Self-supervised Pre-training for Semantic Segmentation in an Indoor Scene

Sulabh Shrestha, Yimeng Li, Jana Košecka
George Mason University

Abstract

The ability to endow 3D models of indoor scenes with semantic information is an integral part of embodied agents performing tasks such as target-driven navigation, object search, and object rearrangement. We propose RegConsist, a method for environment-specific self-supervised pre-training of a semantic segmentation model that exploits the ability of the mobile robot to move and register multiple views in the environment. Using the spatial and temporal consistency cues used for pixel association and a novel efficient region matching approach, we present a variant of contrastive learning to train a DCNN model for predicting semantic segmentation from RGB views in the environment where the agent operates. The approach introduces different strategies for sampling individual pixel pairs from associated regions in overlapping views and an efficient region association method and yields a more robust and better-performing pre-trained model when fine-tuned with a low amount of labeled data. RegConsist outperforms other self-supervised methods that pre-train on single view images and achieves competitive performance with models which are pre-trained for exactly the same task but on a different and larger dataset. We also perform various ablation studies to analyze and demonstrate the efficacy of our proposed method.

1. Introduction

Semantic segmentation has been used extensively for both semantic mapping [5] and also as input representation for training policies for embodied agents (e.g., policies for target driven or point goal navigation) that rely on visual perception [6, 16]. Training semantic segmentation model for a particular environment requires a large amount of per-pixel annotations [51] that is very costly and laborious to obtain.

In this work, we explore the ability of the agent to move around and capture large amounts of visual data, estimate ego-motion, and establish correspondences between multiple views of the same scene. The agent is free to gather information, possibly with informative exploration strategy [6,7,28], viewing the objects from vastly different viewpoints, under environment specific occlusions. We propose to use these corresponding views for self-supervised contrastive pre-training of an environment-specific semantic segmentation model.

We assume that within a single traversal path, the environment remains static, and with the availability of motion and 3D structure estimates it is possible to associate pixels and image regions across widely separated views. Similarly, regions can be computed using various class agnostic segmentation methods such as the efficient graph-based segmentation [15]. The proposed approach *RegConsist* exploits point and region correspondences between multiple views for generating positive examples for contrastive learning framework. We also develop an efficient region matching approach for computing pixel-IoU of class agnostic regions between two overlapping views in linear computation time. The efficacy of our method is shown on Replica [39], AVD [2] and HM3D [36] datasets both qualitatively and quantitatively while using as low as 5% of the annotated data. We perform extensive ablation studies of our method's performance and compare it with alternative single view based self-supervised pre-training methods and models trained on relevant labeled datasets.

2. Related Work

In the past, supervised learning had been the dominant method for pre-training representations useful for downstream tasks in computer vision. Recently, self-supervised learning has emerged as a superior alternative requiring no human-annotated labels. The existing methods typically use various pretext tasks on unlabeled data including masked image modeling [20], object mask prediction [25], instance discrimination [44] and others. These tasks provide objectives to embed semantically similar inputs closer in the embedding space using contrastive learning [4, 11, 18, 47].

Instance discrimination [44] pretext task introduced for contrastive learning considers each image as a separate class. Various enhancements for this baseline method have been introduced [9–11, 17, 21]. More recent methods employ redundancy reduction [47] and covariance regularization [4] which remove the need for large batches and asym-

