Copy-Pasting Coherent Depth Regions Improves Contrastive Learning for Urban-Scene Segmentation

Liang Zeng
liangzeng1996@outlook.com
Attila Lengyel
A.Lengyel@tudelft.nl
Nergis Tomen
N.Tomen@tudelft.nl
Jan van Gemert
J.C.vanGemert@tudelft.nl

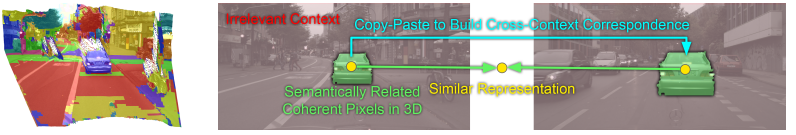
Delft University of Technology
Delft, The Netherlands

Abstract

In this work, we leverage estimated depth to boost self-supervised contrastive learning for segmentation of urban scenes, where unlabeled videos are readily available for training self-supervised depth estimation. We argue that the semantics of a coherent group of pixels in 3D space is self-contained and invariant to the contexts in which they appear. We group coherent, semantically related pixels into coherent depth regions given their estimated depth and use copy-paste to synthetically vary their contexts. In this way, cross-context correspondences are built in contrastive learning and a context-invariant representation is learned. For unsupervised semantic segmentation of urban scenes, our method surpasses the previous state-of-the-art baseline by +7.14% in mIoU on Cityscapes and +6.65% on KITTI. For fine-tuning on Cityscapes and KITTI segmentation, our method is competitive with existing models, yet, we do not need to pre-train on ImageNet or COCO, while we are also more computationally efficient. Our code is available on https://github.com/LeungTsang/CPCDR.

1 Introduction

The lack of labeled data is a bottleneck for deep-learning-based computer vision techniques due to the cost of annotating [12, 39]. To this end, researchers have established the "self-supervised pre-training then fine-tuning" paradigm which can use unlabeled data. One of the most successful approaches is self-supervised contrastive learning [6, 7, 9, 23, 26].



(a) (b)

Figure 1: (a) We group coherent, semantically related pixels into coherent depth regions as a prior for contrastive learning given estimated depth. We argue that the semantics of the coherent pixels are self-contained regardless of different contexts. (b) We copy-paste the regions to build cross-context correspondences. The corresponding representations are pulled together by contrastive learning.

We explore how to use depth to aid self-supervised contrastive learning for urban-scene segmentation [12, 18]. Applying contrastive learning to complex, non-object-centric urban scenes for segmentation is a non-trivial and often overlooked research topic. We argue that the semantic relatedness of pixels correlates with their coherence in 3D space. Therefore, we use estimated depth to group coherent, semantically related pixels into coherent depth regions. We then apply a copy-paste data augmentation strategy [20] to artificially vary the contexts in which the grouped coherent pixels appear and build cross-context correspondences. These correspondences are used as positive pairs in contrastive learning. As self-supervised depth estimation on urban scenes is well-addressed in literature [22, 56, 58] depth can be extracted from images freely.

We explore if a group of coherent pixels is invariant over different contexts and if such invariance benefits representation learning. Deep models for visual recognition are sensitive to contexts [3, 30, 53], where spurious correlations with the contexts learn non-existing shortcuts [19, 34]. This shortcut comes from the large receptive field of modern neural networks [24] and varying contexts by copy-pasting coherent regions in various scenes encourages the learned representations to be invariant to contexts. In Figure 2 we visualize the effect of our approach.

(a) Cars in different contexts (b) w/o copy-paste (c)with copy-paste

Figure 2: Feature maps are visualized as RGB images by PCA reduction [43]. Copy-paste is vital for learning object-specific representations invariant across different contexts.

Our contributions are summarized as follows. (1) We demonstrate that copy-paste augmentation of coherent depth regions is important for unsupervised contrastive learning of object-specific and context-invariant representations. (2) Our method allows effective pre-training on urban scene data, without the need for an additional large-scale dataset. (3) For unsupervised semantic segmentation, our method outperforms the previous state-of-the-art baseline by a significant +7.14% in mIoU on Cityscapes [12] and +6.65% on KITTI [18]. (4) For fine-tuning, our method allows effective pre-training on urban scenes with a single GPU. Our model achieves competitive segmentation performance on Cityscapes [12] and KITTI [18] to existing models that are pre-trained on ImageNet [39] or COCO [32] with more than 8 GPUs.

