Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals

Wouter Van Gansbeke[1*] Simon Vandenhende[1*] Stamatios Georgoulis[2] Luc Van Gool[1,2]
[1]KU Leuven/ESAT-PSI [2]ETH Zurich/CVL, TRACE

Abstract

Being able to learn dense semantic representations of images without supervision is an important problem in computer vision. However, despite its significance, this problem remains rather unexplored, with a few exceptions that considered unsupervised semantic segmentation on small-scale datasets with a narrow visual domain. In this paper, we make a first attempt to tackle the problem on datasets that have been traditionally utilized for the supervised case. To achieve this, we introduce a two-step framework that adopts a predetermined mid-level prior in a contrastive optimization objective to learn pixel embeddings. This marks a large deviation from existing works that relied on proxy tasks or end-to-end clustering. Additionally, we argue about the importance of having a prior that contains information about objects, or their parts, and discuss several possibilities to obtain such a prior in an unsupervised manner.

Experimental evaluation shows that our method comes with key advantages over existing works. First, the learned pixel embeddings can be directly clustered in semantic groups using K-Means on PASCAL. Under the fully unsupervised setting, there is no precedent in solving the semantic segmentation task on such a challenging benchmark. Second, our representations can improve over strong baselines when transferred to new datasets, e.g. COCO and DAVIS. The code is available[1].

1. Introduction

The problem of assigning dense semantic labels to images, formally known as *semantic segmentation*, is of great importance in computer vision as it finds many applications, including autonomous driving, augmented reality, human-computer interaction, etc. To achieve state-of-the-art performance in this task, fully convolutional networks [45] are typically trained on datasets [15, 20, 44] that contain a large number of fully-annotated images. However, obtaining accurate, pixel-wise semantic labels for every image in a dataset is a labor-intensive process that costs significant amounts of time and money [4]. To remedy this situation, weakly-supervised methods leveraged weaker forms of supervision, such as scribbles [43, 65, 66, 72, 80], bounding boxes [16, 37, 56, 80], clicks [5], and image-level tags [56, 66, 80], while semi-supervised methods [16, 26, 28, 56, 57] used only a fraction of the dataset as labeled examples, both of which require substantially less human annotation effort. Despite the continued progress, the vast majority of semantic segmentation works still rely on some form of annotations to train the neural network models.

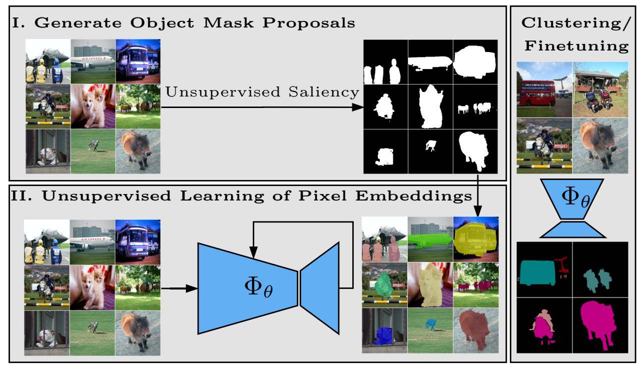

Figure 1. We learn pixel embeddings for semantic segmentation in an unsupervised way. First, we predict object mask proposals using unsupervised saliency. Second, we use the obtained masks as a prior in a self-supervised optimization objective. Finally, the pixel embeddings can be clustered or fine-tuned to a semantic segmentation of the image.

In this paper, we look at the problem from a different perspective, namely self-supervised representation learning. More concretely, we aim to learn pixel-level representations or embeddings for semantic segmentation without using ground-truth. If we obtain a good pixel embedding that is discriminative w.r.t. the semantic classes, we can directly cluster the pixels into semantic groups using K-Means. This tackles the semantic segmentation problem under the fully unsupervised setup. Alternatively, if a limited number of annotated examples are available, the representations can be further fine-tuned under a semi-supervised or transfer learning setup. In this paper, we primarily focus on the fully unsupervised setup, but include additional fine-

[*]Authors contributed equally
[1]github.com/wvangansbeke/Unsupervised-Semantic-Segmentation.git

tuning experiments for the sake of completeness.

Unsupervised or self-supervised techniques [36] were recently being employed to learn rich and effective visual representations without external supervision. The obtained representations can subsequently be used for a variety of purposes, including task transfer learning [24], image clustering [2, 3, 71], semi-supervised classification [12], etc. Popular representation learning techniques used an instance discrimination task [78], that is treating every image as a separate class, to generate representations in an unsupervised way. Images and their augmentations are considered as positive examples of the class, while all other images are treated as negatives. In practical terms, the instance discrimination task is formulated as a non-parametric classification problem, and a contrastive loss [23, 54] is used to model the distribution of negative instance classes.

Purushwalkam and Gupta [61] showed that contrastive self-supervised methods learn to encode semantic information, since two views of the same image will always show a part of the same object, and no objects from other categories. However, under this setup, there is no guarantee that the representations also learn to differentiate between pixels belonging to different semantic classes. For example, when foreground-background pairs frequently co-occur, e.g. cattle grazing on farmland, pixels belonging to the two classes can share their representation. This renders existing works based on instance discrimination less appropriate w.r.t. our goal of learning semantic pixel embeddings. To address these limitations, we propose to learn pixel-level, rather than image-level representations, in a self-supervised way.

