]. Moreover, our method can be used to extract visual concepts from a dataset and offer a global semantic understanding about the dataset.

In summary, the contributions of this paper are threefold:

- We propose a self-supervised pixel representation learning method for semantic segmentation, which learns visual concepts for image segments and uses the relationships between pixels and concepts to improve the pixel representation. Moreover, our method only uses pseudo image segments generated by weak priors (e.g., a contour detector or superpixels) without additional information such as hierarchical groups [Zhang and Maire, 2020] or saliency masks [Van Gansbeke *et al.*, 2021].
- Our approach can produce a set of global visual concepts that are semantically meaningful to humans. Because there are a finite number of discrete visual vocabulary, they are easier to explore and interpret comparing with the high dimensional embeddings.
- We demonstrate the accuracy and generalizability of the learned pixel embeddings on three datasets, including PASCAL VOC 2012, COCO 2017, and DAVIS 2017.

2 Related Work

Unsupervised Semantic Segmentation. Unsupervised semantic segmentation is less studied in the field, and most of the existing work focuses on non-parametric methods [Shi and Malik, 2000; Tighe and Lazebnik, 2010]. Recently, several deep learning-based approaches have been proposed for unsupervised semantic segmentation. Ji *et al.* [Ji *et al.*, 2019] proposed a clustering-based approach for image classification and segmentation. A few self-supervised representation learning approaches [Hwang *et al.*, 2019; Zhang and Maire, 2020; Van Gansbeke *et al.*, 2021] were proposed recently, which learn visual representations at the pixel level and use the learned pixel embeddings to segment images. Our method also belongs to self-supervised representation learning, and the difference between our method and the existing approaches will be discussed in the following section.

Self-supervised Learning. Self-supervised learning aims to learn visual representations by predicting one aspect of the input from another, for which various pretext tasks have been used such as predicting image rotations [Gidaris *et al.*, 2018], solving jigsaw puzzles [Noroozi and Favaro, 2016], inpainting [Pathak *et al.*, 2016], etc. Recently, contrastive learning-based methods [Chen *et al.*, 2020; He *et al.*, 2020] have shown great success in learning visual representations in a self-supervised manner. These methods often use a contrastive loss to map different augmented views of the same input to a common embedding location and distinct from other inputs. However, these methods mainly learn image-level representations and require curated single-object images.

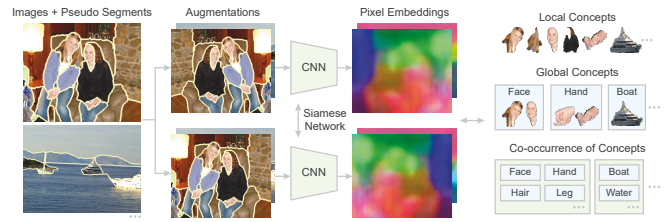


Figure 1: Overview of the proposed method. We learn pixel embeddings for semantic segmentation in a self-supervised setting with pseudo image segments and data augmentations. We define three types of relationships between pixels and visual concepts to guide the self-supervised learning.

A few recent approaches learn pixel-level embeddings for the segmentation of complex scenes. Hwang *et al.* [Hwang *et al.*, 2019] proposed SegSort, which uses pseudo image segments generated by a contour detector to guide the contrastive learning of pixel embeddings. Zhang and Maire [Zhang and Maire, 2020] group the pseudo image segments hierarchically and use the hierarchical structure to guide the sampling of positive and negative pixel pairs. Van Gansbeke *et al.* [Van Gansbeke *et al.*, 2021] use saliency masks to learn pixel embeddings. However, these methods focus on pixels’ local affinity information without considering the global semantics in the whole dataset and thus tend to separate objects of the same semantic class in the embedding space. In this work, we learn visual concepts at both local and global levels only using pseudo image segments. We leverage different types of relationships between pixels and concepts to improve the pixel embeddings for semantic segmentation.