2 Related Work

Self-supervised Contrastive Learning Self-supervised pre-training has gained significant attention in computer vision thanks to its ability to use unlabeled data which is generally available in large quantities. Many self-supervised learning tasks have been explored [14, 21, 27, 34, 35, 45, 54], of which contrastive learning [23] currently most prevalent.

The design of contrastive learning, especially the definition of positive and negative samples, depends on the data and task at hand. Pioneering works [5, 6, 7, 8, 9, 10, 26, 42, 51] are mostly based on instance discrimination on ImageNet [39]. Positive pairs are generated by applying different transformations to the same image, while negative pairs consist of different samples. For dense prediction [33], VADeR [36], DenseCL [46] and PixPro [48] learn dense visual representations by performing instance discrimination at pixel-level. For object detection on complex scenes [37], object-level instance discrimination is possible if object locations are known [40, 47]. For segmentation additional similarity can be defined among certain regions on images, such as classic hierarchical pixel grouping [55], auxiliary labels [52] and salient object estimation [17].

We adopt SwAV [5] to perform contrastive learning as it only needs positive samples. We extend it to learn dense visual representation as our goal is segmentation. The positive samples can be defined at pixel-level or region-level. We compute precise pixel correspondences with respect to the geometric transformations similarly as in PixPro [48] and VADeR [36], but our geometric transformations involve copy-paste in addition. Region-level positive samples are sampled from pixel grouping results as recent works [17, 28, 52, 55], but our pixel grouping is based on the novel 3D coherence given the depth estimation.

Copy-Paste Data Augmentation Copy-Paste is a well-studied augmentation in supervised learning [15, 20, 50]. Copy-paste alike augmentations can also be used for self-supervised contrastive learning as long as we know what to copy-paste. DiLo [57] swaps salient objects from images while RegionCL [49] and CP2 [44] swap random rectangular patches. However, the assumption that the extracted objects or patches are semantically equivalent to the original images is only valid on simple object-centric datasets [39].

Our work utilizes depth to define semantically related regions for copy-paste which improves the learning of dense visual representations in complex urban scenes.

Self-Supervised Depth Estimation Self-supervised depth estimation on videos has gained impressive success [22, 56, 58]. As video clips are generally available in urban-scene datasets [12, 18], depth estimation may help supervised segmentation in different ways [29], including jointly training, data selection and DepthMix data augmentation.

In our work, estimated depth is used for grouping semantically related pixels and generating augmented input samples for self-supervised learning. DepthMix [29] is adopted to enhance copy-paste augmentation.

3 Method

We illustrate the pipeline of our method which consists of four steps, as shown in Figure 3.

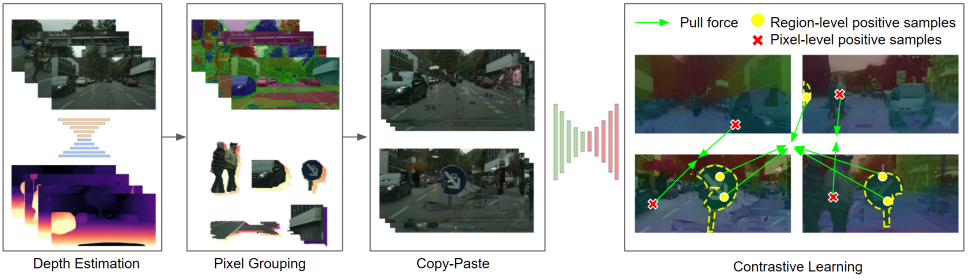


Figure 3: Our method consists of four steps. 1. Training a depth estimator on video clips by self-supervision. 2. Grouping pixels coherent in 3D space given the depth. 3. Building cross-context correspondences by copy-paste. 4. Pulling together the representations of corresponding pixels and regions using contrastive learning.

3.1 Self-Supervised Depth Estimation

To obtain depth, we take advantage of the recent success of self-supervised depth estimation on monocular videos [22, 56, 58]. In our framework, we adopt Monodepth2 [22] in virtue of its simplicity and effectiveness.

3.2 Coherent Pixels Grouping

We propose a heuristic algorithm to group semantically related pixels that are coherent in 3D space into coherent depth regions based on the estimated depth and horizontal pattern of urban-scene images.

For each image, we compute the normal and 3D coordinate of pixels based on the depth D. We use SLIC superpixels [1] to downsample the image as shown in Figure 4 (a) and build a region adjacency graph from the superpixels. Each superpixel node contains the average 3d coordinate $[x,y,z]$ and upward normal N_y perpendicular to the ground. With these attributes we weigh the edges based on how likely the edges lie on spatial boundaries.