The proposed method consists of two steps. First, we leverage an unsupervised saliency estimator to mine object mask proposals from the dataset. This mid-level visual prior transfers well across different datasets. In the second step, we use a contrastive framework to learn pixel embeddings. The object mask proposals are employed as a prior - we pull embeddings from pixels belonging to the same object together, and contrast them against pixels from other objects. The generated representations are evaluated on the semantic segmentation task following standard protocols. The framework is illustrated in Figure 1.

Our contributions are: (1) We propose a two-step framework for unsupervised semantic segmentation, which marks a large deviation from recent works that relied on proxy tasks or end-to-end clustering. Additionally, we argue about the importance of having a mid-level visual prior which incorporates object-level information. This contrasts with earlier works that grouped pixels together based upon low-level vision tasks like boundary detection. (2) The proposed method is the first able to tackle the semantic segmentation task on a challenging dataset like PASCAL under the fully unsupervised setting. (3) Finally, we report promising results when transferring our representations to other datasets. This shows that adopting a mid-level visual prior can be useful for self-supervised representation learning.

2. Related Work

As our method is mostly related to unsupervised semantic segmentation and representation learning, in what follows we discuss representative works from each topic.

Unsupervised semantic segmentation. There have only been a few attempts in the literature to tackle semantic image segmentation under the fully unsupervised setting. Some works [34, 55] followed an end-to-end approach - maximizing the discrete mutual information between augmented views to learn a clustering function. However, these methods were only applied to small-scale datasets, covering a narrow visual domain, e.g. separating sky from vegetation, using satellite imagery, etc. In contrast, our method applies to more challenging scenarios, and decouples feature learning from clustering.

A few works [29, 95] used segments obtained from boundaries to learn pixel embeddings in a self-supervised way. However, it is unclear whether the representations could be post-processed with an off-line clustering criterion to obtain discrete labels. In particular, the evaluation only considered semantic segment retrieval which requires an annotated train set. Furthermore, Hwang *et al*. [29] still relied on additional supervision sources like ImageNet pretraining and boundary annotations [1, 79].

Representation learning. These methods aim at learning visual representations by solving pre-designed *pretext tasks*, which do not require manual annotations. Examples of such pretext tasks include colorizing images [30, 40, 94], predicting context [17, 49], solving jigsaw puzzles [51, 53], generating images [63], clustering [2, 8, 82], predicting noise [6], spotting artifacts [33], using adversarial training [18, 19], predicting optical flow [47, 88], counting [52], inpainting [58], predicting transformation parameters [21, 92], using predictive coding [54], performing instance discrimination [9, 11, 22, 24, 41, 48, 68, 69, 78, 85], and so on. The learned representations can subsequently be transferred to learn a separate down-stream task, e.g. object detection.

In a similar vein, some works tried to learn pixel-level representations for semantic segmentation by solving proxy tasks, e.g. colorization [30, 40, 87, 94], optical flow [47, 88], using co-occurences [31], etc. Differently, in this paper, we avoid the use of a proxy task.

3. Method

In this paper, we aim to learn a pixel embedding function for semantic segmentation from an unlabeled dataset of images. Since the goal of semantic segmentation is to assign a

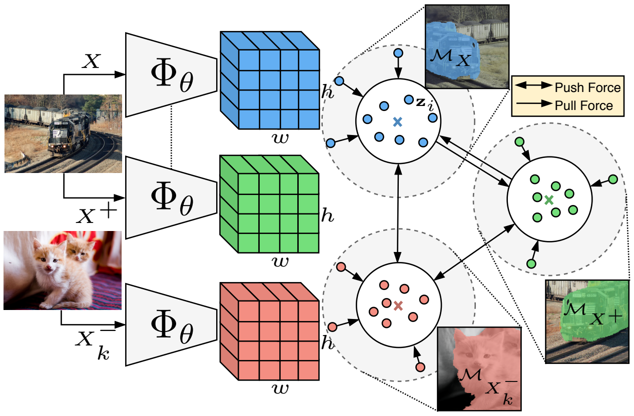

Figure 2. **MaskContrast** learns pixel embeddings for unsupervised semantic segmentation in the following way. We use a saliency estimator to generate positive pairs of object-centric crops (X, X^+) and negative pairs X_k^-. The model Φ_θ is trained to maximize the agreement between embeddings of pixels belonging to the objects in X, X^+, while minimizing the agreement with pixels from objects in X_k^-.

class label to every pixel of an image, a good pixel embedding should be discriminative w.r.t. the semantic classes. If the latter holds true, the embedding function can be directly used to cluster the pixels into semantic groups, or be further fine-tuned under a semi-supervised setup.

To tackle the aforementioned problem, we follow a divide-and-conquer strategy. We argue that it is more difficult to directly cluster the pixels into semantic groups following an end-to-end pipeline, while it is easier to first look for image regions where pixels are likely to belong together. Although this information does not directly result in a semantic segmentation of the scene, it gives us a useful starting point to learn the pixel embeddings. In particular, we can leverage the obtained regions as a prior by grouping their pixels together. Since the prior is determined before the feature learning step, we reduce the dependence on the network initialization. This is an intentional divergence from existing end-to-end learning pipelines [34, 55], which are prone to latch onto low-level image cues - like color, contrast, etc. - as shown in [71].