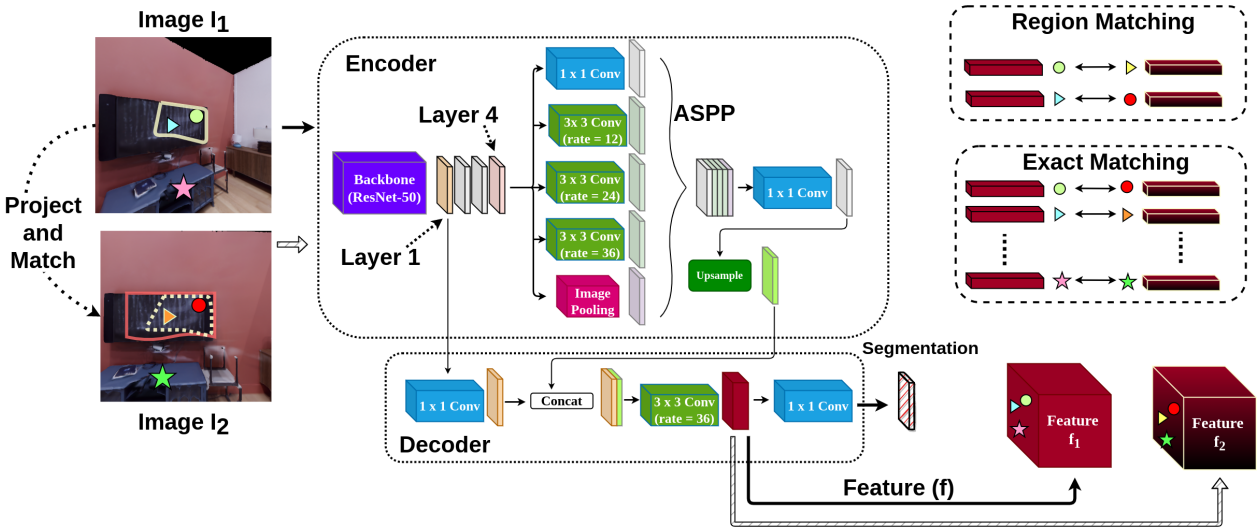

Figure 1. Our proposed method. The segmentation model (DeepLabV3+ [8]) separately processes two views that capture the same part of the environment. Positive pairs are sampled across the two views. Using temporal consistency, we match corresponding points (pixels) from the two views. Regions are estimated for each view separately (yellow in I_1 and red in I_2) using an unsupervised segmentation method [15]. Using spatial consistency, highly overlapping regions across the views (in I_2, the red region with dotted yellow region projected from I_1) are paired. Positive pixel pairs are also sampled from matched regions. Features from paired points and regions are aligned using Barlow Twins loss [47]. Best viewed in color.

metric models. Temporal constraints at the image/frame level have been explored in VideoMoCo [33] that builds upon MoCo [21] to perform learning of video representations for action recognition. CVRL [35] and SCVRL [13] employ 3D ResNets [19] for learning video representations using InfoNCE loss [42] using pairs of temporally close clips from the same video made spatially consistent by using same augmentation in a clip.

Problems such as object detection and semantic segmentation require disambiguation of features at a finer level for bounding box and per pixel predictions respectively calling for different strategies for selecting training examples. PixPro [45] follows SimSiam [11] like training but at the pixel level, where pixels with features that have low cosine distance from each other are chosen as positive pairs. Zhang et al. [48] sample positive pixel pairs within regions obtained by k-means clustering of the initial features; however, to perform well, their model requires clustering in supervised feature representation which is not feasible with a lack of labeled data. PLRC [3] circumvent the need to find regions by dividing the image into a fixed grid where each square grid cell is considered a separate region. DetCon [24] and SoCo [43] both learn over regions obtained through unsupervised bottom-up segmentation methods such as [15, 41] to pre-train object detection models. Our approach also considers pixel and region-level supervision but with regions associated with overlapping views inside the indoor scene.

In settings where motion or multiple views are available, the ability to associate and track objects between multiple views has been used as a source of supervision. Mitash et al. [32] train detectors in simulation and improve them on real unlabeled data, where scenes are observed from different viewpoints; SSOD [34], use contrastive learning followed by clustering on object proposals for object discovery and subsequent fine-tuning of the object detector trained on COCO. We instead focus on semantic segmentation and learning in a specific environment without the need to fine-tune and overcome biases of existing models.

Alternatively, the problem can be tackled using a model trained in a particular domain (say indoor environments), followed by domain adaptation [26]. Since different instances of the environments vary in the encountered labels only the shared subset of semantic labels can be transferred. The majority of unsupervised or self-supervised domain adaptation approaches have been tested in the autonomous driving domain, with a more limited and shared number of classes, using single view approaches, exhibiting smaller view-point variations and less challenging occlusions [40, 46, 49]

Our work is most closely related to the efforts of self-supervised learning for object detection [14, 34] that also uses multiple views and their association to guide the training. We extend these ideas to dense pixel-level prediction tasks such as semantic segmentation however, unlike them, we use an unsupervised segmentation [15] method, so no training is required to obtain the regions. We demonstrate

the approach in challenging indoor scenes with large variations in appearance due to viewpoint, occlusions, and lighting.

3. Method

We assume the availability of multiple registered images and their associated depth maps captured from different locations in a specific indoor environment with significant overlap. This can be achieved with RGB-D sensors, 3D structure and motion estimation techniques [31] or suitable SLAM approach [5]. We demonstrate how these images can be used for self-supervised pre-training of a semantic segmentation model. The goal is to make the model perform well inside *this specific* indoors environment with limited annotations. To instantiate a self-supervised learning approach for semantic segmentation we propose *RegConsist* (*Reg*ion *Consist*ency), a method for temporal and spatial alignment of pixels and 2D regions across overlapping views that forms a basic building block for generation of positive training examples for contrastive learning. We use the (non-learning) efficient graph-based segmentation method [15] to obtain the regions but any class-agnostic segmentation approach can be used.

3.1. Spatial and Temporal Consistency

Let I_1 and I_2 be a pair of images taken inside the fixed indoor environment. Assuming the availability of known intrinsic and extrinsic camera parameters and depth, we can associate the pixels in the overlapping views of the same scene using (1).

$$T_{1\to 2}(I_1) = \{K(T_2{}^{-1}(T_1(K^{-1}(\mathbf{X})) \quad \forall \mathbf{X} \in I_1\} \quad (1)$$

where, K is the intrinsic parameters of the camera, $T_i = [R_i|t_i]$ is the camera pose for the image I_i having rotation R_i and translation t_i with respect to a fixed coordinate system. $\mathbf{X}$ represents the 3D coordinate of the 2D pixel $\mathbf{x}$ in the image along with its known depth. Temporal consistency refers to the fact that corresponding pixels $\mathbf{x}_1^p$ and $\mathbf{x}_2^p$ that are projections of the same 3D point will have the same semantic label. Let $S_t = \{(\mathbf{x}_1^p, \mathbf{x}_2^p)\}$ be the set of all such positive pairs. The learning objective should enforce their features to be aligned across the views. The positive corresponding pixel pairs obtained from neighboring views, while easier to match, look quite similar and do not provide a strong signal for training the model. We instead match pixels belonging to corresponding *regions* from image pairs that are further apart yet have overlapping views to get more varied pairs. Regions can be obtained using unsupervised segmentation methods such as the efficient graph-based segmentation method [15] which we use, similar to DetCon [24].