We define two types of boundaries, namely occlusion and supporting boundaries [41], as shown in Figure 4 (b). Let us consider two adjacent superpixels $\mathbf{S}_1$ and $\mathbf{S}_2$, where $\mathbf{S}_1$ has lower height y_1. To detect occlusion boundaries where the foreground occludes the background we compute the euclidean distance between $\mathbf{S}_1$ and $\mathbf{S}_2$ as in Eqn. 1. Since the 3D scale of superpixels increases proportionally to depth, the distance should be normalized by their joint depth. To detect supporting boundaries where the objects rest on the ground, we assume that the lower $\mathbf{S}_1$ is the ground and $\mathbf{S}_2$ is an object. The pattern of urban-scene images implies that the ground surfaces should be level, which is measured by the normal in the upward direction n_{y1}. The ground should also be low which is measured by the square of y_1 when $y_1 < 0$. The objects on the ground should be upright, measured by the difference of their normals n_{y1} and n_{y2}. The three terms are multiplied resulting in Eqn. 2.

$$D_{ocln} = \frac{\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}}{z_1+z_2} \tag{1}$$

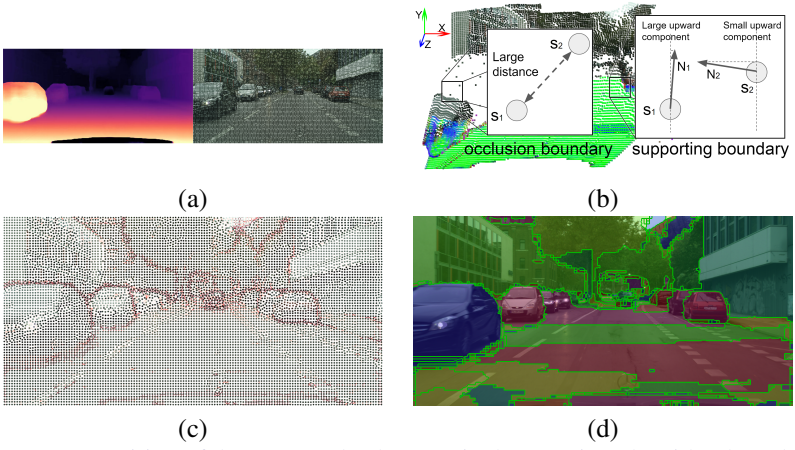


Figure 4: Decomposition of the proposed coherent pixels grouping algorithm based on depth. (a) Estimated depth and RGB image with SLIC superpixels. (b) Typical examples of occlusion and supporting boundaries. The bottom of point cloud is colored based on the normal vector. (c) Weighted boundary graph with weakly connected edges indicated in red. (d) The final grouping results from the community detection.

$$D_{sup} = \begin{cases} \max(n_{y1},0)(\max(n_{y1}-n_{y2},0))y_1^2 & y_1 < 0 \\ 0 & y_1 \geq 0 \end{cases} \tag{2}$$

Finally, the edge weight W between $\mathbf{S}_1$ and $\mathbf{S}_2$ is computed as the weighted sum of the distance terms in Eqns. 1 and 2 and a bias term, normalized by the sigmoid function, as shown in Eqn. 3.

$$W = \sigma(w_{ocln}D_{ocln} + w_{sup}D_{sup} + b), \tag{3}$$

where σ is the sigmoid function. With the weighted graph as shown in Figure 4 (c), our goal is to group strongly connected nodes while separating them from weakly connected nodes from other groups. Weak connections are indicated in red. As such we turn pixel grouping into a community detection problem [4, 16]. We use the classic InfoMap algorithm [38] to fulfill the task. Details are provided in supplementary materials.

3.3 Copy-Paste

We copy-paste the coherent depth regions to build cross-context correspondences. We sample M images and extract all regions above a certain size. Those images also act as backgrounds. We perform copy-paste in two rounds. All regions will be pasted e times on every unmodified background in a round. In the first round, we set a small e to display more parts of the original images. In the second round, e is made large to allow more regions to be present in new contexts. The generated $2M$ images form a training sample. Details are provided in supplementary materials.

Along with copy-paste, random resize, horizontal flip, color jitter, and Gaussian blur are applied. To generate more realistic-looking images, we adopt DepthMix [29] to simulate occlusions. The depth of the copied patch and its target position are compared and only the

pixels with smaller depth, i.e. closer to the camera are kept. Moreover, the height of the target position should be within a threshold h_t of the original height of the patch. All geometric transformations are recorded as transformation matrices so that pixel correspondences can be traced back during training.

