Large-scale Unsupervised Semantic Segmentation

Shang-Hua Gao[1]* **Zhong-Yu Li**[1]* **Ming-Hsuan Yang**[2] **Ming-Ming Cheng**[1]† **Junwei Han**[3] **Philip Torr**[4]
Nankai University[1] University of California at Merced[2]
Northwestern Polytechnical University[3] University of Oxford[4]

Abstract

Powered by the ImageNet dataset, unsupervised learning on large-scale data has made significant advances for classification tasks. There are two major challenges to allow such an attractive learning modality for segmentation tasks: i) a large-scale benchmark for assessing algorithms is missing; ii) unsupervised shape representation learning is difficult. We propose a new problem of **l**arge-scale **u**nsupervised **s**emantic **s**egmentation (LUSS) with a newly created benchmark dataset to track the research progress. Based on the ImageNet dataset, we propose the ImageNet-S dataset with 1.2 million training images and 40k high-quality semantic segmentation annotations for evaluation. Our benchmark has a high data diversity and a clear task objective. We also present a simple yet effective baseline method that works surprisingly well for LUSS. In addition, we benchmark related un/weakly supervised methods accordingly, identifying the challenges and possible directions of LUSS. The benchmark is available on https://github.com/UnsupervisedSemanticSegmentation.

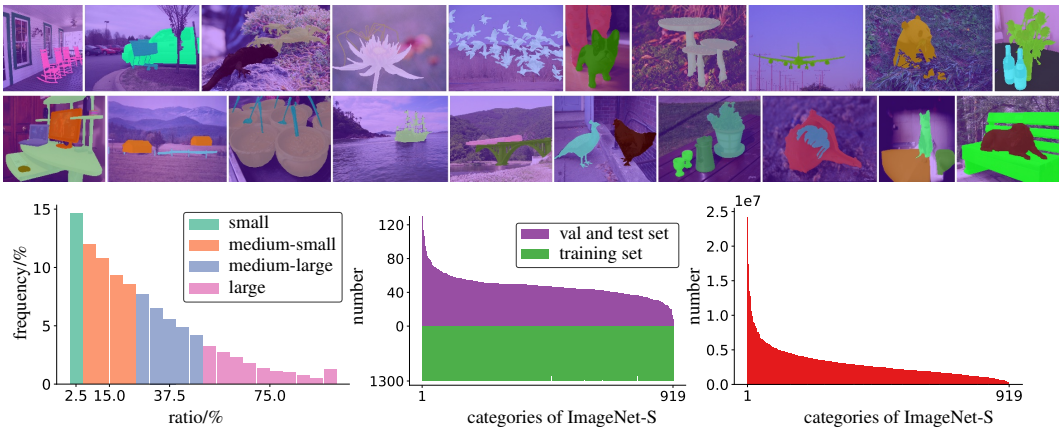
1 Introduction

Semantic segmentation [67, 115, 83, 4, 13, 124, 126], aiming to label image pixels with category information, has drawn much attention in recent years. Due to the inherent challenges of this task, most efforts focus on semantic segmentation under environments with limited diversity [23, 130, 114] and data scale [28, 7]. For instance, the PASCAL VOC segmentation dataset only contains about 2k images, while the BDD100K [114] focuses on road scenes. Numerous approaches have achieved impressive results on these restricted environments [13, 14, 31, 100, 131, 125, 66, 61, 65, 116]. Significantly scale up the problem often results in research modality change, *e.g.,* from PASCAL VOC [28] to ImageNet [84]. This motivates us to consider a far more challenging problem: is semantic segmentation possible for large-scale real-world environments with a wide diversity?

However, due to the huge data scale and privacy issues, annotating images with extremely expensive pixel-level human annotations or even image-level labels is almost impossible. Lacking sufficient benchmark data limits the large-scale semantic segmentation. When trained with millions or even billions of images, *e.g.,* ImageNet [84], JFT-300M [90], and Instagram-1B [68], unsupervised learning of classification model has recently shown a comparable ability to supervised training [93, 37, 16]. To facilitate real-world semantic segmentation, we propose a new problem of **L**arge-scale **U**nsupervised **S**emantic **S**egmentation (LUSS). The LUSS task aims to assign labels to pixels from the large-scale data without human-annotation supervision. Many challenges, *e.g.,* shape/category representation learning, pixel-level network designing, and semantic clustering algorithm under the large-scale data and unsupervised setting, are developed to achieve this goal. Specifically, we need to extract semantic representations with category and shape cues. Category-related representations are required to distinguish different classes, and shape-related representations, *e.g.,* objectness, boundary, are the

*Equal contribution
†M.M. Cheng (cmm@nankai.edu.cn) is the corresponding author.

Preprint. Under review.



(a) Distribution of object size. (b) Number of images per class. (c) Number of pixels per class.

Figure 1: Visualization and distribution statistics of the ImageNet-S (val and test) dataset.

essential pixel-level cues for semantic segmentation. Efficient pixel-level networks are crucial to reducing the huge computational cost brought by large-scale data. Generating categories from large-scale data requires robust and efficient semantic clustering algorithms. Solving these challenges for LUSS also facilitates many related tasks. For example, the learned shape representations from LUSS can be utilized as the pre-training for pixel-level downstream tasks, *e.g.,* semantic segmentation [13, 14] and instance segmentation [38] under restricted data scale and diversity.

To facilitate the LUSS task, we propose a benchmark with large-scale data with high diversity, a clear objective of learning semantic segmentation without direct/indirect human-annotation, and sufficient evaluation protocols from different perspectives for the LUSS task. Large-scale data with sufficient diversity bring challenges to LUSS, but it also provides the source for obtaining extensive representation cues. Due to a lack of sufficient data, a few unsupervised segmentation methods [77, 119, 46, 98] mainly focus on small data with small diversity, thus are not suitable for the LUSS task. We present a large-scale benchmark dataset for the LUSS task, namely ImageNet-S, based on the commonly used ImageNet dataset [84] in category representation learning works. We remove the unsegmentable categories, *e.g.,* bookshop, and utilize 919 categories with about 1.2 million images in ImageNet for unsupervised training. Then we annotate about 40k images in the validation set of ImageNet with precise pixel-level semantic segmentation masks for LUSS evaluation. Following the more precise re-annotated image-level labels in [5], we enable ImageNet with multiple categories within one image. The ImageNet-S dataset provides large-scale and high-diversity data for fairly LUSS training and sufficient evaluation.