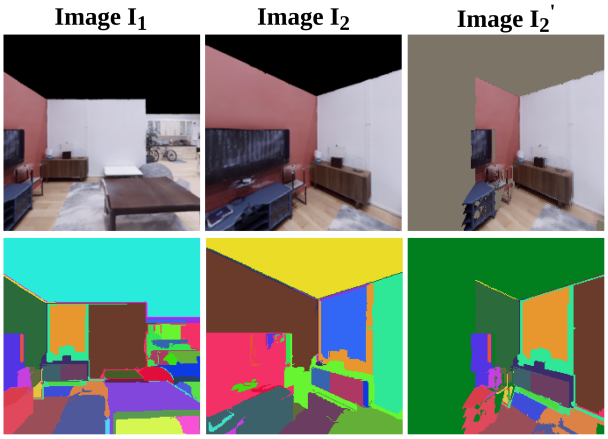

Figure 2. Example of projection from Replica [39] dataset. RGB images on top row and their unsupervised segments on the bottom row. I_2' is obtained by projecting I_1 to I_2. Regions in I_2 and I_2' do not perfectly align, so IoU calculation is required to choose highly overlapping regions.

3.2. Region Matching

Here, we overload I to mean *both* the RGB image and its bottom-up segmentation with regions having unique class agnostic region *labels*. We project image I_1 to image I_2 to get the projected image I_2' using equation 1. I_2' contains same region labels as I_1 but projected to the coordinate frame of I_2. Since regions are independently computed in each image, some regions/pixels in I_2' are not perfectly aligned with those in I_2. For example, in Figure 1, the red boundary region in image I_2 best aligns with the yellow dotted region projected from I_1 but they are not perfectly aligned. Similar examples can be found in Figure 2. We find the intersection over union (pixel IoU) between regions in I_2 and in I_2' and consider those above a threshold IoU_r a match. The brute force approach to calculate the pixel IoU is to iterate over each region from regions $\{r_1^i\}_{i=1}^{R_1} \in I_1, I_2'$ and match it to regions $\{r_2^j\}_{j=1}^{R_2} \in I_2$. This naive approach takes $O(R_1 R_2 N^2)$ time because finding a mask for each of the R_1 and R_2 regions takes $O(N)$ time each, where N is the number of pixels in the two images. This is extremely slow to calculate in each iteration during pre-training and hampers training speed. Therefore we devise a new algorithm to calculate the *class agnostic* pixel IoU in $O(N)$ time using a pairing function, π. A pairing function $\pi : \mathcal{Z}^* \times \mathcal{Z}^* \to \mathcal{Z}^*$ is a *reversible bijective* function that maps non-negative integers (x, y) to a unique integer z. We use the Cantor pairing function π and its inverse π^{-1} that can be computed in $O(1)$ time for an input number-pair and O(N) for the image-pair. The specific form of the function, toy example and detailed pseudo-code can be found in the supplementary material.

3.3. Pixel Pair Sampling and Matching

Given the region matching approach described above, we can now select the regions with large IoU threshold and sample their pixels as positive training examples. While using all positive pairs from Section 3.1 is possible, it is not efficient. So, we sub-sample pixel pairs in each batch. The *first* step is to sample a pixel $\mathbf{x}_1^p$ from I_1. Then, the *second* step is to match it to a viable positive $\mathbf{x}_2^p$ from I_2.

Sampling. In *random* sampling, the pixel $\mathbf{x}_1^p$ is sampled uniformly from the whole image I_1. In *balanced* sampling the pixel $\mathbf{x}_1^p$ is sampled uniformly from each region r_1^p in I_1. This guarantees that each region has the same number of pixels sampled from it unlike in random sampling where it is proportional to the size of the regions.

Matching. Once the first pixel in the pair has been sampled from a view I_1, we need to match it with a positive pixel from I_2. In *exact* matching, we match the $\mathbf{x}_1^p$ with pixel $\mathbf{x}_2^p$ which is the exact correspondence that satisfies Equation (1). This is the same as using only temporal consistency as explained in Section 3.1. To get variability between the pixels in the positive-pair, in *region* matching, we match $\mathbf{x}_1^p \in r_1^p$ with $\tilde{\mathbf{x}}_2^p$ sampled uniformly from matched region r_2^p. This is the same as using spatial consistency as explained in Section 3.1.