The proposed method named *MaskContrast* consists of two steps. In a first step, we determine a prior by identifying objects in the images for which pixels can be grouped together. Mid-level visual groups, like objects, transfer well across datasets, since they do not depend on any pre-defined ground-truth classes. In the second step, we employ the obtained prior in a contrastive loss [23, 54] to generate pixel embeddings. More specifically, we pull pixels belonging to the same object together, and contrast them against pixels from other objects, as shown in Figure 2. This forces the model to map pixels from visually similar objects closer together, while pushing pixels from dissimilar objects further apart. In this way, the model discovers a pixel embedding space that can serve as a dense semantic representation of the scene.

The method section is further organized as follows. Section 3.1 motivates the use of object mask proposals as a prior for semantic segmentation. Section 3.2 analyzes the use of an unsupervised saliency estimator to mine the object masks from unlabeled datasets. Section 3.3 integrates the prior in a contrastive loss to learn pixel embeddings.

3.1. A Mid-Level Visual Prior for Grouping Pixels

As a starting point for unsupervised semantic segmentation, we try to define an appropriate prior. Several works have emerged in the literature that tried to group pixels by solving a proxy task. Examples include colorizing images [30, 40, 94], predicting optical flow [47, 88], using co-occurences [31], etc. Unfortunately, there is no guarantee that the generated representations will align with the semantic classes, as the latter are co-variant to the proxy task's output. For example, a colorization network will be sensitive to color changes, even though these do not necessarily alter the semantics of the scene. This behavior is unwanted for the objective of semantic segmentation.

To overcome these limitations, we follow an alternative route that avoids the use of a proxy task. In particular, we mine object mask proposals which cover patches that are likely to contain an object. A prior can then be defined from the masks based upon *shared pixel ownership*, i.e. if a pair of pixels belongs to the same mask, we assume that they should be grouped together, and maximize the agreement between their pixel embeddings. We hypothesize that this is a more reliable pixel grouping strategy compared to the use of proxy tasks. In particular, our approach builds a high-level image segmentation by first identifying mid-level visual groups, instead of directly producing a complete segmentation by solving a proxy task. A motivation for this bottom-up approach is also provided in [64].

At the same time, the proposed prior can be seen as an object-centric approach to unsupervised semantic segmentation, which brings several advantages to the table. First, using mid-level visual cues, like object information, regularizes the feature representations. In particular, the model can not simply rely on low-level information like color to group the pixels together, but needs to learn more semantically meaningful image characteristics. This differs from competing works [29, 95] that used superpixels or image boundaries as a prior. Second, object cues can be highly informative of the semantic segmentation task. Evidence for the latter has been provided in the literature for weakly-supervised methods that utilized annotations containing object information. As an example, several works [16, 37, 56, 80] reported strong results on the seg-

Figure 3. **Mask Proposals.** We train a supervised (middle) and unsupervised (bottom) saliency estimator on the DUTS and MSRA datasets respectively. We make predictions on PASCAL.

mentation task by employing object bounding boxes.

Next, we show how an unsupervised saliency estimator can be used to generate the object mask proposals.

3.2. Mining Object Mask Proposals

We need to retrieve a set of object mask proposals for the images in our dataset. The literature [1, 50, 60, 70] offers a multitude of ways to do this. We prefer to use a simple strategy to verify whether unsupervised semantic segmentation benefits from adopting a mid-level visual prior. Moreover, we would like to use a method that does not rely on external supervision, or can be trained with a limited amount of annotations. In the latter case, the object mask proposal mechanism should generalize well to new scenes.

Based upon our requirements, we propose the use of saliency estimation [7, 77] to generate object masks proposals. Most importantly, various unsupervised methods can be used for this purpose. Several of these works [50, 89, 91] used predictions obtained with hand-crafted priors [35, 42, 96, 98] as pseudo-labels to train a deep neural network. Others [83, 84] relied on videos to learn a salient object detector. Furthermore, on a variety of datasets [14, 73, 81] unsupervised saliency methods have shown to perform on par with their supervised counterparts [27, 46, 62, 75, 90, 93]. Finally, the model predictions transfer well to novel unseen datasets as shown by [50].

For completeness, in Section 4 we explore both unsupervised [50] and supervised [62] saliency estimation methods to predict the object masks, and showcase the potential of our method. Figure 3 shows some examples.

3.3. MaskContrast: Learning Pixel Embeddings by Contrasting Salient Objects

Consider a dataset of images $\mathcal{X}$ with associated non-overlapping object mask proposals $\{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_N\}$ obtained using a saliency estimator. Our goal is to learn a *pixel embedding function* $\Phi_\theta : \mathcal{X} \to \mathcal{Z}$ parameterized by a neural network with weights θ, that maps each pixel i in an image to a point $\mathbf{z}_i$ on a D-dimensional normalized hyper-sphere. We chose a normalized embedding space, so that the output of Φ_θ is bounded. Note that, the use of such scale-invariant embeddings decouples the loss from other design choices that could implicitly limit the range of distances, e.g. weight decay, as shown in [39].

We construct an optimization objective to learn the embedding function Φ_θ as follows. First, we describe how to learn semantically meaningful image feature using a contrastive learning objective. Second, we modify the criterion to learn pixel embeddings.