3.4 Contrastive Learning and Positive Samples Sampling

SwAV contrastive learning framework We adopt the clustering-based contrastive learning framework introduced by SwAV [5] for its simplicity since it only relies on positive samples. We recap SwAV, with a slight modification in the swap prediction loss since the number of positive samples is inconsistent between each group. With a set of pixel representations denoted by $\mathbf{Z} = [\mathbf{z}_0, \mathbf{z}_1, ..., \mathbf{z}_N]$ sampled from the feature maps, we compute their soft assignments $\mathbf{P} = [\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_N]$ to K prototypes $\mathbf{C} = [\mathbf{c}_0, \mathbf{c}_1, ..., \mathbf{c}_K]$ by Eqn. 4 with temperature τ as follows:

$$\mathbf{p}_n^{(k)} = \frac{\exp\left(\frac{1}{\tau}\mathbf{z}_n^\top \mathbf{c}_k\right)}{\sum_{k'} \exp\left(\frac{1}{\tau}\mathbf{z}_n^\top \mathbf{c}_{k'}\right)} \tag{4}$$

Directly minimizing the cross-entropy for the assignments of positive samples will lead to a trivial solution. Codes $\mathbf{Q}$ are computed by Sinkhorn-Knopp algorithm [13] which is the balanced assignments with the smallest distance to $\mathbf{P}$. We compute the average code within each group of positive samples and get the target Codes $\overline{\mathbf{Q}} = [\overline{\mathbf{q}}_0, \overline{\mathbf{q}}_1, ..., \overline{\mathbf{q}}_N]$ by Eqn. 5.

$$\overline{\mathbf{q}}_n^{(k)} = \frac{1}{|\mathbf{G}^n|} \sum_{i \sim \mathbf{G}^n} \mathbf{q}_i^{(k)} \quad n \sim \mathbf{G}^n \tag{5}$$

The loss for $\mathbf{Z}$ is given by the mean of cross-entropy of all assignments and their corresponding codes as Eqn. 6.

$$\mathcal{L}_\mathbf{Z} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \overline{\mathbf{q}}_n^{(k)} \log \mathbf{p}_n^{(k)} \tag{6}$$

Positive sampling The positive samples are defined in two ways. For pixel-level positive samples, the same pixels under different transformations are positive samples. We randomly sample a certain number of initial coordinates and locate their positive counterparts within the batch using the stored geometric transformations. For region-level positive samples, we assume the pixels inside the same coherent depth region share similarity. We randomly sample coordinates, where coordinates from the same region are considered positive samples.

Both sampling methods can be used separately or jointly during training. Let $\mathcal{L}_{pixel}$ and $\mathcal{L}_{region}$ denote the loss computed among the pixel-level positive samples and region-level positive samples. We balance them with a weight λ as in Eqn. 7.

$$\mathcal{L} = \lambda \mathcal{L}_{pixel} + (1 - \lambda)\mathcal{L}_{region} \tag{7}$$

3.5 Clustering

For unsupervised semantic segmentation, we need to cluster the learned representations. A fine-grained clustering into prototypes is already done by SwAV [5] and we can further group these to match the number of target classes. Empirically, we chose agglomerative clustering with cosine distance and average linkage criterion.

4 Experiments

We evaluate on Cityscapes [12] and KITTI [18]. Cityscapes has 2975 30-frame stereo videos with one frame annotated in the *train* set and 500 images in the *val* set. KITTI has 71 stereo videos and 200 annotated images in the *train* set. We only used the left images, resulting in 89250 images for Cityscapes and 42382 for KITTI. We resized Cityscapes images to 384×768 and KITTI images to 384×1280 pixels. We used semantic FPN [31] with ResNet-50 [24] as the feature extractor. The final classifier was replaced by a 1-layer MLP projection head with 128 output channels. All experiments are conducted on a single 16GB V100 GPU. Further details on training and evaluation setting can be found in supplementary materials.

4.1 Unsupervised Semantic Segmentation

Segmentation can be generated by clustering over the learned dense visual representation directly. The unsupervised semantic segmentation performance of our method and PiCIE [11] with ImageNet pre-training or random initialization is shown in Tab. 1. Using ImageNet initialization [39], our method surpasses PiCIE [11] by +7.14% and 6.65% in mIoU on Cityscapes [12] and KITTI [18], respectively. Our method can better capture the outline of objects, as shown in Figure 5, whereas PiCIE can only predict the coarse masks of big objects and *stuff*, resulting in a good accuracy but poor mIoU. Our method performs more consistently when training on Cityscapes [12] then testing on KITTI [18], or vice versa. We also observe that PiCIE [11] performs poorly when initialized from scratch and diverges to a trivial solution as training progresses. Our method does not suffer from these shortcomings and is able to learn semantic clustering also when initialized from scratch, outperforming PiCIE [11] by a large margin in both accuracy and mIoU.