3.4. Losses

We use Barlow Twins loss [47] for pre-training the models because of its simplicity, memory efficiency, and demonstrated effectiveness even with relatively smaller batch sizes compared to other approaches. Let $F = [f^{(1)}, f^{(2)}, ..., f^{(B)}]$ be the features which need to be aligned with features $G = [g^{(1)}, g^{(2)}, ..., g^{(B)}]$ element-wise, *i.e* each pair $(f^{(b)}, g^{(b)})$ is a positive pair and B is the batch size. The Barlow Twins loss is then given by equation (2).

$$\mathcal{L}_{barlow} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} {\mathcal{C}_{ij}}^2 \qquad (2)$$

where $\mathcal{C}$ is the cross-correlation matrix computed between F and G and each of its elements $\mathcal{C}_{ij}$ is cross-correlation between $f(i)$ and $g(j)$. The first term in the loss aligns each input feature-pairs $(f^{(b)}, g^{(b)})$ while the second term minimizes the redundancy between each dimension of the features. We use three types of positive pairs and calculate each of their losses separately. In pixel loss $\mathcal{L}_{pix}$, we use B_{pix} batch of pixel-pairs obtained via temporal consistency. This is same as random sampling with exact matching as explained in section 3.3. In region loss $\mathcal{L}_{reg}$, the loss is calculated over B_{reg} batch of pixel-pairs from matched regions. Finally, in pool loss $\mathcal{L}_{pool}$, in order to align features of regions as a whole, we adopt masked feature pooling of regions similar to DetCon [24] to match B_{pool} region pairs.

The total loss $\mathcal{L}$ is obtained by summing all three losses

$$\mathcal{L} = \mathcal{L}_{pix} + \mathcal{L}_{reg} + \mathcal{L}_{pool} \qquad (3)$$

The $\mathcal{L}_{reg}$ loss aligns pixels from overlapping regions of varied views while $\mathcal{L}_{pool}$ loss helps to align the features of the overlapping regions as a whole. The $\mathcal{L}_{pix}$ loss works on exact correspondences so, it helps to mitigate the noise from matching regions of possible different categories in the other two losses. When using selected labeled images in the fine-tuning phase, we use focal loss [29] to train the models.

4. Experiments

We perform our experiments on Replica dataset [39], Active Vision Dataset (AVD) [2] and Habitat-Matterport 3D (HM3D) [36]. AVD is a real-world dataset that consists of scenes from different apartments. Each scene contains images taken by a robot in a grid-like manner and a few of the images are annotated. We use *Home_006_1* which contains 2412 images among which 43 images are annotated. Replica is a photo-realistic dataset that consists of indoor environments. HM3D is also a photo-realistic dataset similar to Replica but with more 3D reconstruction artifacts. The datasets contain ground-truth depth as well as intrinsic and extrinsic parameters of the camera. Since AVD and HM3D contain more noise, we start with the Replica dataset to validate our approach and demonstrate that it also works for the other two datasets. Unless otherwise stated, we experiment on *frl_apartment_1* scene from Replica and *00820-mL8ThkuaVTM* scene from HM3D. We use the Habitat simulator [38] to move the agent in the environment and generate views similar to AVD. Exact details on the data generation can be found in the *supplemental materials*. Examples of images can be seen in Figure 3.

We heuristically sample informative pairs of images, by considering uniformly sampled views on a grid and selecting neighboring views with varying degrees of overlap as characterized by the Intersection over Union (IoU) measure. The view pairs with IoU in the range of $[IoU_l, IoU_h]$ are selected for training. This sampling process reduces computation during training as it needs only be done once per environment for all the experiments. More details about view generation and view-pair selection can be found in the *supplemental materials*. To compare our pre-training method with the model pre-trained on the semantic segmentation dataset, we use ADE20K dataset [51, 52]. The class labels from Replica and AVD are both separately mapped to those in the ADE20K dataset resulting in 52 and 66 classes respectively. We discard classes that do not have an unambiguous overlap. The exact mappings can be found in the *supplemental materials*. For HM3D, its default classes are used.

Supervision	Sampling	Matching	mIoU
gt-labels	random	region	48.6
gt-labels	random	exact	60.5
gt-labels	balanced	exact	61.2
gt-labels	balanced	region	**73.4**

Table 1. Supervised Pre-training on Replica [39] dataset. We pre-train the model assuming regions overlap with unique ground truth class labels to get an upper bound of our proposed approach.

4.1. Implementation Details

We use a DeeplabV3+ [8] with ResNet50 [23] backbones as our segmentation model as shown in Figure 1 and modify it by adding another Conv2D(256,256,1) layer before the final layer similar to [1]. All weights are randomly initialized by default.

Pre-training. We use a batch size of 16 image pairs. In each batch, we sample $B_{pix} = 81920$ batch of pixel-pairs for loss $\mathcal{L}_{pix}$ and $B_{reg} = 81920$ for loss $\mathcal{L}_{reg}$ while B_{pool} is left unbound to include all region pairs with IoU overlap above $IoU_r = 0.2$. To generate regions, we use the efficient graph-based segmentation method [15] with $scale = 85$ and $\sigma = 2000$. We obtain this value by generating segments with different hyper-parameters and empirically observing the segments on a handful of images from the Replica dataset. We use this default value for all other datasets. We take the output before the final layer as the feature to calculate our loss. We resize the feature map to the original input resolution using bilinear interpolation before projecting and matching across views. We use pixel IoU thresholds of $[IoU_l, IoU_h] = [0.3, 0.7]$ to get the overlapping image-pairs. We use the same augmentations for I_1 and I_2 as in [47] and $\lambda = 0.005$. To make pre-training more stable, we use a norm gradient clipping of 5 when using Barlow Twins loss. We use a learning rate of 0.01 with a cosine decay scheduler [30] without restarts. We pre-train for 20K (1x) iteration by default but also try 50K iterations (2.5x) schedules for Replica to compare with others. We use a learning rate warm-up period of 5% of total training iterations. We use a single V100 GPU on which 20K iterations take approximately 8 hours for the Replica dataset. Similar to [43], we pre-train the whole model excluding the final classification layer.