Learning Image-Level Representations. Existing contrastive self-supervised methods (e.g. [11, 24, 78]) learn visual representations through an instance discrimination task defined at the image-level. Positive views (X, X^+) of the same image are acquired for which it is guaranteed that both images contain a part of the same object. Similarly, examples of negative pairs $\left\{(X, X_0^-), (X, X_1^-), \ldots, (X, X_K^-)\right\}$ can be found that never contain the same object. In practice, we impose additional invariances by applying augmentations. The positives and negatives can now be used in a contrastive framework to learn image representations that encode semantic information about the objects.

We realize this concept by training an *image embedding function* Ψ_η to maximize the agreement between positive pairs (X, X^+), while minimizing the agreement between negative pairs $\left\{(X, X_0^-), (X, X_1^-), \ldots, (X, X_K^-)\right\}$. If we measure the similarity between pairs using a dot product, the contrastive loss [23, 54] is defined as

$$\mathcal{L} = -\log \frac{\exp(\Psi_\eta(X)^T \cdot \Psi_\eta(X^+)/\tau)}{\sum_{k=0}^{K} \exp(\Psi_\eta(X)^T \cdot \Psi_\eta(X_k^-)/\tau)}, \quad (1)$$

where the temperature τ relaxes the dot product. As shown by [61], the model learns to encode object information because the positive examples always preserve a part of the same object. Moreover, since the representational capacity of the network is intentionally limited, visually similar objects will tend to be mapped closer together by Ψ_η. The combination of these two properties results in image representations that can be directly clustered into semantic groups (see also [71] for a more detailed explanation).

The above observations showed how to train a model that encodes semantic object information. Next, we modify the contrastive loss from Equation 1 to learn representations at the pixel level.

Learning Pixel-Level Representations. We adopt the following notation. Let i be a pixel with $\mathbf{z_i}$ its pixel embedding. Let $m(i)$ be the index of the object mask that pixel i belongs to, i.e. $i \in \mathcal{M}_{m(i)}$. Finally, let the mean pixel embedding $\mathbf{z}_{\mathcal{M}_n}$ of an object mask $\mathcal{M}_n$ be defined as

$$\mathbf{z}_{\mathcal{M}_n} = \frac{1}{|\mathcal{M}_n|} \sum_{i \in \mathcal{M}_n} \mathbf{z}_i. \tag{2}$$

The optimization objective is derived from a pull- and push-force in the pixel embedding space.

Pull-force. In Section 3.1, we motivated the use of a prior based upon *shared pixel ownership* to pull pixels together in the embedding space. More concretely, if two pixels i, j belong to the same object, i.e. $m(i) = m(j)$, we maximize the agreement between their pixel embeddings $\mathbf{z}_i, \mathbf{z}_j$. In practice, the agreement is maximized between pixels and the mean embedding of their object mask in order to obtain a criterion that scales linearly with the number of pixels, rather than quadratically.

Push-force. Additionally, we require a push-force to avoid mode collapse in the embedding space. Moreover, the push-force should drive pixels from visually similar objects to lie close together in the embedding space, while pixels from dissimilar objects to be mapped further apart. As motivated in the previous paragraph, this can be achieved by adopting a contrastive loss that takes augmented views of objects as positive pairs, and views of other objects as negatives. In this case, the push-force is found between different objects. We represent the objects by their mean pixel embedding.

Optimization objective. We modify the contrastive loss from Equation 1 to include the proposed pull- and push-forces. Positive pairs of object-centric crops $(\Psi_\eta(X), \Psi_\eta(X^+))$ are replaced with positive pairs of pixel embeddings: $(\mathbf{z}_i, \mathbf{z}_{\mathcal{M}_{X^+}})$ for $i \in \mathcal{M}_X$. In a similar way, the negative pairs $(\Psi_\eta(X), \Psi_\eta(X_k^-))$ are replaced with $(\mathbf{z}_i, \mathbf{z}_{\mathcal{M}_{X_k^-}})$. We obtain the following optimization criterion for a pixel $i \in \mathcal{M}_X$

$$\mathcal{L}_i = -\log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_{\mathcal{M}_{X^+}} / \tau\right)}{\sum_{k=0}^{K} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_{\mathcal{M}_{X_k^-}} / \tau\right)}. \tag{3}$$

The pixel embedding function Φ_θ maximizes the agreement between pixels and an augmented view of the object they belong to, while minimizing the agreement with other objects. We apply the pixel-wise loss $\mathcal{L}_i$ to all foreground pixels. The background pixels are not contrasted, since there could be multiple background objects on which we have no conclusive information. In this case, however, the network does not need to discriminate between pixels that fall inside or outside the object masks. As a consequence, the pixel embeddings can collapse to a single vector across an image. To prevent this, we regularize the feature space by including a separate linear head that predicts the saliency masks.

The supplementary materials provide pseudo-code of MaskContrast.

Interestingly, the proposed objective can also be viewed in an alternative way. Wang and Isola [76] showed that a contrastive loss optimizes two properties: (1) alignment of features from positive pairs and (2) uniformity of the feature distribution on a normalized hyper-sphere. From this viewpoint, our optimization objective can also be interpreted as optimizing the alignment of pixel embeddings based upon shared pixel ownership, while spreading pixel embeddings uniformly across the hyper-sphere $\mathcal{Z}$.