Method	Init.	Training Data	CS-Sem.		KT-Sem.	
			Acc	*mIoU*	*Acc*	*mIoU*
PiCIE	scratch	Cityscapes	33.56	8.33	32.20	6.52
Ours($\lambda = 0.5$)	scratch	Cityscapes	65.42	20.49	**68.37**	21.03
PiCIE	scratch	KITTI	30.28	6.81	30.62	7.54
Ours($\lambda = 0.5$)	scratch	KITTI	49.18	17.20	49.58	18.22
PiCIE	ImageNet	Cityscapes	**68.50**	16.24	56.74	13.54
Ours($\lambda = 0.5$)	ImageNet	Cityscapes	66.70	**23.38**	68.25	**22.50**
PiCIE	ImageNet	KITTI	53.24	12.55	61.74	12.92
Ours($\lambda = 0.5$)	ImageNet	KITTI	56.96	18.85	59.11	19.57

Table 1: Unsupervised semantic segmentation performance on Cityscapes *val* set and KITTI *train* set. We retrained PiCIE with equivalent setting to ours.

Figure 5: Qualitative comparison of unsupervised semantic segmentation on Cityscapes.

4.2 Semi-Supervised Segmentation

In semi-supervised segmentation the self-supervised pre-trained models are used as initialization for the following supervised learning. For urban scene segmentation a common practice is to pre-train on large scale datasets, such as ImageNet [39] and COCO [32]. Our method, in contrast, allows directly pre-training on the target urban scene data, without the need for extra curated datasets.

We compared our pre-trained model with existing models [5, 40, 46, 47, 48] trained on ImageNet [39] and COCO [32] by their original authors. However, pre-trained models on Cityscapes [12] and KITTI [18] are not publicly available for the baseline methods and neither are we able to train them using the original setting due to limited compute power. We trained SwAV [5] and PixPro [48] on Cityscapes [12] or KITTI [18] on a single 16 GB GPU using their official implementations. In this experiment, all models were initialized from scratch in pre-training and used ResNet-50 [24] backbone.

Pre-training Method	Pre-training Dataset	CS-Sem.	CS-Inst.		KT-Sem.	KT-Inst.	
		mIoU	*AP*	AP_{50}	*mIoU*	*AP*	AP_{50}
scratch	-	65.11	24.30	46.98	32.99	8.15	15.79
supervised	ImageNet	70.54	27.34	50.59	40.09	12.38	23.42
SwAV	ImageNet	71.07	28.08	52.25	40.52	**13.78**	**27.90**
DenseCL	ImageNet	72.09	28.97	51.93	40.88	12.63	22.74
PixPro	ImageNet	72.66	29.04	52.59	40.50	13.04	24.95
ORL	COCO	72.32	**29.94**	52.55	41.88	12.02	23.48
CAST	COCO	69.92	27.33	51.31	38.78	10.67	20.13
SwAV	Cityscapes	61.69	23.62	46.21	36.10	9.22	18.01
PixPro	Cityscapes	61.64	23.78	46.45	36.99	9.61	18.73
SwAV	KITTI	60.74	23.51	46.08	36.90	9.42	18.11
PixPro	KITTI	61.25	23.23	46.23	37.28	9.45	18.57
Ours($\lambda = 1$)	Cityscapes	**73.55**	**29.94**	**52.88**	**42.70**	12.58	24.98
Ours($\lambda = 0.5$)	Cityscapes	73.03	29.11	51.87	42.32	12.22	23.16
Ours($\lambda = 1$)	KITTI	71.62	28.77	52.71	41.17	11.74	20.57
Ours($\lambda = 0.5$)	KITTI	71.45	27.86	51.16	41.03	11.36	20.49

Table 2: Segmentation performance over Cityscapes *val* set and 5-fold validation of KITTI *train* set. SwAV and PixPro on Cityscapes and KITTI are trained with limited GPU compute as ours.

Semantic segmentation We show semi-supervised semantic segmentation performance on Cityscapes [12] and KITTI [18] in Tab. 2. Our method pre-trained on Cityscapes [12] with $\lambda = 1$ outperforms other existing pre-trained models for semantic segmentation by a noticeable margin on both datasets. It exceeds the SwAV [5] pre-trained model by $+3.01\%$ on Cityscapes [12] and $+2.18\%$ on KITTI [18] in mIoU. The KITTI pre-trained model also achieves a comparable result on both datasets with less training data. Overall, our method demonstrates better performance and generalization across different urban scene datasets.