Visual Concepts. Concepts are human interpretable abstractions extracted from images, which are typically represented as image segments with semantic meanings [Bau *et al.*, 2017]. Visual concepts have been widely used in explainable AI to interpret and explain what has been learned by a CNN model. Most of the existing work along this line of research uses human-specified concepts [Bau *et al.*, 2017; Kim *et al.*, 2018] for model interpretation. Recently, several data-driven approaches have been proposed to extract concepts from images using superpixels [Ghorbani *et al.*, 2019], prototypes [Chen *et al.*, 2019], and a dictionary of object parts [Huang and Li, 2020]. Our work bridges the gap between concept extraction and semantic segmentation with pixel embeddings. On one hand, we propose a new approach for visual concept extraction. On the other hand, we leverage the visual concepts to improve self-supervised representation learning for semantic segmentation.

3 Method

In this work, we learn a pixel embedding function, namely a convolutional neural network (CNN), for semantic segmentation with a contrastive self-supervised learning framework. For each pixel p , the embedding function ϕ generates a feature representation \mathbf{z}_p in an embedding space of dimension D , which are then used to derive the semantic segmentation of input images. Like existing contrastive self-supervised methods [Chen *et al.*, 2020; He *et al.*, 2020], we use data aug-

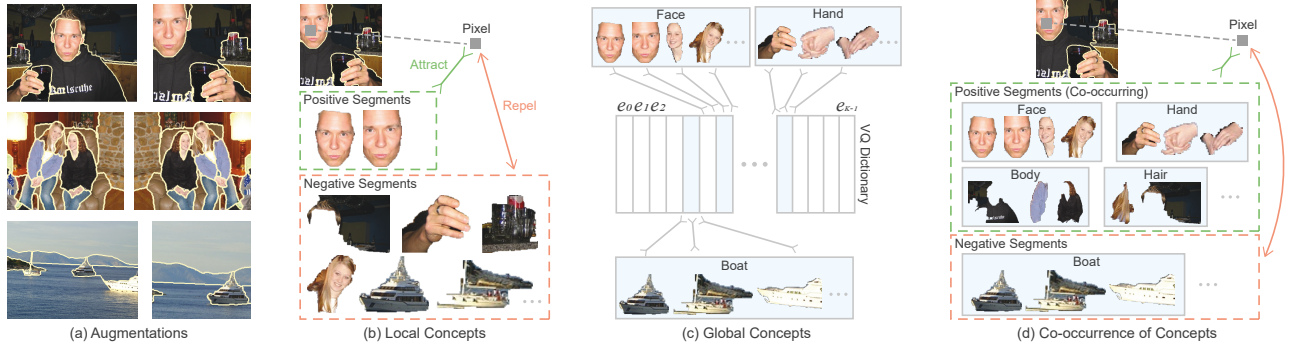


Figure 2: Given the augmented views of a set of images with pseudo segments (a), we train pixel embeddings to capture three types of relationships between pixels and visual concepts. (b) To capture the relationship between pixels and local concepts (i.e., visually coherent regions in each image), we attract the representation of a pixel with the pseudo segments that it belongs to in different augmented views and repel other segments. (c) We group local concepts with similar feature representations into global concepts with VQ. The representations of the global concepts form a VQ dictionary that captures the semantic meanings of segment clusters over the entire dataset. (d) We attract the representations of pixels and segments whose global concepts often co-occur in the same image, such as the human face, body, and hand.

mentation to improve the learning of visual representations. Specifically, we generate two augmented views for each image with multiple data augmentation operations such as random cropping and random color jittering. We then force consistent pixel embeddings between the augmented views.

As pixel-wise labels are not available, our method is based on pseudo image segments of visually coherent regions, which could be derived from super-pixels or contours. Then we train the pixel embedding function by leveraging important properties of visual concepts learned from the pseudo image segments (Figure 1). Specifically, we use three types of relationships between pixels and concepts to guide the self-supervised learning, including the relationships between pixels and local concepts, local and global concepts, as well as the co-occurrence of different concepts as detailed below.