4. Experiments

4.1. Experimental Setup

Datasets. We conduct the bulk of our experimental analysis on the PASCAL [20] dataset following prior work [29, 95]. The `train_aug` and `val` splits are used during training and evaluation respectively. We perform additional experiments on the COCO [44] and DAVIS-2016 [59] datasets to verify if the pixel embeddings transfer to novel scenes. We use the annotations from Kirillov *et al.* [38] for the semantic segmentation task on COCO and evaluate on the PASCAL classes. On DAVIS-2016, the representations are used to compute correspondences for propagating object masks in videos. Only the first frame is annotated and we evaluate the propagated masks on the rest of the frames. We adopt the evaluation protocol from [32], and report the region similarity $\mathcal{J}$ and contour-based accuracy $\mathcal{F}$ scores.

Training setup. We use a DeepLab-v3 [10] model with dilated [86] ResNet-50 backbone [25]. The backbone is initialized from MoCo v2 [13] pre-trained on ImageNet, unless defined otherwise. We train the model for 60 epochs using batches of size 64. The model weights are updated through SGD with momentum 0.9 and weight decay $1e^{-4}$. The initial learning is set to 0.004 and decayed with a poly learning rate scheme. We use the same set of augmentations as SimCLR [11] to generate positive pairs (X, X^+), while making sure that each image contains at least a part of the salient object (area $> 10\%$). The features of negatives $\left\{\mathbf{z}_{\mathcal{M}_{X_0^-}}, \ldots, \mathbf{z}_{\mathcal{M}_{X_K^-}}\right\}$ are saved in a memory bank, with K set to 128. The negatives are encoded with a momentum-updated version of the network following [24]. We use dimension $D = 32$ and temperature $\tau = 0.5$.

Saliency estimation. We test both unsupervised and supervised saliency estimators to mine the object mask proposals. We adopt the BAS-Net [62] architecture. The *supervised saliency model* is trained on DUTS [74]. Differently, the *unsupervised saliency model* is trained on MSRA [14] using the approach from DeepUSPS [50]. MSRA considers less complex scenes from which the unsupervised training benefits. However, directly transferring the predictions

Method	LC (MIoU)
Supervised Saliency Model	6.5
MoCo v2 [13] (Unsupervised)	45.0
ImageNet (IN) Classifier (Supervised)	53.1
MaskContrast (MoCo v2 Init. - Unsup. Sal. Model)	58.4
MaskContrast (MoCo v2 Init. - Sup. Sal. Model)	62.2
MaskContrast (IN Classifier Init. - Unsup. Sal. Model)	61.0
MaskContrast (IN Classifier Init. - Sup. Sal. Model)	**63.9**

Table 1. **Baseline comparison** under the linear evaluation protocol on PASCAL.

to our target datasets, e.g. PASCAL, results in low-quality mask proposals when using the unsupervised model. We employ a simple bootstrapping procedure to improve the predictions on the target datasets. In particular, we obtain our final saliency estimator from training BAS-Net on pseudo-labels generated with the unsupervised DeepUSPS model on MSRA.

Implementation. We provide the implementation details of every method in the supplementary materials. The code and pre-computed saliency masks will be made available.

Scope. We adopt standard evaluation protocols [34, 95] for unsupervised semantic segmentation to benchmark our method. More specifically, we use linear probes (Sec. 4.3), direct clustering (Sec. 4.4) and a segment retrieval approach (Sec. 4.5) to quantify if the pixel embeddings are disentangled according to the semantic classes. This experimental setup differs from the typical setting used in self-supervised representation learning, where the evaluation focuses on fine-tuning the feature representations to various down-stream tasks. For completeness, we include additional fine-tuning experiments in Sections 4.6 - 4.7.

4.2. Ablation Studies

We examine the influence of the different components of our framework under the linear evaluation protocol following existing work [95]. The network weights are kept fixed and we train a 1 x 1 convolutional layer on top to predict the class assignments. Since the discriminative power of a linear classifier is low, the pixel embeddings need to be informative of the semantic class to solve the task in this way.

Baseline comparison. Table 1 compares several baselines. Applying a linear classifier on top of the saliency features results in the lowest performance (6.5%). This is to be expected since the saliency estimator only discriminates between two groups of pixels, i.e. the salient object vs. background. Differently, our method discovers a semantically structured embedding space, where pixels from visually similar objects lie close together, while pixels from dissimilar objects end up far apart. This allows a linear classifier to correctly group the pixels ($> 58.4\%$). Importantly, the results improve over the models from which the backbone weights were initialized (45.0% to 58.4% for MoCo and 53.1% to 61.0% for supervised pre-training). We conclude that the performance of our method can not be attributed to the use of a specific initialization. Also, it is beneficial to learn representations at pixel-, rather than at image-level, for the segmentation task. Finally, we observe further performance gains when including additional supervision, e.g. supervised pre-training on ImageNet (58.4% to 61.0%), or a supervised saliency estimator (58.4% to 62.2% and 61.0% to 63.9%).