Instance segmentation We show semi-supervised instance segmentation performance in Tab. 2 on Cityscapes [12] and KITTI [18]. Our Cityscapes pre-trained model with $\lambda = 1$ still slightly outperforms other baselines on Cityscapes [12]. It still exceeds SwAV [5] by

+1.86% in AP and +0.63% in AP_{50} on Cityscapes [12]. The performance on KITTI [12] is below SwAV [5] by −1.20% in AP and −2.92% in AP_{50}. This is expected as some layers of semantic FPN [31] are discarded to adapt the Mask-RCNN [25] architecture for instance segmentation.

Training baseline methods on urban scenes Pre-training on Cityscapes and KITTI with SwAV [5] and PixPro [48] yields inferior results. It indicates that the improved performance of our method is not only due to pre-training on similar data to the downstream tasks, but also from our effective design, especially under limited compute. Training details and discussion are provided in the supplementary material.

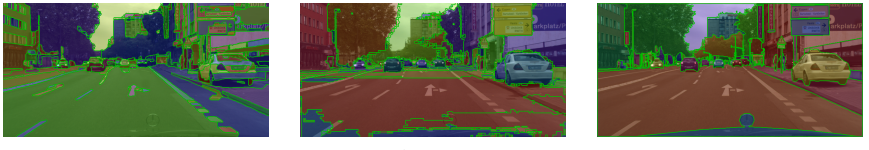
4.3 Ablation Studies

We performed ablation studies to validate our design choices. In ablation experiments, the models were trained on the 2975 images from the *train* set of the Cityscapes [12] for 100 epochs. The semantic segmentation performance is reported after fine-tuning on $1/16$ of the *train* set for 6k iterations.

Region proposal To study the significance of our coherent depth region proposal we compared different region proposal methods, including owt-ucm [2] used by Zhang et al. [53], as shown in Figure 6. We also split the instance label in Cityscapes [12] *train* into only connected regions as the optimal region proposals (ground truth). We show the fine-tuning performance in Tab. 4. Our coherent depth region proposal produces the closest performance to the ground truth.

Weight	Fine-tuning			Clustering		
	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$
owt-ucm [2]	44.83	47.01	47.24	13.64	15.70	12.11
Ours	46.91	48.87	48.55	15.94	20.71	13.94
GT	49.92	52.26	48.82	24.05	27.35	14.59

Table 3: Effects of different types of object proposals. Measurement is based on mIoU by fine-tuning on a subset of Cityscapes [12] for semantic segmentation.



(a)owt-ucm [2] (b)Ours (c)GT

Figure 6: Examples of different region proposals.

Copy-Paste To further investigate the effectiveness of copy-paste, we altered the number of images M for copy-paste. We report the fine-tuning and clustering performance in Tab. 4. Mixing more images to build richer cross-context correspondence is beneficial in all cases, especially when region-level positive samples are sampled ($\lambda < 1$).

Weight	M / Proposal	0	2	4	8	0	2	4	8
		Fine-tuning				Clustering			
$\lambda = 0$	Ours	35.18	42.34	46.75	46.91	5.93	13.92	13.65	15.94
	GT	37.70	46.74	48.64	49.92	9.39	21.31	23.51	24.05
$\lambda = 0.5$	Ours	34.35	45.92	48.23	48.87	3.97	14.03	20.32	20.71
	GT	38.01	49.26	50.06	52.26	8.20	24.26	24.82	27.35
$\lambda = 1$	Ours	39.93	48.24	48.09	48.55	4.89	10.59	13.65	13.94
	GT		48.20	49.25	48.81		10.36	14.08	14.59

Table 4: Effects of loss weighting (Eqn. 7) and number of sample images M for copy-paste. 0 means copy-paste is disabled. Measurement is based on mIoU by fine-tuning on subset of Cityscapes [12] for semantic segmentation.

Loss Weight The effects of the loss weight in Eqn. 7 are shown in Tab. 4. With high-quality proposals like the ground truth, combining the pixel-level and region-level positive samples remarkably increases the performance for both fine-tuning and unsupervised clustering. Our coherent depth region proposals fail to strictly contain a single class as Figure 6b. Thus, including region-level positive samples takes little effect on fine-tuning performance as shown in Tab. 4 or even slightly degrades the performance as shown in Tab. 2. However, for unsupervised segmentation performance, combining both pixel and region level positive samples is still beneficial.