We present one basic pipeline for the LUSS task, including unsupervised representation learning, label generation, fine-tuning, and evaluation. Based on the pipeline and ImageNet-S dataset, we benchmark related methods, *i.e.,* category-related unsupervised learning [37, 16, 11, 48, 110], weakly supervised semantic segmentation [103, 12, 56], and identify the challenges and possible directions of LUSS. We then present a baseline method for LUSS by enhancing the shape representation learning, which composes an efficient network, pixel attention for pixel-level representation discrimination, and image-to-pixel/pixel-to-pixel representation alignment. In this work, we make two main contributions:

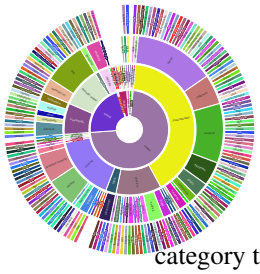
- We propose a new large-scale unsupervised semantic segmentation problem and a benchmark containing the ImageNet-S dataset with 40k pixel-level annotated evaluation images.
- We present one basic pipeline, benchmarks of related works, and a baseline method for LUSS.

2 Large-scale Unsupervised Semantic Segmentation Benchmark

Prior to the recent advances of deep learning, a plethora of approaches have been developed to segment objects with non-parametric methods (*e.g.,* label transfer [64], matching [85, 96], and distance evaluation [69]), and handcrafted features (*e.g.,* boundary [70], and superpixels [32]). Numerous data-driven deep learning models have recently been developed for semantic segmentation [67, 115, 83, 4, 13, 124, 60, 63, 62, 43, 17]. Based on pre-trained representations [77, 119, 46, 98], a few unsupervised semantic segmentation (USS) models have been proposed based on segment sorting [46], mutual information maximization [77], and region contrastive learning[98]. However,

Table 1: Left: Categories and number of images comparison between the ImageNet-S dataset and existing semantic segmentation datasets. Right: Category structure tree of the ImageNet-S dataset.

Dataset	category	train	val	test
PASCAL VOC 2012 [28]	20	1464	1449	1456
CityScapes [23]	19	2975	500	1525
ADE20K [130]	150	20210	2000	3000
ImageNet-S_{50}	50	64431	760	1703
ImageNet-S_{300}	300	384862	4125	9172
ImageNet-S	919	1183322	12466	27717



category tree

several issues limit the applicability of USS to the LUSS task. 1) Existing methods focus on small datasets [77, 119, 46, 98] and easy categories [77] (*e.g.,* sky and ground). Because of the insufficient data, the advantages of the large-scale data (*e.g.,* unsupervised learning of rich representations from the large-scale data) are not explored, and the challenges of large-scale data (*e.g.,* huge computational cost) are also ignored. 2) Due to the lack of clear problem definition and standardized evaluation, some methods utilize supervised prior knowledge, *e.g.,* supervised edge detection [46] and supervised saliency detection [98], making it difficult to evaluate these methods. As such, we develop a LUSS benchmark with a clear objective, large-scale training data, and comprehensive evaluation protocols.

The LUSS task aims to learn semantic segmentation from large-scale images without any direct/indirect human annotations. Given a large set of images, a LUSS model assigns self-learned labels to each pixel of all images. We give one of the possible pipelines for LUSS as follows: First, a model learns both category and shape representations from large-scale data without human annotation. Using the learned feature representations, the model conducts label clustering and assignment to get the pseudo labels. Then, the model is fine-tuned on the generated pseudo labels to refine the segmentation results. Ideally, label assignment and refining can be implicitly contained in the unsupervised representation learning process.

2.1 Large-scale ImageNet-S Dataset

The LUSS task requires large-scale data to learn rich representations. Existing segmentation datasets, *e.g.,* PASCAL VOC [28], ADE20K [130], and CityScapes [23], contain a limited number of images under few scenes. To facilitate the LUSS task, we present a large-scale ImageNet-S dataset by collecting data from the ImageNet dataset [84], which has been widely used for visual classification and segmentation [39, 109, 33, 37, 16, 11, 34]. As shown in Tab. 1, the ImageNet-S dataset is much larger than existing datasets in terms of image amount and category diversity. Due to the limited space, we discuss the main properties of the dataset, and more details are presented in the supplementary.

Annotation. We annotate the validation/testing sets in the ImageNet-S dataset for LUSS evaluation. As Lucas *et al.* [5] observe that the ImageNet dataset has incorrect annotations, *e.g.,* incorrect labels and missing multiple categories, we annotate the pixel-level semantic segmentation mask of images following the relabeled image-level annotations [5]. The objects indicating the image-level labels are annotated, and other parts are annotated as the background. We annotate 40183 images with precise pixel-level masks, and some visualized annotations are shown in Fig. 1.

Statistics and distribution. As shown in Tab. 1, after removing the unsegmentable categories in the ImageNet dataset, *e.g.,* bookshop, valley, and library, the ImageNet-S dataset contains 1183322 training, 12466 validation, and 27717 testing images from 919 categories. As it is more difficult to segment smaller objects, we divide the objects into groups, *i.e.,* small, medium-small, medium-large, and large object size, according to the ratio of objects to the image. The distribution shown in Fig. 1(a) indicates the majority of objects are relatively small. Fig. 1(b) shows the number of images for most categories is balanced, while the number of pixels per category shown in Fig. 1(c) presents the long-tail distribution. To facilitate the research under a low computational budget, we develop two subsets contains 50 and 300 categories, namely ImageNet-S_{50} and ImageNet-S_{300}, in which categories are overlapped with the ImageNet-S dataset. Considering the great difficulty of the LUSS task, we choose 50 common and distinguishable categories in daily life for ImageNet-S_{50}. The ImageNet-S_{300} is composed of ImageNet-S_{50} and 250 randomly sampled categories.

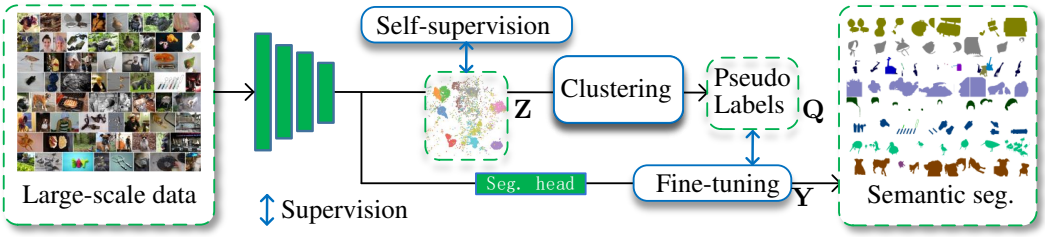


Figure 2: One of the possible basic pipelines of the LUSS task. 1) Representation learning with self-supervision. 2) Clustering learned features to generate pseudo categories and labels. 3) Fine-tuning on segmentation head with pseudo label supervision. 4) Inference semantic segmentation mask.