Local Concepts. We first train the pixel embeddings to conform with visually coherent regions, namely local concepts, in each image. The idea is that pixels within visually coherent regions should have similar representations in the embedding space [Hwang *et al.*, 2019; Zhang and Maire, 2020; Ke *et al.*, 2021]. To this end, we define the **local segmentation loss** \mathcal{L}_s to learn pixel representations with a contrastive representation learning approach following [Hwang *et al.*, 2019]. Given the augmented views for a batch of images (Figure 2a), we first derive each pixel’s positive and negative segments, denoted by \mathcal{S}^+ and \mathcal{S}^- respectively, based on the pseudo image segments. For a pixel p , its positive segments include the segments it belongs to in both augmented views, and other segments are negative ones (Figure 2b). Then, the local segmentation loss is defined as the pixel-to-segment contrastive loss:

$$\mathcal{L}_s(p) = -\log \frac{\sum_{s \in \mathcal{S}^+} \exp(\text{sim}(\mathbf{z}_p, \mathbf{z}_s) \kappa)}{\sum_{s \in \mathcal{S}^+ \cup \mathcal{S}^-} \exp(\text{sim}(\mathbf{z}_p, \mathbf{z}_s) \kappa)}, \quad (1)$$

where κ is the concentration constant and $\text{sim}(\mathbf{z}_p, \mathbf{z}_s)$ is the cosine similarity between the feature representation \mathbf{z}_p of a pixel p and the feature representation \mathbf{z}_s of a segment s . The feature representation \mathbf{z}_s is defined as the average representation of the pixels within s , namely $\mathbf{z}_s = \sum_{p \in s} \mathbf{z}_p / |s|$.

Global Concepts. We group the local concepts extracted from individual images into global concepts (i.e., clusters of image segments with similar semantic meanings) for the entire dataset. We introduce the global concepts based on the following observations. Image segments with similar visual appearance may locate at different regions of the same image or even different images, such as the human faces in Figure 2a. Since they belong to different local concepts, those segments are considered as negative examples with each other. Hence, their representations will be pushed away from each other if only considering local concepts, which will eventually hurt the performance of semantic segmentation. Moreover, as focusing on individual images, local concepts do not capture visual concepts across the entire dataset.

We use VQ [van den Oord *et al.*, 2017] to learn global visual concepts that extract clusters of image segments with similar semantic meanings from the dataset (Figure 2c). For the segments that belong to the same global concept, the **VQ loss** \mathcal{L}_v makes their representations close to each other. Specifically, we train a VQ dictionary that contains a set of discrete representations for the concept centers, denoted as $\mathbf{e}_0, \mathbf{e}_1, \dots, \mathbf{e}_{K-1}$, where K is the number of concepts and \mathbf{e}_i is the representation of i -th concept. For each training iteration, we first assign each segment s to the global concept k with the nearest representation:

$$k = \text{argmax}_i \text{sim}(\mathbf{z}_s, \mathbf{e}_i), \quad (2)$$

where $\text{sim}(\mathbf{z}_s, \mathbf{e}_i)$ is the cosine similarity between representations. We use the cosine similarity instead of the Euclidean distance used in [van den Oord *et al.*, 2017] is because we learn representations on a hypersphere. Then, we maximize the cosine similarity between the representations of the segments and the corresponding concepts:

$$\mathcal{L}_v = (1 - \text{sim}(\text{sg}(\mathbf{z}_s), \mathbf{e}_k)) + \beta(1 - \text{sim}(\mathbf{z}_s, \text{sg}(\mathbf{e}_k))), \quad (3)$$

where the first half of this function is used to update the VQ dictionary (i.e., cluster centers) as the stop gradient operator sg is applied on the representation of the segments. Similarly, the second half is used to update the segments’ representation while fixing the VQ dictionary. In addition, we use

the commitment constant β to control the learning rate of the segments' representation.