5 Discussion and Limitations

In this work we leverage estimated depth to group semantically related coherent pixels, which allows copy-paste to build cross-context correspondences for contrastive learning. Experiments on Cityscapes [12] and KITTI [18] show that our method trained with one GPU surpasses previous state-of-the-art on unsupervised semantic segmentation and achieves similar fine-tuning performance to existing methods that are pre-trained on ImageNet [39] or COCO [32] with 8 GPUs. Our experiments also demonstrate the importance of using high quality region proposals for our method. The current pixel grouping algorithm is handcrafted and requires some task- and dataset-dependent parameter tuning. Future work therefore includes investigating data-driven approaches. Lastly, our method is particularly well suited to urban scene related datasets but its not clear if it is also applicable in other scenarios where depth is available or can be extracted. One inherent limitation of relying on self-supervised depth estimation is that these methods are developed and evaluated almost exclusively on urban scene images. Further experiments are needed to evaluate how reliably depth can be extracted in e.g. indoor scenes, and how well our method generalizes in these new settings. We hope our work inspires community to explore new applications and uses of depth in self-supervised representation learning. We invite researchers to explore and build upon our publicly available codebase.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, 2012. doi: 10.1109/TPAMI.2012.120. URL https://doi.org/10.1109/TPAMI.2012.120.

[2] Pablo Arbelaez, Michael Maire, Charless C. Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, 2011. doi: 10.1109/TPAMI.2010.161. URL https://doi.org/10.1109/TPAMI.2010.161.

[3] Ehud Barnea and Ohad Ben-Shahar. Exploring the bounds of the utility of context for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7412–7420. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00759. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Barnea_Exploring_the_Bounds_of_the_Utility_of_Context_for_Object_CVPR_2019_paper.html.

[4] Arnaud Browet, Pierre-Antoine Absil, and Paul Van Dooren. Community detection for hierarchical image segmentation. In Jake K. Aggarwal, Reneta P. Barneva, Valentin E. Brimkov, Kostadin Koroutchev, and Elka Korutcheva, editors, *Combinatorial Image Analysis - 14th International Workshop, IWCIA 2011, Madrid, Spain, May 23-25, 2011. Proceedings*, volume 6636 of *Lecture Notes in Computer Science*, pages 358–371. Springer, 2011. doi: 10.1007/978-3-642-21073-0_32. URL https://doi.org/10.1007/978-3-642-21073-0_32.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/fcbc95ccdd551da181207c0c1400c655-Abstract.html.

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15750–15758. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html.

[9] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. URL https://arxiv.org/abs/2003.04297.

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9620–9629. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00950. URL https://doi.org/10.1109/ICCV48922.2021.00950.

[11] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16794–16804. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Cho_PiCIE_Unsupervised_Semantic_Segmentation_Using_Invariance_and_Equivariance_in_Clustering_CVPR_2021_paper.html.

[12] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.350. URL https://doi.org/10.1109/CVPR.2016.350.

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2292–2300, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

[14] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1422–1430. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.167. URL https://doi.org/10.1109/ICCV.2015.167.

[15] Haoshu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yonglu Li, and Cewu Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul,*

Korea (South), October 27 - November 2, 2019, pages 682–691. IEEE, 2019. doi: 10.1109/ICCV.2019.00077. URL https://doi.org/10.1109/ICCV.2019.00077.

[16] Santo Fortunato. Community detection in graphs. *CoRR*, abs/0906.0612, 2009. URL http://arxiv.org/abs/0906.0612.

[17] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10032–10042. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00990. URL https://doi.org/10.1109/ICCV48922.2021.00990.

[18] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[19] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

[20] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2918–2928. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Ghiasi_Simple_Copy-Paste_Is_a_Strong_Data_Augmentation_Method_for_Instance_CVPR_2021_paper.html.

[21] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=S1v4N2l0-.

[22] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3827–3837. IEEE, 2019. doi: 10.1109/ICCV.2019.00393. URL https://doi.org/10.1109/ICCV.2019.00393.

[23] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 1735–1742. IEEE Computer Society, 2006. doi: 10.1109/CVPR.2006.100. URL https://doi.org/10.1109/CVPR.2006.100.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.322. URL https://doi.org/10.1109/ICCV.2017.322.

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL https://doi.org/10.1109/CVPR42600.2020.00975.

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. URL https://arxiv.org/abs/2111.06377.

[28] Olivier J. Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aäron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10066–10076. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00993. URL https://doi.org/10.1109/ICCV48922.2021.00993.

[29] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Köring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11130–11140. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Hoyer_Three_Ways_To_Improve_Semantic_Segmentation_With_Self-Supervised_Depth_Estimation_CVPR_2021_paper.html.