2.2 Evaluation Protocols

Unlike the supervised vision tasks, categories are generated by the model in the LUSS task, which needs to be the same as the categories in the ground-truth (GT) during evaluation. The training and testing set of the ImageNet-S dataset contains the same categories. We present the default image-level matching scheme, while an effective matching scheme should improve LUSS performance. Suppose the set for matching (normally validation set) has N images and C categories. Given the image set $\mathbf{D} = \{\mathbf{D}_k, k \in [1, N]\}$ with GT labels $\mathbf{G} = \{\mathbf{G}_k, k \in [1, N]\}$ and predicted labels $\mathbf{P} = \{\mathbf{P}_k, k \in [1, N]\}$, where $\mathbf{G}_k$ and $\mathbf{P}_k$ are the GT and predicted category sets of the image $\mathbf{D}_k$. We calculate the matching matrix $\mathbf{S} \in R^{C\times C}$ between generated and GT categories, in which $\mathbf{S}_{ij}$, representing the matching degree between i-th generated category and j-th GT category, is larger when two categories are more likely to be the same category. That is, $\mathbf{S}_{ij}$ is computed by $\mathbf{S}_{ij} = \sum_{k=1}^{N} \mathbb{I}\{(i, j) \in \mathbf{P}_k \times \mathbf{G}_k\}$, where $\mathbf{P}_k \times \mathbf{G}_k$ is the Cartesian product of $\mathbf{P}_k$ and $\mathbf{G}_k$, and the indicator $\mathbb{I}$ equals 1 when (i, j) belongs to $\mathbf{P}_k \times \mathbf{G}_k$. With the matching matrix $\mathbf{S} \in R^{C\times C}$, we find the bijection $\mathbf{f} : i \mapsto j$ between generated and GT categories by maximizing $\sum_{i=1}^{C} \mathbf{S}_{i,\mathbf{f}(i)}$ using the Hungarian algorithm [53].

We use the image-level accuracy (img-acc), normalized mutual information (NMI), mean intersection over union (mIoU), and F-measure (F_β) as the evaluation metrics used for the benchmarking of the LUSS task, and the detailed implementations are shown in the supplementary document. The img-acc can evaluate the category representation ability of models. As many images contain multiple labels, we follow [5] and treat the predicted label as correct if the prediction with the largest area is within the GT label list. NMI can reflect the distribution difference between the generated pseudo labels and GT labels. Similar to the supervised semantic segmentation task [28, 130], we utilize the mIoU metric to evaluate the segmentation mask quality. In addition to category-related representation, we utilize F_β to evaluate the shape quality, which ignores the semantic categories. Specifically, we treat all pixels of the foreground as a single category to calculate F_β.

3 Benchmarking Related Tasks

3.1 Basic LUSS Pipeline

Based on the proposed benchmark described, we present one basic pipeline for LUSS (see Fig. 2). **Step 1.** A randomly initialized model, *e.g.,* ResNet, is trained with self-supervision of pretext tasks, *e.g.,* contrastive learning, to implicitly learn shape and category representations. We use ResNet-18 for the ImageNet-S_{50} dataset, and ResNet-50 for ImageNet-S_{300} as well as ImageNet-S datasets. After representation learning, we obtain the features set $\mathbf{Z} = \{\mathbf{z}_k \in R^{L\times H\times W}, k \in [1, N]\}$ for all training images, where L, H, and W are the number of dimensions, height, and width of features. **Step 2.** We cluster $\mathbf{Z}$ to obtain C pseudo category and assign generated categories to each image. To save the clustering cost, we use the image-level feature $\hat{\mathbf{Z}} = \{\hat{\mathbf{z}}_k \in R^L, k \in [1, N]\}$ by pooling $\mathbf{Z}$ on the spatial dimension, and utilize the k-means clustering algorithm over $\hat{\mathbf{Z}}$ to assign pseudo labels $\mathbf{Q} = \{\mathbf{q}_k, k \in [1, N]\}$ to images. A proper pixel-level label clustering and assignment scheme should give more accurate segmentation results. **Step 3.** We add a conv 1×1 layer with channels of $L \times C$ and a global averaged pooling layer following the model output as the segmentation head. The output features $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_k \in R^C, k \in [1, N]\}$ from this head are supervised with $\mathbf{Q}$ to fine-tune

Table 2: Performance comparison between unsupervised representation learning methods and our proposed shape-enhanced representation learning methods on the LUSS benchmark using ImageNet-S_{50} and ImageNet-S_{300} dataset. Except for [98] that is designed for segmentation, other methods share the same basic pipeline (Sec. 3.1) but have different representation learning schemes. Our-D/Our-S4 denotes conducting segmentation using output features from the final output of the modified DeeplabV3+* model or stage4 of the ResNet.

Types	**Methods**	ImageNet-S_{50} mIoU val	mIoU test	NMI train	Img-Acc val	Img-Acc test	F_β val	F_β test	ImageNet-S_{300} mIoU val	mIoU test	NMI train	Img-Acc val	Img-Acc test	F_β val	F_β test
Supervised	supervised.	33.9	34.7	87.1	85.9	86.8	55.6	55.3	31.7	31.7	87.1	86.8	87.0	56.8	57.0
Basic	Rotation[35]	14.4	14.3	47.2	38.9	37.6	50.3	50.5	4.9	4.7	39.6	16.0	15.1	40.6	40.5
Contrastive	NPID[107]	13.5	13.0	35.5	37.9	35.5	51.3	51.0	6.1	5.9	37.6	21.7	19.4	47.8	47.9
	SimCLR[16]	22.3	22.4	55.1	57.8	55.5	52.0	51.5	11.6	11.4	49.7	36.2	34.2	51.5	51.5
	MoCov2[18, 37]	19.6	19.4	50.9	49.3	49.8	52.9	52.5	18.7	18.0	62.3	49.1	47.8	53.6	53.7
Clustering	DeepClustering[9]	17.3	17.9	44.3	41.7	41.2	51.0	50.9	13.2	12.9	51.9	36.6	34.8	54.0	54.1
	ODC[120]	20.3	18.3	51.8	49.3	48.3	54.4	54.1	7.5	7.3	43.9	25.4	24.0	51.0	50.9
	PCL[57]	20.0	19.7	53.2	47.9	47.9	54.6	53.7	16.3	15.5	62.8	44.6	42.4	54.8	54.8
	SwAV[11]	21.7	20.8	63.3	57.4	57.7	51.0	50.1	23.7	23.4	71.9	59.8	58.3	55.5	55.6
Other	BYOL[48]	20.6	20.3	55.2	51.6	50.1	51.4	51.2	14.4	14.3	55.2	38.3	36.8	49.6	49.5
	PixelPro[110]	25.4	24.3	56.1	54.1	53.1	53.8	53.4	14.0	13.8	51.6	35.9	35.2	49.7	49.6
	MaskContrast[98]	23.9	22.6	51.8	47.6	48.0	58.9	57.4	15.7	15.4	62.4	42.4	40.6	56.6	56.3
	Ours-S4	35.7	35.2	75.3	72.8	73.1	60.5	59.4	26.3	25.6	73.5	60.7	59.1	60.3	60.2
	Ours-D	38.7	39.4	75.1	72.4	72.5	60.2	59.0	26.8	26.4	74.0	57.5	55.4	59.0	59.0

Table 3: Performance comparison between unsupervised representation learning methods and our proposed shape-enhanced representation learning methods on the LUSS benchmark using the ImageNet-S dataset. The mIoU under different object size range is calculated in the test set.