The VQ dictionary captures discrete representations with semantic meanings such as human faces and boats for the entire dataset. It can be used to learn the relationships between different global concepts such as boats and water, which can be exploited to further improve the pixel representations as detailed in the following section.

Co-occurrence of Concepts. We utilize the co-occurrence of different global concepts to further improve the pixel embeddings. The motivation is that global concepts with relevant semantic meanings tend to co-occur in the same image, such as the human face, hair, and body in Figure 2a. However, the representation of relevant concepts (e.g., different human body parts) will be pushed away as they belong to different global concepts if no co-occurrence constraints.

Inspired by [Ke *et al.*, 2021], we introduce the **co-occurrence loss** \mathcal{L}_c to attract the representations of pixels and segments whose global concepts often co-occur in the same image. Different from [Ke *et al.*, 2021], which uses image tags to obtain the co-occurrence, we exploit the VQ dictionary without additional supervision. Given a pixel p , we redefine its positive and negative segments based on the co-occurrence of the global concepts derived from the VQ dictionary. Specifically, we first determine which global concept the pixel belongs to by looking up the segment containing that pixel in the VQ dictionary. Then, the pixels positive segments \mathcal{C}^+ are defined as the segments that co-occur with the pixel's concept in the same image, and other segments are defined as negative ones \mathcal{C}^- . For example, in Figure 2d, because the pixel belongs to the concept of human face, all image segments that co-occur with human faces are its positive segments, such as different body parts of a person. The pixel's embedding is then trained based on the contrastive loss similar to the local segmentation loss:

$$\mathcal{L}_c(p) = -\log \frac{\sum_{s \in \mathcal{C}^+} \exp(\text{sim}(\mathbf{z}_p, \mathbf{z}_s) \kappa)}{\sum_{s \in \mathcal{C}^+ \cup \mathcal{C}^-} \exp(\text{sim}(\mathbf{z}_p, \mathbf{z}_s) \kappa)}, \quad (4)$$

where the only difference is the definition of the positive and negative segments.

In the end, we use the three types of relationship between pixels and visual concepts to regularize the self-supervised learning process. The total loss to train the representation of a pixel p is the weighted combination of the aforementioned three loss terms:

$$\mathcal{L}(p) = \lambda_s \mathcal{L}_s(p) + \lambda_v \mathcal{L}_v(p) + \lambda_c \mathcal{L}_c(p), \quad (5)$$

where λ_s , λ_v , and λ_c are the weights for each loss term.

4 Experiments

4.1 Experimental Setup

Datasets. We mainly experiment on the Pascal VOC 2012 dataset, which contains 20 object classes and one background class. Following the prior work [Hwang *et al.*, 2019], we train networks on the *train_aug* set with 10,582 images and test on the *val* set with 1,449 images. For self-supervised pre-training, we use pseudo segmentations generated by HED-owt-ucm [Hwang *et al.*, 2019], unless otherwise stated.

We also perform experiments on COCO 2017 and DAVIS 2017 to evaluate the generalizability of the learned pixel embeddings. Note that the pixel embeddings are trained only on the *train_aug* set of Pascal VOC 2012 without including any images from COCO 2017 or DAVIS 2017.