Baselines. The ResNet50 weights are loaded from the official Pytorch library which was trained for image classification on the ImageNet-1K dataset [12]. DeepLabV3+ [8] is our implementation trained on the ADE20K dataset [51,52] for semantic segmentation for 200 epochs which reaches a mIoU of 39.8 in the ADE20K validation set. We use this same architecture but with randomly initialized weights for our approach. Baseline self-supervised methods MoCov2 [10], SimSiam [11], PixPro [45] and PLRC [3] work on two augmented versions of the same image and trained in this manner. Owing to their large memory and batch-size requirements, these baselines are pre-trained on a single 80 GB A100 GPU. For a fairer comparison, following [22], we pre-train the self-supervised baselines for 1250 iteration (1x) and 3125 iteration (2.5x) schedules. The number of images seen by these models for 1x and 2.5x are 320K (256 x 1250) and 800K (256 x 3125) respectively, same as ours 320K (16 x 20K) and 800K (16 x 50K) images. The baselines take dramatically longer hours to train if the same number of iterations as ours are used and are computationally restrictive to perform. We also use our own version of DetCon [24] which is also trained using Barlow Twins loss and using our default hyper-parameters.

Fine-tuning. For fine-tuning, we use ground truth annotation from 5% of all the images in the pre-trained Replica scene and HM3D scene resulting in 16 and 30 images respectively. For AVD, we create 2 sub-datasets. We label 43 of the 2412 images from the chosen scene in AVD. In *AVD-easy* and *AVD-hard*, we fine-tune on 38 and 24 labeled images respectively. For **evaluation**, the model for each of the *specific* scenes is evaluated on the remaining labeled images from the scene. For a fair comparison, we use the same set of images for training and testing across all the methods in the experiments and use a learning rate of 0.01 with a polynomial scheduler for 20K iterations and a weight decay of $5e^{-4}$ for all models.

Supervised Pre-training To get an upper bound of our approach, we assume ground truth labels are available during pre-training such that each region exactly overlaps a single class in the environment. This is followed by our default fine-tuning. The results are shown in Table 1. The models that use random sampling are the worst performing models. We hypothesize their poor performance is due to inherent class imbalance in indoor environments with more pixels being sampled from the classes with large extent (e.g., walls, floors) that dominate the smaller ones. In region matching, the class imbalance is further enhanced with more regions coming from larger classes. Balanced sampling mitigates this problem. Region matching is better in balanced sampling because the positive pixel-pairs capture more variability compared to exact matching which may look very similar, especially across images captured from close locations. Matching pixels across regions does not require exact correspondences and allows us to sample more positive pixels-pairs within a batch.

4.2. Results

Replica We compare our RegConsist approach on the Replica dataset [39] with other supervised models and pre-training methods. The results are shown in Table 2. Our model performs the best, beating even the ADE20K super-

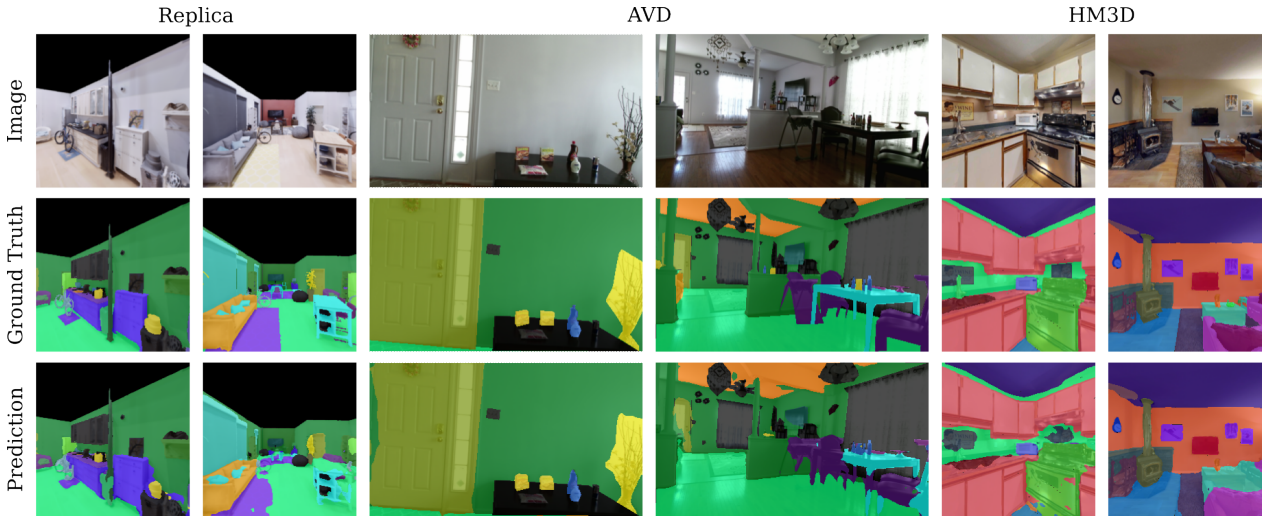

Figure 3. Segmentation Results from our model on the respective datasets. Dark pixels in ground truth and predictions are those without valid class labels. Best viewed digitally or in color.