Methods	ImageNet-S mIoU val	mIoU test	S.	M.S.	M.L.	L.	NMI train	Img-Acc val	Img-Acc test	F_β val	F_β test
supervised.	29.2	28.9	5.5	25.4	31.6	24.1	81.4	83.6	84.1	54.8	54.6
MoCov2 [18, 37]	11.0	10.1	1.5	8.3	11.1	8.7	57.2	33.7	31.1	52.7	52.8
PCL [57]	14.4	13.9	1.9	11.5	15.1	12.0	63.2	39.3	37.1	55.2	55.1
SwAV [11]	15.3	14.7	2.9	13.0	15.5	11.5	67.2	46.9	45.2	56.6	56.6
Ours-S4	18.3	17.4	3.2	15.0	18.9	14.0	68.2	47.4	45.5	60.0	60.0
Ours-D	19.7	18.4	2.6	14.3	20.8	16.1	68.8	45.0	42.2	59.5	59.4

the model. **Step 4.** During inference, since $\mathbf{Q}$ only contains image-level labels, we follow the mechanism of class activation maps (CAM) [129] to remove the pooling layer and obtain the output features $\mathbf{Y} = \{\mathbf{y}_k \in R^{C\times H\times W}, k \in [1, N]\}$. For each pixel embedding $\mathbf{u} \in R^C$ in $\mathbf{y}_k$, we get the segmentation labels as follow: $\mathbf{p} = \arg\max_{i\in[1,C]}(\mathbf{u}_i) \cdot \mathbb{I}\{\max(\mathbf{u}) \geq \tau\}$, where $\mathbb{I}\{\max(\mathbf{u}) \geq \tau\}$ is the indicator that equals 1 if $\max(\mathbf{u}) > \tau$, and τ is the threshold between semantic foreground and background. More details can be found in the supplementary document. Noted that this pipeline is not the only option, and other pipelines are also encouraged for the LUSS task. We evaluate the existing methods, *e.g.,* unsupervised representation learning, weakly supervised segmentation methods, with the basic pipeline on the LUSS task.

3.2 Unsupervised Representation Learning

Unsupervised representation learning approaches with pretext tasks [49, 6, 121], *e.g.,* colorization [122, 47, 54], jigsaw puzzles [73, 75, 71], inpainting [79], adversarial learning [25, 26], context prediction [24, 72], counting [74], rotation predictions [35, 71], cross-domain prediction [80], contrastive learning [76, 40, 37, 16, 93, 94] and clustering [113, 58, 11], facilitate models learning semantic features from training examples. As the basic requirement of LUSS, category-related representation is used to distinguish scenes from different classes. To analyze the category representation ability of unsupervised learning methods, we categorize and benchmark some representative methods on the LUSS benchmark.

Benchmarking unsupervised learning methods. Numerous pretext tasks, *e.g.,* colorization [122, 47, 54], jigsaw puzzles [73, 75, 71], and rotation predictions [35, 71], enable models to learn task-

specific semantic representations. As these pretext tasks are not designed for category representation learning, the learned features are not necessarily effective for the LUSS tasks. As the core of unsupervised contrastive learning methods [107, 76, 3, 112, 8, 37, 16, 48, 22, 99, 78], instance discrimination with the contrastive loss [21, 36, 27] considers images from different views [93, 94, 71] or augmentations [16, 37] as pairs. In addition, it forces the model to learn representations by pushing "negative" pairs away and pull "positive" pairs closer. As shown in Tab. 2, the contrastive-based methods improve the model performance by a large margin compared to the rotation pretext task, *e.g.,* SimCLR [16] outperforms Rotation [35] by 17.9% img-acc on the ImageNet-S_{50} dataset. The learned representations are more related to object categories as the nature of contrastive loss is to distinguish among images. Some approaches [41, 48, 110], *e.g.,* BYOL [48], maximize the similarity of different output versions of the image and avoid negative pairs, and achieve similar performance as contrastive methods. Nevertheless, as the concept of the category is not included in contrastive methods, they are less effective for category-related tasks. Another line of work introduces the clustering strategy to unsupervised learning [132, 9, 10, 50, 111, 113, 58, 120, 57, 11], and encourages a group of images to have feature representations close to that of a cluster center. We observe that the clustering-based methods further outperform the contrastive methods in terms of image-level accuracy, *e.g.,* SwAV [11] achieves 57.7% test img-acc compared to the 55.5% of SimCLR. The clustering strategy encourages stronger category-related representations with category centroids compared to distinguishing positive/negative pairs in contrastive methods.

What role does the category play in the LUSS task? To answer this question, we use the models trained with image-level supervision as the baseline. As shown in Tab. 2, the supervised model performs better than the unsupervised models in terms of mIoU. In addition, it outperforms unsupervised models in terms of image-level metrics, *i.e.,* NMI and image-level accuracy, by a large margin. In contrast, the performance gap in shape-related metrics, *i.e.,* F_β, is quite small. These results show that category features indeed facilitates the LUSS task. However, shape features cannot be learned solely by category representation learning.

3.3 Unsupervised vs. Weakly Supervised Semantic Segmentation

Weakly supervised semantic segmentation (WSSS) [105, 52, 59, 2, 12] aims to carry out the task using weak annotations, *e.g.,* image-level labels, which is related to LUSS as both require shape features. However, some modules in typical WSSS methods, *e.g.,* supervised ImageNet$_{1k}$ pre-trained models [2, 1, 30, 92, 12], image-level GT labels [88, 92], and large network architectures [88, 103, 12], are not applicable to the LUSS tasks. To analyze the influence of these typical settings in WSSS methods on LUSS, we evaluate some state-of-the-art WSSS methods on the LUSS benchmark by using self-generated pseudo image-level labels.

Pre-trained models. One of the main challenges in LUSS is to learn effective representations without supervision. However, the effect of representation learning, *i.e.,* using weights pre-trained with different approaches, is less explored in the WSSS methods. Existing WSSS methods mostly utilize supervised ImageNet$_{1k}$ pre-trained models and fine-tune models on the semantic segmentation dataset [2, 1, 30, 92, 12], *e.g.,* PASCAL VOC [28]. To understand the importance of pre-training, we use different pre-trained models for SEAM [103], SC-CAM [12], and AdvCAM [56], as shown in Tab. 4. We observe that replacing the supervised ImageNet$_{1k}$ with the supervised ImageNet$_{50}$ dataset in SEAM [103] reduces the test mIoU from 44.5% to 35.8%. Replacing the supervised models with unsupervised models, *i.e.,* MoCo, and SwAV, further reduces the test mIoU to 19.1% and 22.3%, respectively. Both SC-CAM and AdvCAM suffered from the same issue, indicating WSSS methods rely heavily on supervised pre-training. The lack of supervised pre-training makes the representation learning crucial to the LUSS task. And our ImageNet-S dataset provides a basis for fairly evaluating the representation quality of pre-trained models.