Training. For all the experiments, we use PSPNet [Zhao *et al.*, 2017] with a dilated ResNet-50 backbone as the network architecture. The backbone is pre-trained on the ImageNet dataset. For self-supervised pre-training, the hyper-parameters are set as follows. The embedding dimension is set to 32, and the concentration constant κ is set to 10. For VQ, we use a dictionary of size 512 and set the commitment constant β to 0.5. The weights λ_s , λ_v , and λ_o of each loss term are set to 1, 2, and 1, respectively. We train the network on the *train_aug* set of Pascal VOC 2012 for 5k iterations with a batch size of 8. We set the initial learning rate to 0.001 and decay it with a poly learning rate policy. We use additional memory banks to cache the embeddings of the previous 2 batches. We use the same set of data augmentations as SimCLR [Chen *et al.*, 2020] during training, including random resizing, cropping, flipping, color jittering, and Gaussian blurring. Note that in the experiments, the same pseudo segment within different augmented views are merged and considered as one segment for loss computation to save time.

4.2 Results and Analysis

Pascal VOC 2012 and COCO 2017: Benchmarking Results. We evaluate the learned pixel embeddings with two approaches, including k -means clustering [Hwang *et al.*, 2019] and linear classification [Zhang and Maire, 2020]. For k -means clustering, we follow the procedure of SegSort [Hwang *et al.*, 2019]. We first segment each image by clustering the pixels based on the embeddings. We then assign each segment a semantic class label by the majority vote of its nearest neighbors from the training set. The hyper-parameters are defined as follows. Each image is clustered into 25 segments through 50 iterations, and 15 nearest neighbors are used for predicting class labels during inference. For linear classification, we train an additional softmax classifier, namely a 1×1 convolutional layer, while fixing the learned pixel embeddings. We train the classifier for 60k iterations with a batch size of 16. The learning rate starts at 0.1 and decays by 0.1 at 20k and 50k iterations.

We compare our method with the three methods that learn pixel embeddings for semantic segmentation include SegSort [Hwang *et al.*, 2019], MaskContrast [Van Gansbeke *et al.*, 2021], and Hierarchical Grouping [Zhang and Maire, 2020]. Both SegSort and MaskContrast are trained on a backbone network pre-trained on the ImageNet dataset, which is the same as our approach. For comparison, we train a SegSort model using the same hyper-parameter setting as our approach. For MaskContrast, we take the best model provided by the authors and evaluate the performance of the model the same as our approach. Hierarchical Grouping is a slightly different approach, whose goal is to learn pixel embeddings from scratch for general purposes. For semantic segmentation, Hierarchical Grouping requires training a complete atrous spatial pyramid pooling (ASPP) [Chen *et al.*,

Method	Pseudo Seg.	mIoU (k -means)	mIoU (LC)
SegSort	HED	49.17	58.85
Hierarch. Group.	SE	–	46.51
Hierarch. Group.	HED	–	48.82
MaskContrast	Unsup. Sal.	49.67	62.22†
MaskContrast	Sup. Sal.	52.35	65.08†
Ours <i>w/o</i> \mathcal{L}_v & \mathcal{L}_c	HED	61.02	62.31
Ours <i>w/o</i> \mathcal{L}_c	HED	61.77	62.72
Ours	Felz-Hutt	60.38	61.90
Ours	HED	62.56	63.47

Table 1: **Benchmarking results on Pascal VOC 2012.** Our method outperforms previous work by a large margin when using k -means. Meanwhile, our approach performs the second best for linear classification. The method that outperforms our approach is MaskContrast with a supervised saliency estimator trained on annotated data. † The number is slightly higher than the original paper because we evaluate the models on images at multiple scales.

Method	Pseudo Seg.	mIoU (k -means)	mIoU (LC)
SegSort	HED	31.02	36.07
MaskContrast	Unsup. Sal.	27.47	39.77
MaskContrast	Sup. Sal.	29.82	43.19
Ours <i>w/o</i> \mathcal{L}_v & \mathcal{L}_c	HED	36.45	38.92
Ours <i>w/o</i> \mathcal{L}_c	HED	37.16	40.22
Ours	HED	38.00	40.84

Table 2: **Benchmarking results on COCO 2017.** Our method outperforms previous work on k -means and performs the second best on linear classification. The method that outperforms our approach is MaskContrast that trained on annotated data.