Mask proposals. Table 2a compares three mask proposal strategies. Better numbers are reported when using salient object masks. We found that the regions extracted with the hierarchical segmentation algorithm were often too small to be representative of an object or part. In this way, the model does not learn useful information for the segmentation task. This confirms the hypothesis from Section 3.1, i.e. a good prior expresses object information.

Training mechanisms. Table 2b ablates some of the included training mechanisms. First, using augmented views to sample positive pairs improves the results, as we learn additional invariances. Second, including a memory bank results in further performance gains, because we can better estimate the distribution of negatives. Third, it is helpful to encode the negatives with a momentum-updated version of the network Φ_θ, as this enforces consistency in the memory bank (see also [24]). In summary, all three mechanisms positively contribute to the results.

Hyperparameter study. Table 2c studies the influence of the used temperature τ and number of negatives K. We conclude that the proposed algorithm is not very hyperparameter sensitive based upon the reported standard deviations.

4.3. Linear Classifier

Table 3a compares our method against competing works under the linear evaluation protocol on PASCAL.

MaskContrast vs. proxy tasks. The method substantially outperforms works based on proxy tasks. It is unlikely that a proxy task aligns the embeddings with the semantic groups in the dataset. In contrast, combining our proposed prior, i.e. shared pixel ownership, with a contrastive loss results in more semantically meaningful pixel embeddings.

MaskContrast vs. clustering. We outperform IIC [34] which used a clustering objective. As discussed earlier, the clusters strongly depend on the network initialization, which negatively impacts the learned features as the network can latch onto low-level information, like color, texture, contrast, etc. Differently, we suppress these problems by decoupling the prior from the network initialization.

Mask Proposals	LC (MIoU)
Hierarchical Seg. [1, 79]	30.5
Unsupervised Sal. Model	58.4
Supervised Sal. Model	62.2

(a) Comparison of three mask proposal mechanisms.

Augmented Views	Memory	Momentum Encoder	LC (MIoU)
✗	✗	✗	52.4
✓	✗	✗	54.0
✓	✓	✗	55.0
✓	✓	✓	58.4

(b) Analysis of the used training mechanisms.

Hyperparameter	Range	LC (MIoU)
Temperature τ	[0.1-1]	56.2 ± 1.4
Negatives K	[64-1024]	57.0 ± 0.6

(c) Hyperparameter study. We report the mean and standard deviation.

Table 2. **Ablation studies** of our method under the linear evaluation protocol on PASCAL. Tables 2b- 2c report results with masks from the unsupervised saliency estimator. We use MoCo v2 initial weights.

MaskContrast vs. contrastive learning. The method reports higher accuracy compared to existing contrastive self-supervised approaches. This group of works defined the contrastive loss at the global image- or patch-level. Naturally, our pixel embeddings are more predictive of the semantic segmentation task as we defined a contrastive learning objective at the pixel-level.

MaskContrast vs. boundary based. Finally, we outperform methods that relied on boundary detectors to group pixels together. We argue that the employed saliency masks incorporate higher level visual information compared to the regions obtained from boundary detectors.

4.4. Clustering

We verify whether the feature representations can be directly clustered in semantically meaningful groups using an off-line clustering criterion like K-Means. The number of clusters equals the number of ground-truth classes. The Hungarian matching algorithm is used to match the predicted clusters with the ground-truth classes and the results are averaged across five runs. Table 3b shows the results. Our learned pixel embeddings can be successfully clustered using K-Means on PASCAL. In contrast, the features representations obtained in prior works do not exhibit this behavior. We include additional results in the suppl. materials when applying overclustering.

4.5. Semantic Segment Retrieval

Next, we adopt a retrieval approach to examine our representations on PASCAL. First, we compute a feature vector for every salient object by averaging the pixel embeddings within the predicted mask. Next, we retrieve the nearest neighbors of the `val` set objects from the `train_aug` set. Table 4 shows a quantitative comparison with the state-of-the-art for the following 7 classes: bus, airplane, car, person, cat, cow and bottle. As before, we outperform prior works by significant margins. To facilitate future comparison, we also include results when evaluating on all 21 PASCAL classes. Figure 4 shows some qualitative results.

Method	LC	K-Means
Proxy task based:		
Co-Occurence [31]	13.5	4.0
CMP [88]	16.5	4.3
Colorization [94]	25.5	4.9
Clustering based:		
IIC [34]	28.0	9.8
Contrastive learning based:		
Inst. Discr. [78]	26.8	4.4
MoCo v2 [24]	45.0	4.3
InfoMin [69]	45.2	3.7
SWAV [9]	50.7	4.4
Boundary based:		
SegSort [29]†	36.2	-
Hierarch. Group. [95]†	48.8	-
ImageNet (IN) Classifier (Supervised)	53.1	4.7
MaskContrast (MoCo Init. + Unsup. Sal.)	58.4	35.0
MaskContrast (MoCo Init. + Sup. Sal.)	62.2	38.9
MaskContrast (IN Sup. Init. + Unsup. Sal.)	61.0	41.6
MaskContrast (IN Sup. Init. + Sup. Sal.)	**63.9**	**44.2**

(a) Linear classifier. (b) K-Means.

Table 3. **State-of-the-art comparison** on PASCAL *val* (MIoU). (†) Indicates results taken from [95]. Note that the authors use a slightly different evaluation protocol, i.e. without ImageNet pretraining, but with finetuning of the complete ASPP decoder.