Network architectures. Numerous network architectures have been developed to improve WSSS, including multi-scale enhancement [106] and affinity prediction [2, 1, 30]. Due to the small size of the PASCAL VOC dataset, many state-of-the-art WSSS methods improve the performance using large models with extensive parameters and computational cost, *e.g.,* wide ResNet-38 [108, 88, 103, 12] and ResNet with small output strides [15, 56]. As the proposed ImageNet-S datasets are 44 to 800 times larger than PASCAL VOC, the computational cost of training LUSS models with large models used by WSSS methods is prohibitively high. To analyze the effect of model architectures, we change the network in SEAM [103] for both WSSS and LUSS (see Tab. 4). We remove the Deeplab re-training step used in WSSS methods for fair comparisons. In WSSS, when replace the

Table 4: Ablation of shipping WSSS methods to the LUSS task. Properties in WSSS, *i.e.,* supervised pre-trained models, image-level GT labels, and large networks, that are not applicable in LUSS, make WSSS methods have a large performance drop in the LUSS task.

ImageNet-S_{50}	Arch.	Param./MACC	pre-train	Labels	mIoU		Img-Acc		F_β	
					val	test	val	test	val	test
SEAM[103]	ResNet-38 [108]	105.5M/100.4G	Sup. ImageNet$_{1k}$	GT	49.7	49.6	96.6	95.7	61.5	60.9
	ResNet-18 [39]	11.3M/1.9G	Sup. ImageNet$_{1k}$	GT	45.2	44.5	90.9	90.4	55.9	54.5
	ResNet-18 [39]	11.3M/1.9G	Sup. ImageNet-50	GT	35.1	35.8	81.2	81.5	46.3	46.5
	ResNet-18 [39]	11.3M/1.9G	MoCo. ImageNet-S_{50}	-	19.0	19.1	45.1	46.7	45.1	45.3
	ResNet-18 [39]	11.3M/1.9G	SwAV. ImageNet-S_{50}	-	22.1	22.3	54.6	53.5	41.1	41.1
	DeeplabV3+ [14]	16.7M/3.3G	SwAV. ImageNet-S_{50}	-	29.1	28.7	67.1	66.6	42.6	41.9
SC-CAM[12]	ResNet-18[39]	11.5M/1.8G	Sup. ImageNet$_{1k}$	GT	38.5	39.3	81.9	83.8	49.4	49.6
	ResNet-18[39]	11.5M/1.8G	Sup. ImageNet$_{50}$	GT	31.3	32.1	70.2	71.0	44.1	44.4
	ResNet-18 [39]	11.5M/1.8G	MoCo. ImageNet-S_{50}	-	17.7	18.1	43.7	45.7	39.7	40.0
	ResNet-18 [39]	11.5M/1.8G	SwAV. ImageNet-S_{50}	-	19.0	19.7	50.0	49.1	38.6	40.8
SEAM[103] +AdvCAM [56]	ResNet-18 [39]	11.3M/1.9G	Sup. ImageNet$_{1k}$	GT	46.9	46.2	90.9	90.4	58.4	57.5
	ResNet-18 [39]	11.3M/1.9G	Sup. ImageNet$_{50}$	GT	36.9	37.6	81.2	81.5	49.2	49.6
	ResNet-18 [39]	11.3M/1.9G	MoCo. ImageNet-S_{50}	-	19.2	19.5	45.1	46.7	46.8	47.3
	ResNet-18 [39]	11.3M/1.9G	SwAV. ImageNet-S_{50}	-	23.7	23.3	54.6	53.5	44.2	43.9

ResNet-38 [108] with a standard ResNet-18 [39], the test mIoU drops from 49.6% to 44.5%. In the unsupervised setting, replacing ResNet-18 with our modified DeepLabV3+ with an acceptable extra computational cost (detailed in Sec. 4) improves the test mIoU from 22.3% to 28.7%, indicates that designing a more efficient architecture for the LUSS task is essential.

Image-level GT labels. One essential difference between WSSS and LUSS tasks is that WSSS requires image-level GT labels. Class activation maps [129, 86], commonly treated as the initial segment regions, usually cover the most discriminative small area of objects. Numerous WSSS methods heavily rely on GT labels to extend the CAM region to the whole object and remove wrong region [91] by image erasing [89, 59, 104, 42], regions growing [52, 45, 101, 51], stochastic feature selection [123, 55], gradients manipulation [56]. To analyze the effect of GT labels in the LUSS task on WSSS methods, we apply the recent work AdvCAM [56] to SEAM [103]. AdvCAM anti-adversarially refines the CAM results by perturbing the images along pixel gradients according to GT labels. Tab. 4 shows AdvCAM using GT labels improves the baseline of ImageNet$_{1k}$ and ImageNet$_{50}$ pre-trained models with 1.7% and 1.8% in test mIoU. However, when using clustered labels and MoCo pre-trained model, the performance gain is only 0.4%. Using the SwAV pre-trained model with better image-level accuracy, AdvCAM improves the model performance by 1.0%. Similarly, the SEAM and SC-CAM with GT labels outperform the unsupervised settings with a large margin. Thus, the GT label reliance makes WSSS methods cannot be directly shipped to the LUSS task due to the image-level GT label absence.

We note it is possible to use other alternative WSSS modules *e.g.,* affinity prediction [2, 1, 30], region separation [88, 29], boundary refinement [1, 15], joint learning [118], and sub-category exploration [12] to further improve LUSS.

3.4 Challenges of LUSS

We summarize the main challenges of the LUSS tasks: 1) The model should learn category-related representations without image-level label supervision. 2) Extracting a semantic segmentation mask requires the model to learn the shape representations. 3) With learned representations, the model should assign self-learned labels to each pixel in the image. 4) The large-scale training data helps to learn rich representations in an unsupervised learning manner but inevitably causes a large amount of training cost, which requires improving the representation learning efficiency.

4 A Baseline Method for LUSS with Shape Representation Learning

Shape representations play a key role in the LUSS task. As most approaches for unsupervised representation focus on image-level models, less effort has been made to exploit pixel-level visual information. In this section, we present a baseline method for the LUSS task to improve model representation ability regarding shape-related features, *i.e.,* efficient architectures, pixel attention, unsupervised shape prior knowledge, pixel-to-pixel alignment, and image-to-pixel alignment. We utilize SwAV [11] as the baseline method as its clustering strategy provides good category representations.