2018b] module based on the learned embeddings and the annotated data. In contrast, our method as well as SegSort and MaskContrast only train a linear classifier on the annotated data. For the sake of completeness, we take the resulting numbers from [Zhang and Maire, 2020] and compare with other methods. Note that comparing with methods that produce image-level representations [Chen *et al.*, 2020; He *et al.*, 2020] is out the scope of this work. The interested reader may consult the previous work [Zhang and Maire, 2020; Van Gansbeke *et al.*, 2021] for the comparison results.

Table 1 shows the benchmarking results of the aforementioned methods on Pascal VOC 2012. We find that our method outperforms previous methods by a large margin on k -means. The performance of our method on linear classification outperforms MaskContrast with an unsupervised saliency estimator (63.47 vs. 62.22). Meanwhile, MaskContrast with a supervised saliency estimator has a better performance comparing with ours (65.08 vs. 63.47). However, it requires training a saliency estimator with a large number of annotated images [Qin *et al.*, 2019]. Figure 3 compares the learned pixel embeddings and the semantic segmentation generated by k -means for our approach and MaskContrast. We find that our method can produce pixel embeddings with clear boundaries between different objects. While MaskContrast focuses only on the salient objects and the pixel embeddings are less informative in other regions. We also examine the effects of the three losses by introducing them one by one

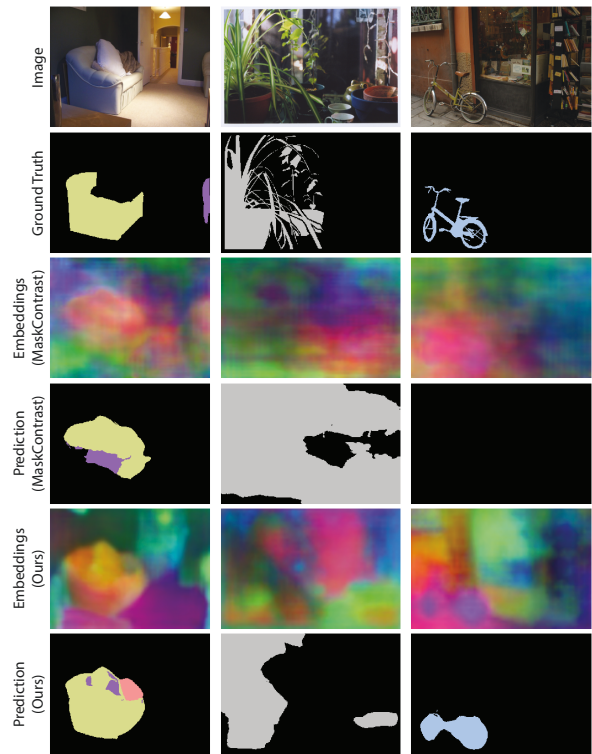


Figure 3: **Visual comparison on PASCAL VOC 2012 validation set.** Comparing with MaskContrast, our method can produce less noisy pixel embeddings with clear boundaries between different objects, hence generates better semantic segmentation using k -means clustering. Note that the pixel embeddings are visualized by projecting to a 3-dimensional space using PCA.

(shown in Table 1). With global concepts, the model is improved by 0.75% on k -means. With concepts co-occurrence, the model is further improved by 0.8%. Similar results can also be found in linear classification. In addition, our method can achieve comparable performance to previous work even with pseudo segments generated by a non-parametric contour detector [Felzenszwalb and Huttenlocher, 2004]. **Detailed ablation study results are included in the supplementary.**

Table 2 shows the benchmarking results on COCO 2017. Again, our method outperforms previous work on k -means and performs the second best on linear classification.