Method	MIoU (7 classes)	MIoU (21 classes)
SegSort [29]	10.2	-
Hierarch. Group. [95]	24.6	-
MoCo v2 [13]	48.0	39.0
MaskContrast (Unsup. Sal.)	53.4	43.3
MaskContrast (Sup. Sal.)	**62.3**	**49.6**

Table 4. **State-of-the-art comparison** for semantic segment retrieval on the PASCAL `val` set. We use MoCo v2 initial weights.

4.6. Transfer Learning

We study the transferability of our pixel embeddings. Table 5 shows the results when pretraining on ImageNet and evaluating the generated pixel embeddings on a different target dataset. Interestingly, our representations transfer well across various datasets. Training a linear classifier to solve the segmentation task on PASCAL improves over the MoCo v2 baseline (55.4% for MaskContrast vs. 45.0% for MoCo when using an unsupervised saliency model). A similar effect can be observed on COCO (45.0% for MaskCon-

Figure 4. **Nearest neighbors** for queries (1st col.) on PASCAL.

Model	PASCAL (MIoU)↑	COCO (MIoU)↑	DAVIS '16 $\mathcal{J}_m$ ↑	DAVIS '16 $\mathcal{F}_m$ ↑
MoCo v2	45.0	35.2	77.1	77.2
MaskContrast (Unsup. Sal.)	55.4	45.0	78.0	77.8
MaskContrast (Sup. Sal.)	**57.2**	**47.2**	**82.0**	**80.9**

Table 5. **Transfer learning setup.** All models were pre-trained on ImageNet. We use MoCo v2 initial weights. Results on PASCAL and COCO are reported for a linear classifier. On DAVIS, we freeze the representations and adopt the protocol from [32].

trast vs. 35.2% for MoCo). Finally, our representations also transfer well to the semantic object segmentation task on DAVIS-2016. This dataset covers a rich set of natural image augmentations like viewpoint changes, occlusions, etc., for which our pixel embeddings have learned invariances.

The gains observed across all three benchmarks show that the learned representations are not limited to a specific dataset. We conclude that the use of a mid-level visual prior can be useful for self-supervised representation learning.

4.7. Semi-Supervised Learning

The proposed method can alternatively be used as a pre-training strategy for semantic segmentation. That is, the model is fine-tuned in a semi-supervised way on PASCAL. We use 1%, 2%, 5%, 12.5% and 100% of the `train_aug` split as labeled examples. We initialize our model from supervised pre-training on ImageNet. This weight initialization is commonly used in semantic segmentation. Furthermore, directly fine-tuning a model initialized in the same way serves as a strong baseline. Table 6 shows the results.

The representations generated with our method yield higher performance after fine-tuning, compared to supervised pre-training on ImageNet. This holds true when using both an unsupervised and supervised saliency estimator to predict the object mask proposals. Predictably, the gains become smaller when more labeled examples are available (see also [97]). In conclusion, unsupervised learning of pixel embeddings can complement a pre-training strategy based on an image-level optimization criterion.

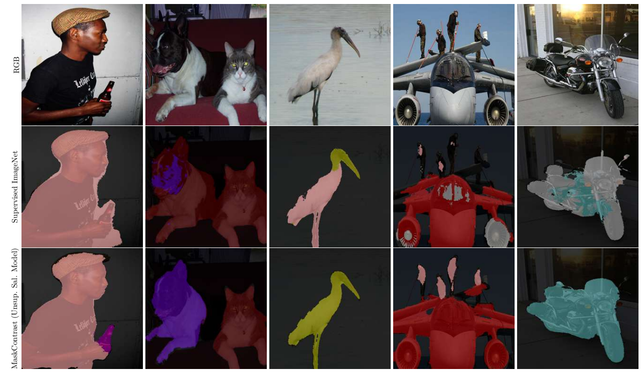

Figure 5. **Qualitative comparison** after fine-tuning on PASCAL using 1 % of labeled data. We use supervised pre-training on ImageNet (middle) or our method (bottom) to initialize the weights.

Label Fraction	1%	2%	5%	12.5%	100%
ImageNet Classifier Init.	43.4	55.2	62.7	68.4	78.0
+ MaskContrast (Unsup. Sal.)	50.5	57.2	64.5	69.0	78.4
+ MaskContrast (Sup. Sal.)	**51.5**	**59.6**	**65.3**	**69.4**	**78.6**

Table 6. **Semi-supervised fine-tuning** on PASCAL (MIoU).

5. Discussion and Limitations

This work presented a general two-step framework based upon a mid-level visual prior for tackling unsupervised semantic segmentation. The proposed setup prevents the model from latching onto low-level image features, a problem present in prior works that relied on end-to-end clustering, proxy tasks or low-level visual cues. Instead, MaskContrast learns pixel embeddings which incorporate more semantically meaningful information (see Figure 4). As a result, we were able to tackle the semantic segmentation task under a fully unsupervised setup on a diverse dataset like PASCAL. Further, experimental evaluation showed that our pixel embeddings have several other interesting properties: the ability for semantic segment retrieval, transfer learning and semi-supervised fine-tuning.