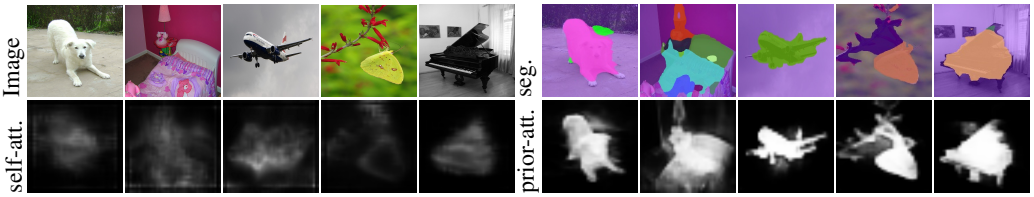


Figure 3: Visualization of pixel attention with/without unsupervised prior knowledge. Pixel attention highlights the semantic regions (self-att) during representation learning, and unsupervised prior knowledge further enhances the attention effect (prior-att).

Table 5: Ablation of the proposed shape and category representation learning method for LUSS. The efficient architecture, pixel attention with/without shape prior, image-to-pixel alignment, and pixel-to-pixel alignment enhance the shape and category representation.

ImageNet-S_{50}	mIoU		NMI	Img-Acc		F_β		view.cos↑	cat.cos↑
	train	val	test	val	test	val	test	val	val
SwAV[11]	21.7	20.8	63.3	57.4	57.7	51.0	50.1	.8355	.7455
+ DeepLabV3+*	28.1	27.4	72.0	67.4	69.1	49.5	48.4	.8585	.7535
+ Pixel attention	31.0	30.4	72.6	68.9	70.5	54.5	53.1	.8597	.7623
+ Unsupervised shape prior	35.2	35.3	73.6	67.4	68.4	60.8	59.4	.9064	.8300
+ Image2Pixel	38.0	38.0	74.9	69.2	69.3	60.3	59.1	.9862	.9708
+ Image2Pixel + Pixel2Pixel	38.7	39.4	75.1	72.4	72.5	60.2	59.0	.9886	.9758

Efficient architecture for shape representation learning. Due to the small size of many pixel-level datasets, existing methods mostly achieve better performance at the cost of large network complexity[87, 127, 128], *e.g.,* DeeplabV3+ [14] is about 8× larger than ResNet-50 [39] in computational cost (32.4G vs. 4.1G). For LUSS, the conflict between high-performance but costly segmentation models and the extremely large-scale training data requires designing effective segmentation models. As a baseline, we present a modified low-cost DeepLabV3+ with only 9.5G MACC, denoting DeepLabV3+*, by down-sampling the high-resolution feature maps in the deep stages as detailed in the supplementary. Tab. 5 shows adding an effective segmentation model improves the test mIoU from 20.8% to 27.4%, thereby demonstrating the value of designing effective segmentation models for the LUSS task.

Pixel attention for pixel-level representation discrimination. Given the feature $\mathbf{z}$ predicted by the model, SwAV learns image-level representations by clustering the image-level embeddings that are obtained with $\mathbf{M}\left[\mathbf{Pooling}(\mathbf{z})\right]$, where $\mathbf{Pooling}$ is the global average pooling operation to pool the feature in the spatial dimension, and $\mathbf{M}$ is the projection layer composed of the two fully-connected layers and activation layers. However, SwAV does not take shape information into account. To enforce SwAV with shape representation ability, we add a simple yet effective pixel attention scheme that implicitly learns the region of interest used for SwAV representation learning. The pixel attention module, $\mathbf{c}(\mathbf{z}) = \mathbf{Sigmoid}\left[\mathbf{Conv1} \times \mathbf{1}(\mathbf{z})\right]$, is composed of a conv 1×1 layer with 1 output channel and a sigmoid function. As shown in Fig. 5, we multiply the pixel attention to feature $\mathbf{z}$ and obtain the pixel attention enhanced image-level embedding $\hat{\mathbf{v}} = \mathbf{M}\left\{\mathbf{Pooling}\left[\mathbf{c}(\mathbf{z}) \cdot \mathbf{z}\right]\right\}$. Same as the SwAV, $\hat{\mathbf{v}}$ is used to calculate the clustering loss in SwAV. Fig. 3 shows that the pixel attention self-learns to highlight the semantic regions instead of the whole image. The test mIoU is improved from 27.4% to 30.4%, and notably, the F_β achieves 4.7% improvement. The self-learned pixel attention roughly shows the semantic regions. To guide the representation learning with more clear attention, we use the unsupervised generated prior knowledge, *i.e.,* foreground saliency maps [20], to supervise the pixel attention. Fig. 3 shows the pixel attention under the prior knowledge supervision focuses more on the semantic-related objects. The shape representation is further improved as the F_β is improved from 53.1% to 59.4%, and the test mIoU has improved 4.9% accordingly. We observe the shape representation learning improves the shape quality (59.4% in F_β) based on the unsupervised foreground saliency maps with the F_β of 56.4%.

Image-to-pixel and pixel-to-pixel representation alignment. Pixel-level representation requires enhancing the consistency among pixels, *i.e.,* pixels within the same category or from the same image position of different views should have consistent representations. Therefore, we propose the image-to-pixel alignment to enhance the representation consistency between pixel-level embeddings and image-level embedding. The pixel-to-pixel alignment is also proposed to align the pixels from the same image position of different views. As shown in Fig. 5, given the feature pair $(\mathbf{z_1}, \mathbf{z_2})$ predicted from two views of the image, we obtain the pixel attention enhanced image-level embeddings

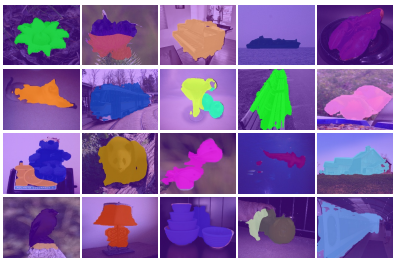


Figure 4: Visualization of unsupervised semantic segmentation results.

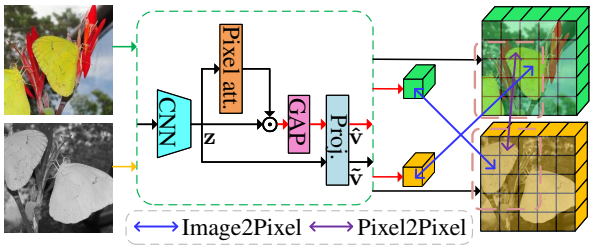


Figure 5: Illustration of pixel attention enhanced projection, Pixel2Pixel, and Pixel2Image alignment.

Table 6: Test mIoU results under different object scales.