Pascal VOC 2012: Analysis of Visual Concepts. We further analyze the learned visual concepts, whose quality plays an important role in the performance of our approach. To this end, we segment each image based on the discrete representation of the global concepts, namely the VQ dictionary. Given an input image, we first generate the embedding of each pixel. Then we cluster the pixels into segments based on the pixel embeddings. In the end, we map each segment into one of the concepts based on their distance to the concepts. We also merge the segments with the same concept if they are connected. Based on the extracted image segments, the visual concepts are evaluated from different aspects.

To evaluate the visual concepts, we first answer the question that if each visual concept captures image features with

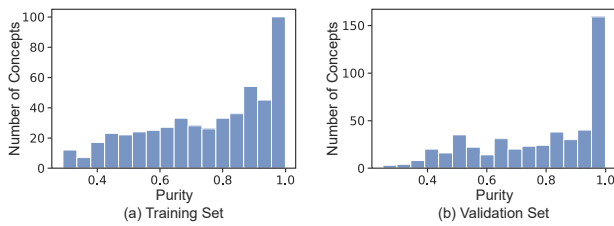


Figure 4: **Distribution of the purity for the visual concepts on the (a) training and (b) validation set.** More than half of the concepts have the purity higher than 80% on both sets.



Figure 5: **Image segments randomly sampled from a few concepts (one concept per column).** Concepts of high purity capture image segments with clear semantic meanings. Moreover, fine-grained visual concepts can be extracted from the dataset using our method, such as human face, cloth with stripes, and car window. Concepts of low purity can also capture image features with similar meanings such as farm animals and round structures in vehicles.

the same semantic meaning. To this end, we define the *purity* of each concept as follows. For each concept, we collect the corresponding image segments and assign each image segment with a class label based on the ground truth. Then we calculate the percentage of the image segments that belong to the majority class of the current concept. The distribution of the purity for the visual concepts is shown in Figure 4. We find that more than half of the concepts have the purity higher than 80% on both the training and validation set.

Figure 5 visualizes image segments randomly sampled from a few concepts (one concept per column). The first 6 columns are from concepts with purity higher than 80%, and the last two columns are from concepts with purity lower than 40%. We find that the visual concepts of high purity contain image segments with clear semantic meanings such as airplane, cat, and bird. Moreover, our method can extract fine-grained visual concepts such as human face, cloth with stripes, and car window. For the concepts with low purity, the image segments still capture similar features such as farm animals and round structures in vehicles. However, without annotations, the network has difficulty differentiating horses and cows or determining the type of vehicles.

Method	$\mathcal{J}(\text{Mean})\uparrow$	$\mathcal{F}(\text{Mean})\uparrow$
MaskTrack (fine-tuned) [†]	51.2	57.3
OSVOS (fine-tuned) [†]	55.1	62.1
MaskTrack-B	35.3	36.4
OSVOS-B	18.5	30.0
Video Colorization	34.6	32.7
CycleTime	41.9	39.4
mgPFF	42.2	46.9
Hierarch. Group.	47.1	48.9
MaskContrast (Sup.)	34.3	36.7
Ours (Felz-Hutt)	46.3	49.0
Ours	50.4	53.9

Table 3: **Performance of instance mask tracking on DAVIS 2017 validation set.** The performance is measured by the region similarity \mathcal{J} and the contour-based accuracy \mathcal{F} . Our method outperforms recent supervised and unsupervised methods on both metrics. [†] indicates the models are fine-tuned on the first frame of the test videos.

DAVIS 2017: Instance Mask Tracking. We evaluate the generalizability of our method on instance mask tracking over the DAVIS 2017 validation set, where the instance masks at the first frame are given for each video. We propagate the instance masks to the rest of the frames based on the similarity between pixel embeddings following the method proposed in [Zhao *et al.*, 2017]. The performance is measured by the region similarity \mathcal{J} (IoU) and the contour-based accuracy \mathcal{F} .