method	Pre-Training Settings dataset	supervision	level	input image(s)	Fine-tune mIoU (2.5x)
Random	–	–	–	single	40.5
ResNet50 [23]	ImageNet-1K	classif.	image	single	55.9
DeepLabV3+ [8]	ADE20K	segment.	pixel	single	57.2
MoCov2 [10]	Replica	self	image	single (x2)	40.0 (41.2)
SimSiam [11]	Replica	self	image	single (x2)	43.3 (44.8)
BarlowTwins [47]	Replica	self	image	single (x2)	43.8 (37.8)
PixPro [45]	Replica	self	pixel	single (x2)	45.7 (46.4)
PLRC [3]	Replica	self	pixel	single (x2)	42.7 (42.5)
*DetCon$_{Barlow}$ [24]	Replica	self	region	single (x2)	38.9 (38.5)
RegConsist (Ours)	Replica	self	pix. + reg.	view-pairs	**59.7 (62.7)**

Table 2. Results on Replica dataset. Pre-trained models were fine-tuned on 5% of the images from the scene and evaluated on other 95% of the images. *DetCon$_{Barlow}$ is our own implementation of DetCon [24] but using Barlow twins loss. For self-supervised pre-training methods, mIoU shown outside brackets are for 1x iterations and those inside are for 2.5x iterations of pre-training. In input image(s) column, *single(x2)* means the method uses two augmented versions of the same view while *view-pairs* means two overlapping views from the scene are used.

vised model. Among the self-supervised models, [45] is the best performing model, which is also trained on pixel-level supervision similar to ours. However, our model aligns pixel features (paired using temporal and spatial consistency) as well as pooled features of overlapping regions and beats other self-supervised baselines pre-trained on the same dataset. Furthermore, the model pre-trained with our approach using a 1x pre-training schedule beats even the 2.5x schedule pre-trained baselines.

AVD Similarly, we compare our method pre-trained on AVD dataset [2] with a randomly initialized model and a model supervised on ADE20K [52] dataset. From Table 3, we can see that *AVD-easy* is easy even for the random initialization but *AVD-hard* is more difficult. We observe that our method does not perform better than the ADE20K model on AVD-hard. We suspect that AVD and ADE20K share some similar characteristics (real world, indoor scenes) so ADE20K model is able to utilize its existing knowledge unlike in Replica. This bolsters our proposed method of pre-training the models on the same dataset, especially when the domain difference is large. Also unlike for Replica, the bottom-up region segmentation for AVD is not tuned for fairer comparison.

HM3D Similarly, we compare our method with a model

method\dataset	**AVD**(easy)	**AVD**(hard)	**HM3D**
Random	66.7	49.1	41.2
ADE20K	69.3	**69.1**	49.8
RegConsist	**69.5**	64.8	**51.1**

Table 3. Results (mIoU) on AVD [2] and HM3D [36]. Model pre-trained using our approach (RegConsist) versus models initialized randomly (Random) and with supervised labels from ADE20K [52]. A small subset of images is used for fine-tuning each model while the remaining images from the scene are used for evaluation. All models are fine-tuned using the same images.

$\mathcal{L}_{pix}$	$\mathcal{L}_{reg}$	$\mathcal{L}_{pool}$	mIoU
✓			52.4
	✓		57.2
		✓	57.4
✓	✓		58.4
✓		✓	58.8
	✓	✓	58.5
✓	✓	✓	59.7

Table 4. Effect of using different combination of the three losses. Each row represents a separate instance of our model where only the respective losses are used.

randomly initialized and a model supervised on ADE20K [52]. From Table 3, we can see that, unlike Replica, we do not map HM3D classes to ADE20K. So, the ADE20K supervised model has a harder time learning newer classes. This demonstrates a more realistic scenario where classes do not overlap between existing models and the target dataset.

4.3. Ablations

We present a detailed ablation study on the contribution of different loss terms, the effect of the pixel batch size, the number of labeled examples, and the sensitivity of IoU threshold in our ablations experiments. All ablation experiments are performed on Replica [39] dataset with default hyper-parameters unless otherwise stated.

Losses Contribution. We try pre-training the model by using different combinations of the three losses $\mathcal{L}_{pix}$, $\mathcal{L}_{reg}$ and $\mathcal{L}_{pool}$, taking one combination at a time. This is followed by our default fine-tuning regime. Results are shown in Table 4. We observe that using only $\mathcal{L}_{pix}$ *i.e.* matching the exact corresponding pixels based on temporal consistency is the worst. Such positive pixel pairs are very similar to each other and do not possess enough variability to learn about regions. Both $\mathcal{L}_{reg}$ and $\mathcal{L}_{pool}$ losses individually perform better than $\mathcal{L}_{pix}$. Using a combination of any two of the losses is better than using only individual loss. Best performance is achieved when using all three losses together, surpassing all other combinations of losses. This shows that all three losses contribute to the performance of the model.