ImageNet-S_{300}	test	mIoU S.	M.S.	M.L.	L.
SimCLR[16]	11.4	1.5	9.5	12.6	9.3
MoCov2[18, 37]	18.0	3.0	15.3	19.6	15.1
PCL[57]	15.5	2.6	13.6	17.0	12.3
SwAV[11]	23.4	3.5	19.1	25.1	21.5
BYOL[48]	14.3	1.8	10.8	15.7	13.3
Ours	26.4	3.8	20.5	30.2	24.5

Table 7: Transfer learning results (mIoU) from LUSS pre-trained models to the downstream task, *i.e.,* semantic segmentation on the PASCAL VOC dataset using the original DeeplabV3+ [14].

Pre-train	Dataset	Sup.	MoCo	SwAV	Ours
ImageNet-S_{50}	PASCAL VOC	61.5	62.1	59.0	65.1
ImageNet-S_{300}	PASCAL VOC	74.4	72.4	74.3	75.4

$(\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2)$. Then, we extract the overlapped regions $(\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2)$ and get the pixel-level embedding pairs $(\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2)$ by the projection $\tilde{\mathbf{v}} = \mathbf{M}(\tilde{\mathbf{z}})$. We show in the supplementary that learning with pixel-level embeddings worsens the performance, indicating pixel-level embeddings may contain more noises. To alleviate the effect of noise to feature projection, the gradients to $\mathbf{M}$ are detached when obtaining the pixel-level embeddings. The image-to-pixel alignment is achieved by aligning the pixel-level embeddings to the image-level embeddings: $L_{I2P} = L_s(\hat{\mathbf{v}}_1, \tilde{\mathbf{v}}_2) + L_s(\hat{\mathbf{v}}_2, \tilde{\mathbf{v}}_1)$, where L_s is the online clustering loss in SwAV. 5 Tab. 5 shows that image-to-pixel alignment improves the test mIoU from 35.3% to 38.0%. The mean cosine similarity between the pixel embeddings and the averaged embeddings over all pixels of certain categories becomes much larger (from 0.83 to 0.97) when using image-to-pixel alignment, showing that image-to-pixel alignment enforces the similarity of pixels within a category. Furthermore, we utilize the pixel-to-pixel alignment to align the overlapped pixel-level embeddings from two views: $L_{I2I} = L_s(\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2) + L_s(\tilde{\mathbf{v}}_2, \tilde{\mathbf{v}}_1)$, forming a more robust pixel-level representation across different views. The pixel-to-pixel alignment improves the test mIoU to 39.4%. The mean cosine similarity among pixels between two views is further enhanced, as shown in Tab. 5.

Comparison with existing methods. As shown in Tab. 2 and Tab. 3, our shape-enhanced representation learning method outperforms existing methods in mIoU on ImageNet-S_{50}, ImageNet-S_{300}, and ImageNet-Sdatasets. Notably, the F_β is improved with a large margin as we enhance the shape representation ability. The visualization shown in Fig. 4 proving that unsupervised semantic segmentation with the large-scale dataset is achievable. In addition, our method performs favorably against the unsupervised method [98] designed for the small-scale dataset. We also evaluate the testing mIoU under different object scales, as shown in Tab. 6. The performance on large objects is much better than on small objects, indicating that small objects need a model with a more precise pixel-level representation and segmentation ability.

4.1 Transfer Learning on the Pixel-level Downstream Task

The LUSS task can promote the shape-related representation ability of pre-training and benefit the downstream pixel-level task. As shown in Tab. 7, we transfer the LUSS pre-trained models to the semantic segmentation task on the PASCAL VOC dataset using the original DeeplabV3+ [14]. For a fair comparison, only the ResNet backbone is loaded with pre-trained weights. Our method surpasses the supervised pre-trained model with 3.6% and 1.0% in mIoU when pre-trained on ImageNet-S_{50} and ImageNet-S_{300}. Also, our shape representation enhanced method outperforms other unsupervised representation learning methods, *e.g.,* MoCo and SwAV, on the downstream task.

5 Conclusions

In this work, we propose a new problem of large-scale unsupervised semantic segmentation to facilitate semantic segmentation in real-world environments with a large diversity and large-scale

data. We present a benchmark for LUSS to provide large-scale data with high diversity, a clear task objective, and sufficient evaluation. We present one basic pipeline of LUSS to assign labels to pixels with category/shape representations learned from the large-scale data without human-annotation supervision. In addition, we benchmark and analyze unsupervised representation learning methods and weakly supervised semantic segmentation methods on the LUSS benchmark, and summarize the challenges and possible directions of LUSS. Furthermore, we introduce a baseline LUSS method with an enhanced ability for pixel-level shape representation learning, and reveal the potential of LUSS to pixel-level downstream tasks, *e.g.,* semantic segmentation. We mainly use ImageNet-$S_{50/300}$ in this work due to the limited computational budget. In our future work, we will benchmark more recent proposed unsupervised works [82, 97, 19, 117, 44, 102, 95, 81] on the largest-scale full ImageNet-S dataset. In addition, we will annotate a part of training images of the ImageNet-S dataset to serve the semi-supervised learning tasks.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2209–2218, 2019. 6, 7

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4981–4990, 2018. 6, 7

[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 6

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(12):2481–2495, 2017. 1, 2

[5] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 2, 3, 4

[6] Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *International Conference on Machine Learning (ICML)*, pages 517–526. PMLR, 2017. 5

[7] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[8] Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric instance classification for unsupervised visual feature learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 6

[9] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 5, 6

[10] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2959–2968, 2019. 6

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3, 5, 6, 7, 8, 9

[12] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 6, 7

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):834–848, 2017. 1, 2

[14] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 7, 8, 9

[15] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *European Conference on Computer Vision (ECCV)*, pages 347–362. Springer, 2020. 6, 7

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020. 1, 2, 3, 5, 6, 9

[17] Wanli Chen, Xinge Zhu, Ruoqi Sun, Junjun He, Ruiyu Li, Xiaoyong Shen, and Bei Yu. Tensor low-rank reconstruction for semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[18] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 5, 9

[19] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 10

[20] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015. 8

[21] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 539–546, 2005. 6

[22] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 8765–8775, 2020. 6

[23] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3

[24] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015. 5

[25] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *International Conference on Learning Representations (ICLR)*, 2017. 5

[26] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 5

[27] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 6

[28] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136, 2015. 1, 3, 4, 6

[29] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[30] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 6, 7

[31] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[32] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision (IJCV)*, 59(2):167–181, 2004. 2

[33] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–1, 2020. 3

[34] Shang-Hua Gao, Qi Han, Duo Li, Pai Peng, Ming-Ming Cheng, and Pai Peng. Representative batch normalization with feature calibration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[35] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 5, 6

[36] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1735–1742, 2006. 6

[37] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 5, 6, 9

[38] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 2

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 7, 8

[40] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning (ICML)*, pages 4182–4192. PMLR, 2020. 5