We compare our method with recent supervised and unsupervised methods in Table 3. The supervised methods MaskTrack-B [Perazzi *et al.*, 2017] and OSVOS-B [Caelles *et al.*, 2017] train models with ImageNet pre-training and annotated masks. MaskTrack [Perazzi *et al.*, 2017] and OSVOS [Caelles *et al.*, 2017] further fine-tune the models on the first frame of the test video. Comparing to the supervised methods, our method outperforms MaskTrack-B and OSVOS-B by a large margin without using any annotated masks. Also, our method achieves more than 91% and 87% performance of the fine-tuned models in terms of the region similarity \mathcal{J} and contour accuracy \mathcal{F} , respectively. Our method also outperforms recent video-based [Vondrick *et al.*, 2018; Wang *et al.*, 2019; Kong and Fowlkes, 2019] and image-based [Zhang and Maire, 2020; Van Gansbeke *et al.*, 2021] unsupervised approaches by more than 3% and 5% in \mathcal{J} and \mathcal{F} , respectively. Moreover, our method can achieve comparable performance to previous work even with pseudo segments generated by a non-parametric contour detector [Felzenszwalb and Huttenlocher, 2004]. Due to limited space, visual results are included in the supplementary.

5 Conclusion

We propose a novel unsupervised semantic segmentation method based on self-supervised representation learning at the pixel level. Our method uses three types of relationships between pixels and visual concepts to regularize the self-supervised representation learning and hence improve the learned pixel embeddings. We demonstrate the accuracy and generalizability of the learned pixel embeddings on PASCAL VOC2012, COCO 2017, and DAVIS 2017.

References

- [Bau *et al.*, 2017] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, pages 3319–3327, 2017.
- [Caelles *et al.*, 2017] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR*, pages 5320–5329, 2017.
- [Chen *et al.*, 2018a] 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. *PAMI*, 40(4):834–848, 2018.
- [Chen *et al.*, 2018b] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018.
- [Chen *et al.*, 2019] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K. Su. This looks like that: Deep learning for interpretable image recognition. In *NeurIPS*, pages 8930–8941, 2019.
- [Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.
- [Felzenszwalb and Huttenlocher, 2004] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.
- [Ghorbani *et al.*, 2019] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations. In *NeurIPS*, pages 9277–9286, 2019.
- [Gidaris *et al.*, 2018] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv:1803.07728*, 2018.
- [He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735, 2020.
- [Huang and Li, 2020] Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *CVPR*, pages 8659–8669, 2020.
- [Hwang *et al.*, 2019] 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*, pages 7333–7343, 2019.
- [Ji *et al.*, 2019] Xu Ji, Andrea Vedaldi, and Joao Henriques. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, pages 9864–9873, 2019.
- [Ke *et al.*, 2021] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X. Yu. Universal weakly supervised segmentation by pixel-to-segment contrastive learning. In *ICLR*, 2021.
- [Kim *et al.*, 2018] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, pages 4186–4195, 2018.
- [Kong and Fowlkes, 2019] Shu Kong and Charless Fowlkes. Multigrid predictive filter flow for unsupervised learning on videos. *arXiv:1904.01693*, 2019.
- [Noroozi and Favaro, 2016] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84, 2016.
- [Pathak *et al.*, 2016] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016.
- [Perazzi *et al.*, 2017] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, pages 3491–3500, 2017.
- [Qin *et al.*, 2019] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7471–7481, 2019.
- [Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000.
- [Tighe and Lazebnik, 2010] Joseph Tighe and Svetlana Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV*, pages 352–365, 2010.
- [van den Oord *et al.*, 2017] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, pages 6309–6318, 2017.
- [Van Gansbeke *et al.*, 2021] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. *arXiv:2102.06191*, 2021.
- [Vondrick *et al.*, 2018] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *ECCV*, pages 402–419, 2018.
- [Wang *et al.*, 2019] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, pages 2561–2571, 2019.
- [Zhang and Maire, 2020] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. In *NeurIPS*, pages 16579–16590, 2020.
- [Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 6230–6239, 2017.