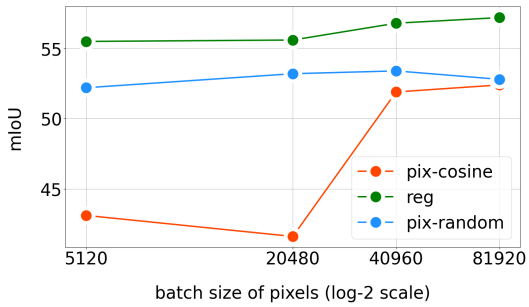

Figure 4. Effect of changing pixel batch size for $\mathcal{L}_{reg}$ and $\mathcal{L}_{pix}$. Model using $\mathcal{L}_{reg}$ loss is represented by *reg* while *pix-cosine* and *pix-random* are models that use $\mathcal{L}_{pix}$ loss with cosine distance sampling and random sampling of pixels respectively.

labeled images (%)	5%	10%	20%	30%
labeled images (count)	16	32	64	96
ADE20K	59.2	67.2	76.8	80.1
RegConsist (Ours)	**62.7**	**67.8**	**77.3**	**80.5**

Table 5. Number of Labeled Images. Increasing number of images on which the models are fine-tuned.

Pixel Batch Size. We experiment by changing the batch sizes of pixel pairs B_{pix} and B_{reg} we use in $\mathcal{L}_{pix}$ and $\mathcal{L}_{reg}$ when using each of the losses separately for each experiment. The results are shown in Figure 4. $\mathcal{L}_{reg}$ benefits from the increase in a batch size of pixels (reg in figure). For $\mathcal{L}_{pix}$ we perform random sampling between all temporal correspondences available by default (pix-random). Alternatively, we can choose the pairs that have a high cosine distance between their pixel features (pix-cosine). We found that using cosine distance for sampling is worse, especially for smaller batch sizes. We conjecture learning from only hard samples is a difficult task. When using larger B_{pix} values, however, both pixel sampling methods work almost equally well as there is a good mixture of hard and easy pairs.

Number of Labeled examples. We experiment by changing the number of labeled images used for fine-tuning. We fine-tune for 80K iterations instead of 20K. The results are shown in table 5. As can be seen, our method performs better than the ADE20K supervised model in every case. The performance gap decreases as more labeled images are available. This shows that our model is more suited when the number of annotations is low.

Image IoU Threshold. To prove our hypothesis that varied image pairs (I_1, I_2) are better than the ones where they are taken from similar location and pose, we try using a differ-

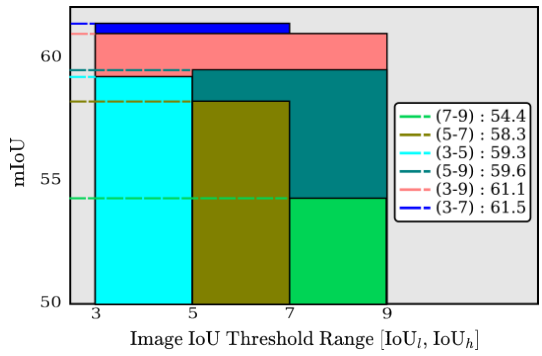

Figure 5. Image overlap IoU threshold range for guiding selection of image pairs (I_1, I_2) vs mIoU.

ent range of IoU thresholds $[IoU_l, IoU_h]$ when selecting the (I_1, I_2) image pairs (not regions) for pre-training. We follow this by fine-tuning for 80K iterations. From Figure 5, we can see that threshold $[7, 9]$ produces the worst result of 54.4 because the image-pairs (I_1, I_2) in this range overlap highly with each other meaning they were obtained from very similar poses of the camera. We find that threshold $[3, 7]$ produces the best results which shows that it is important to keep a balance between similar and dissimilar view pairs.

4.4. Discussion

The proposed method requires diverse pairs of overlapping views for effective self-supervision as demonstrated in the ablation studies. During the pre-training stage, the training data is gathered through the association of pixels across these diverse views from the specific environment. At the moment, we have achieved this by assuming the availability of camera poses and depth maps for ease of training and evaluation. This assumption can be relaxed by having alternate methods for computing correspondences between the views. The training data is collected and registered off-line, so, integration with on-line mapping and exploration strategies would enhance the applicability of our approach.

While Replica and HM3D datasets, have almost perfect depth and camera pose measurements, in the real world, these measurements are more prone to errors. Such errors can be encountered in the AVD dataset where COLMAP (a state-of-the art SLAM system) is used. Due to these errors, the gathered training data may contain wrong pixel/region association. Nevertheless, our proposed method works in all three datasets, including AVD, which demonstrates that few incorrect associations can be mitigated by enough correct associations.

The performance of the model depends on the quality of the class agnostic regions being used. The quality of such regions can be improved by using more recent class-agnostic segmentation methods such as the learning-based bottom-up segmentation model (SAM) [27]. There are also approaches [37,50] which can be utilized for scene completion to further gather labeling data. These approaches are complementary to our approach as they are tailored towards gathering more labels while we are proposing a method to learn better representation for the semantic segmentation model in the low data regime through self-supervised pre-training.