Still, there are some limitations to our method. The object mask proposals were obtained using a salient object estimator - which can retrieve only a limited number of objects per image. Alternative ways to mine the object mask proposals can be explored for tackling even more challenging datasets where many objects can exist per image. In particular, we could see additional sensory data [67] or other techniques [60] being used that are better suited for this type of images. The optimization criterion from Equation 3 could then be extended accordingly. Given the viability of our framework, we believe these are interesting research directions.

Acknowledgment. The authors thankfully acknowledge support by Toyota via the TRACE project and MACCHINA (KU Leuven, C14/18/065).

References

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *T-PAMI*, 2010.

[2] Yuki M Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *NeurIPS*, 2020.

[3] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020.

[4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016.

[5] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016.

[6] Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *ICML*, 2017.

[7] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12):5706–5722, 2015.

[8] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

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

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

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

[12] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020.

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

[14] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *T-PAMI*, 2015.

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[16] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.

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

[18] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *ICLR*, 2017.

[19] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *NIPS*, 2019.

[20] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

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

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

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

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

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[26] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *NeurIPS*, 2015.

[27] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *CVPR*, 2017.

[28] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018.

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

[30] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. In *SIGGRAPH*, 2016.

[31] Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H Adelson. Learning visual groups from co-occurrences in space and time. *arXiv preprint arXiv:1511.06811*, 2015.

[32] Allan Jabri, Andrew Owens, and Alexei A. Efros. Spacetime correspondence as a contrastive random walk. In *NeurIPS*, 2020.

[33] Simon Jenni and Paolo Favaro. Self-supervised feature learning by learning to spot artifacts. In *CVPR*, 2018.

[34] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, 2019.

[35] Bowen Jiang, Lihe Zhang, Huchuan Lu, Chuan Yang, and Ming-Hsuan Yang. Saliency detection via absorbing markov chain. In *ICCV*, 2013.

[36] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *T-PAMI*, 2020.

[37] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017.

[38] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019.

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

[40] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, 2017.

[41] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[42] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, 2013.

[43] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016.

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

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

[46] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *CVPR*, 2017.

[47] Aravindh Mahendran, James Thewlis, and Andrea Vedaldi. Cross pixel optical-flow similarity for self-supervised learning. In *ACCV*, 2018.

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

[49] T Nathan Mundhenk, Daniel Ho, and Barry Y Chen. Improvements to context based self-supervised learning. In *CVPR*, 2018.

[50] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction via self-supervision. In *NeurIPS*, 2019.

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

[52] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *ICCV*, 2017.

[53] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *CVPR*, 2018.

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

[55] Yassine Ouali, Céline Hudelot, and Myriam Tami. Autoregressive unsupervised image segmentation. In *ECCV*, 2020.

[56] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015.

[57] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015.

[58] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[59] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

[60] Pedro OO Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *NeurIPS*, 2015.

[61] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. In *NeurIPS*, 2020.

[62] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019.

[63] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *CVPR*, 2018.

[64] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *T-PAMI*, 2000.

[65] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *CVPR*, 2018.

[66] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *ECCV*, 2018.

[67] Hao Tian, Yuntao Chen, Jifeng Dai, Zhaoxiang Zhang, and Xizhou Zhu. Unsupervised object detection with lidar clues. *arXiv preprint arXiv:2011.12953*, 2020.

[68] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[69] Yonglong Tian, C. Sun, Ben Poole, Dilip Krishnan, C. Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arxiv preprint arXiv:2005.10243*, 2020.

[70] Koen EA Van de Sande, Jasper RR Uijlings, Theo Gevers, and Arnold WM Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011.

[71] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *ECCV*, 2020.

[72] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *CVPR*, 2017.

[73] Jingdong Wang, Huaizu Jiang, Zejian Yuan, Ming-Ming Cheng, Xiaowei Hu, and Nanning Zheng. Salient object detection: A discriminative regional feature integration approach. *IJCV*, 2017.

[74] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017.

[75] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, 2017.

[76] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

[77] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *arXiv preprint arXiv:1904.09146*, 2019.

[78] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[79] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015.

[80] Jia Xu, Alexander G Schwing, and Raquel Urtasun. Learning to segment under various forms of weak supervision. In *CVPR*, 2015.

[81] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. *CVPR*, 2013.

[82] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *CVPR*, 2020.

[83] Yanchao Yang, Brian Lai, and Stefano Soatto. Time-supervised primary object segmentation. *arXiv preprint arXiv:2008.07012*, 2020.

[84] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *CVPR*, 2019.

[85] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, 2019.

[86] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[87] Xiaohang Zhan, Ziwei Liu, Ping Luo, Xiaoou Tang, and Chen Change Loy. Mix-and-match tuning for self-supervised semantic segmentation. In *AAAI*, 2018.

[88] Xiaohang Zhan, Xingang Pan, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised learning via conditional motion propagation. In *CVPR*, 2019.

[89] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *CVPR*, 2017.

[90] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *CVPR*, 2020.

[91] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *CVPR*, 2018.

[92] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *CVPR*, 2019.

[93] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *CVPR*, 2017.

[94] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016.

[95] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. *NeurIPS*, 2020.

[96] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *CVPR*, 2014.

[97] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. In *NeurIPS*, 2020.

[98] Wenbin Zou and Nikos Komodakis. Harf: Hierarchy-associated rich features for salient object detection. In *ICCV*, 2015.