[41] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations (ICLR)*, 2019. 6

[42] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 7

[43] Hanzhe Hu, Deyi Ji, Weihao Gan, Shuai Bai, Wei Wu, and Junjie Yan. Class-wise dynamic graph convolution for semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[44] Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 10

[45] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7014–7023, 2018. 7

[46] 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 *IEEE International Conference on Computer Vision (ICCV)*, pages 7334–7344, 2019. 2, 3

[47] 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. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 5

[48] Grill Jean-Bastien, Strub Florian, Altché Florent, Tallec Corentin, Pierre Richemond H., Buchatskaya Elena, Doersch Carl, Bernardo Pires Avila, Zhaohan Guo Daniel, Mohammad Azar Gheshlaghi, Piot Bilal, Kavukcuoglu Koray, Munos Rémi, and Valko Michal. Bootstrap your own latent - a new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 5, 6, 9

[49] Simon Jenni and Paolo Favaro. Self-supervised feature learning by learning to spot artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2733–2742, 2018. 5

[50] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9865–9874, 2019. 6

[51] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2070–2079, 2019. 7

[52] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 695–711. Springer, 2016. 6, 7

[53] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4

[54] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6874–6883, 2017. 5

[55] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[56] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 6, 7

[57] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C.H. Hoi. Prototypical contrastive learning of unsupervised representations. *International Conference on Learning Representations (ICLR)*, 2021. 5, 6, 9

[58] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *International Conference on Learning Representations (ICLR)*, 2021. 5, 6

[59] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9215–9223, 2018. 6, 7

[60] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[61] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[62] Yanwei Li, Lin Song, Yukang Chen, Zeming Li, Xiangyu Zhang, Xingang Wang, and Jian Sun. Learning dynamic routing for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[63] Peiwen Lin, Peng Sun, Guangliang Cheng, Sirui Xie, Xi Li, and Jianping Shi. Graph-guided architecture search for real-time semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[64] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing via label transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(12):2368–2382, 2011. 2

[65] Jianbo Liu, Junjun He, Jimmy S. Ren, Yu Qiao, and Hongsheng Li. Learning to predict context-adaptive convolution for semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[66] Jianbo Liu, Junjun He, Jiawei Zhang, Jimmy S. Ren, and Hongsheng Li. Efficientfcn: Holistically-guided decoding for semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[67] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 1, 2

[68] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *European Conference on Computer Vision (ECCV)*, pages 181–196, 2018. 1

[69] Tomasz Malisiewicz and Alexei A Efros. Recognition by association via learning per-exemplar distances. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2

[70] David R Martin, Charless C Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(5):530–549, 2004. 2

[71] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6707–6717, 2020. 5, 6

[72] T Nathan Mundhenk, Daniel Ho, and Barry Y Chen. Improvements to context based self-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9339–9348, 2018. 5

[73] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, pages 69–84. Springer, 2016. 5

[74] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5898–5906, 2017. 5

[75] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9359–9367, 2018. 5

[76] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5, 6

[77] Yassine Ouali, Céline Hudelot, and Myriam Tami. Autoregressive unsupervised image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 142–158. Springer, 2020. 2, 3

[78] Massimiliano Patacchiola and Amos Storkey. Self-supervised relational reasoning for representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 6

[79] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016. 5

[80] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 762–771, 2018. 5

[81] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations (ICLR)*, 2021. 10

[82] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 10

[83] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 234–241. Springer, 2015. 1, 2

[84] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 2, 3

[85] Bryan C Russell, Alexei Efros, Josef Sivic, William T Freeman, and Andrew Zisserman. Segmenting scenes by matching image composites. In *European Conference on Computer Vision (ECCV)*, 2009. 2

[86] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 7

[87] Dingguo Shen, Yuanfeng Ji, Ping Li, Yi Wang, and Di Lin. Ranet: Region attention network for semantic segmentation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 13927–13938. Curran Associates, Inc., 2020. 8

[88] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5208–5217, 2019. 6, 7

[89] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3544–3553, 2017. 7

[90] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017. 1

[91] Guolei Sun, Salman Khan, Wen Li, Hisham Cholakkal, Fahad Khan, and Luc Van Gool. Fixing localization errors to improve image classification. *European Conference on Computer Vision (ECCV)*, 2020. 7

[92] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2020. 6

[93] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 5, 6

[94] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 5, 6

[95] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning (ICML)*. PMLR, 2021. 10

[96] Joseph Tighe and Svetlana Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *European Conference on Computer Vision (ECCV)*, pages 352–365. Springer, 2010. 2

[97] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations (ICLR)*, 2021. 10

[98] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. *arxiv preprint arxiv:2102.06191*, 2021. 2, 3, 5, 9

[99] Feng Wang, Huaping Liu, Di Guo, and Fuchun Sun. Unsupervised representation learning by invariance propagation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 6

[100] Li Wang, Dong Li, Yousong Zhu, Lu Tian, and Yi Shan. Cross-dataset collaborative learning for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[101] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1354–1362, 2018. 7

[102] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 10

[103] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 6, 7

[104] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1568–1576, 2017. 7

[105] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(11):2314–2320, 2016. 6

[106] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7268–7277, 2018. 6

[107] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742, 2018. 5, 6

[108] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 6, 7

[109] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017. 3

[110] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 5, 6

[111] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6509–6518, 2020. 6

[112] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6210–6219, 2019. 6

[113] Asano YM., Rupprecht C., and Vedaldi A. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020. 5, 6

[114] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2636–2645, 2020. 1

[115] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*, 2015. 1, 2

[116] Yuhui Yuan, Xiaokang Chen, Xilin Chen, and Jingdong Wang. Segmentation transformer: Object-contextual representations for semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2021. 1

[117] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 10

[118] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 7

[119] Xiaohang Zhan, Ziwei Liu, Ping Luo, Xiaoou Tang, and Chen Loy. Mix-and-match tuning for self-supervised semantic segmentation. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, 2018. 2, 3

[120] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5, 6

[121] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2547–2555, 2019. 5

[122] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, pages 649–666. Springer, 2016. 5

[123] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1325–1334, 2018. 7

[124] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017. 1, 2

[125] Shuai Zhao, Yang Wang, Zheng Yang, and Deng Cai. Region mutual information loss for semantic segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1

[126] Mingmin Zhen, Jinglu Wang, Lei Zhou, Shiwei Li, Tianwei Shen, Jiaxiang Shang, Tian Fang, and Long Quan. Joint semantic segmentation and boundary detection using iterative pyramid contexts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[127] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8

[128] Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, and Alexander Wong. Squeeze-and-attention networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[129] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 5, 7

[130] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision (IJCV)*, 2018. 1, 3, 4

[131] Lanyun Zhu, Deyi Ji, Shiping Zhu, Weihao Gan, Wei Wu, and Junjie Yan. Learning statistical texture for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[132] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6002–6012, 2019. 6