5. Conclusion

We have demonstrated the effectiveness of self-supervised pre-training for semantic segmentation models in an indoor environment by exploiting spatial and temporal consistency between overlapping views. The method exploits the ability to register neighboring views of an indoor scene and uses efficient generation of positive training examples for a contrastive learning framework using unsupervised segmentation approaches. The proposed approach was validated through several experiments and ablation studies, demonstrating the effects of different choices of sampling strategies, amounts of labeled data and comparing with other self-supervised approaches. We also demonstrate that our approach allows the agent to learn as well as a supervised model trained on labeled images from a similar dataset. The assumption of the availability of such relevant labeled data is not always valid and we argue that our approach is especially beneficial in such scenarios.

6. Acknowledgements

This material is based upon work supported by National Science Foundation under grant IIS 1925231 NSF NRI. The experiments were run on ARGO and HOPPER clusters provided by the Office of Research Computing at George Mason University.

References

[1] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C. Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8219–8228, October 2021. 5

[2] Phil Ammirato, Patrick Poirson, Eunbyung Park, Jana Kosecka, and Alexander C. Berg. A dataset for developing and benchmarking active vision. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017. 1, 4, 6, 7

[3] Yutong Bai, Xinlei Chen, Alexander Kirillov, Alan Yuille, and Alexander C. Berg. Point-level region contrast for object detection pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16061–16070, June 2022. 2, 5, 6

[4] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. 1

[5] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, Jose Neira, Ian Reid, and John J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6), 2016. 1, 3

[6] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4247–4258. Curran Associates, Inc., 2020. 1

[7] Devendra Singh Chaplot, Helen Jiang, Saurabh Gupta, and Abhinav Gupta. Semantic curiosity for active visual learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 309–326, Cham, 2020. Springer International Publishing. 1

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2, 5, 6

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 1

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 5, 6

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, June 2021. 1, 2, 5, 6

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[13] Michael Dorkenwald, Fanyi Xiao, Biagio Brattoli, Joseph Tighe, and Davide Modolo. Scvrl: Shuffled contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4132–4141, June 2022. 2

[14] Zhaoyuan Fang, Ayush Jain, Gabriel Sarch, Adam W. Harley, and Katerina Fragkiadaki. Move to see better: Self-improving embodied object detection. In *BMVC*, 2021. 2

[15] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59:167–181, 2004. 1, 2, 3, 5

[16] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation. In *International Conference on Learning Representations*, 2022. 1

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 1

[18] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006. 1

[19] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022. 1

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[22] Kaiming He, Ross Girshick, and Piotr Dollar. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 5

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5, 6

[24] Olivier J. Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10086–10096, October 2021. 2, 3, 4, 5, 6

[25] Olivier J. Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 123–143, Cham, 2022. Springer Nature Switzerland. 1

[26] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998. PMLR, 10–15 Jul 2018. 2

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Alexander C. Berg Spencer Whitehead, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *https://arxiv.org/pdf/2304.02643.pdf*, 2023. 8

[28] Yimeng Li, Arnab Debnath, Gregory Stein, and Jana Kosecka. Learning-augmented model-based planning for visual exploration. *arXiv preprint arXiv:2211.07898*, 2022. 1

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 4

[30] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 5

[31] Yi Ma, Stefano Soatto, Jana Kosecka, and S. Shankar Sastry. *An Invitation to 3-D Vision: From Images to Geometric Models*. SpringerVerlag, 2003. 3

[32] Chaitanya Mitash, Kostas E Bekris, and Abdeslam Boularias. A self-supervised learning system for object detection using physics simulation and multi-view pose estimation. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*, pages 545–551. IEEE, 2017. 2

[33] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11205–11214, June 2021. 2

[34] Etienne Pot, Alexander Toshev, and Jana Kosecka. Self-supervisory signals for robotic object discovery and detection. *arXiv preprint arXiv:1806.03370*, 2018. 2

[35] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6964–6974, June 2021. 2

[36] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 1, 4, 7

[37] Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. 3D Semantic Scene Completion: A Survey. *International Journal of Computer Vision*, 130(8):1978–2005, Aug. 2022. 8

[38] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 4

[39] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1, 3, 4, 5, 7

[40] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2

[41] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective Search for Object Recognition. *International Journal of Computer Vision*, 104(2):154–171, Sept. 2013. 2

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. 2

[43] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 2, 5

[44] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[45] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16684–16693, June 2021. 2, 5, 6

[46] Jinyu Yang, Weizhi An, Chaochao Yan, Peilin Zhao, and Junzhou Huang. Context-aware domain adaptation in semantic

segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2

[47] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 18–24 Jul 2021. 1, 2, 4, 5, 6

[48] Feihu Zhang, Philip Torr, Rene Ranftl, and Stephan Richter. Looking beyond single images for contrastive semantic segmentation learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3285–3297. Curran Associates, Inc., 2021. 2

[49] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[50] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15838–15847, October 2021. 8

[51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 4, 5

[52] 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*, 127(3):302–321, 2019. 4, 5, 6, 7