[30] Osman Semih Kayhan and Jan C. van Gemert. Evaluating context for deep object detectors. *CoRR*, abs/2205.02887, 2022. doi: 10.48550/arXiv.2205.02887. URL https://doi.org/10.48550/arXiv.2205.02887.

[31] Alexander Kirillov, Ross B. Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6399–6408. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00656. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Kirillov_Panoptic_Feature_Pyramid_Networks_CVPR_2019_paper.html.

[32] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. doi: 10.1109/CVPR.2015.7298965.

[34] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, volume 9910 of *Lecture Notes in Computer Science*, pages 69–84. Springer, 2016. doi: 10.1007/978-3-319-46466-4_5. URL https://doi.org/10.1007/978-3-319-46466-4_5.

[35] Deepak Pathak, Ross B. Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6024–6033. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.638. URL https://doi.org/10.1109/CVPR.2017.638.

[36] Pedro O. Pinheiro, Amjad Almahairi, Ryan Y. Benmalek, Florian Golemo, and Aaron C. Courville. Unsupervised learning of dense visual representations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3000311ca56a1cb93397bc676c0b7fff-Abstract.html.

[37] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html.

[38] M. Rosvall, D. Axelsson, and C. T. Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, nov 2009. doi: 10.1140/epjst/e2010-01179-1. URL https://doi.org/10.1140%2Fepjst%2Fe2010-01179-1.

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[40] Ramprasaath R. Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11058–11067. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/

html/Selvaraju_CASTing_Your_Model_Learning_To_Localize_Improves_Self-Supervised_Representations_CVPR_2021_paper.html.

[41] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V*, volume 7576 of *Lecture Notes in Computer Science*, pages 746–760. Springer, 2012. doi: 10.1007/978-3-642-33715-4_54. URL https://doi.org/10.1007/978-3-642-33715-4_54.

[42] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c2e5eaae9152079b9e95845750bb9ab-Abstract.html.

[43] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, volume 11217 of *Lecture Notes in Computer Science*, pages 402–419. Springer, 2018. doi: 10.1007/978-3-030-01261-8_24. URL https://doi.org/10.1007/978-3-030-01261-8_24.

[44] Feng Wang, Huiyu Wang, Chen Wei, Alan L. Yuille, and Wei Shen. CP2: copy-paste contrastive pretraining for semantic segmentation. *CoRR*, abs/2203.11709, 2022. doi: 10.48550/arXiv.2203.11709. URL https://doi.org/10.48550/arXiv.2203.11709.

[45] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2794–2802. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.320. URL https://doi.org/10.1109/ICCV.2015.320.

[46] 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, virtual, June 19-25, 2021*, pages 3024–3033. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wang_Dense_Contrastive_Learning_for_Self-Supervised_Visual_Pre-Training_CVPR_2021_paper.html.

[47] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. In *NeurIPS*, 2021.

[48] 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, virtual, June 19-25, 2021*, pages 16684–16693. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Xie_Propagate_Yourself_Exploring_Pixel-Level_Consistency_for_Unsupervised_Visual_Representation_Learning_CVPR_2021_paper.html.

[49] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Regioncl: Can simple region swapping contribute to contrastive learning? *CoRR*, abs/2111.12309, 2021. URL https://arxiv.org/abs/2111.12309.

[50] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6022–6031. IEEE, 2019. doi: 10.1109/ICCV.2019.00612. URL https://doi.org/10.1109/ICCV.2019.00612.

[51] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane 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, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 2021. URL http://proceedings.mlr.press/v139/zbontar21a.html.

[52] Feihu Zhang, Philip Torr, Rene Ranftl, and Stephan Richter. Looking beyond single images for contrastive semantic segmentation learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[53] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12982–12991. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01300. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Zhang_Putting_Visual_Object_Recognition_in_Context_CVPR_2020_paper.html.

[54] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, volume 9907 of *Lecture Notes in Computer Science*, pages 649–666. Springer, 2016. doi: 10.1007/978-3-319-46487-9_40. URL https://doi.org/10.1007/978-3-319-46487-9_40.

[55] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*,

2020. URL https://proceedings.neurips.cc/paper/2020/hash/c1502ae5a4d514baec129f72948c266e-Abstract.html.

[56] Chaoqiang Zhao, Qiyu Sun, Chongzhen Zhang, Yang Tang, and Feng Qian. Monocular depth estimation based on deep learning: An overview. *CoRR*, abs/2003.06620, 2020. URL https://arxiv.org/abs/2003.06620.

[57] Nanxuan Zhao, Zhirong Wu, Rynson W. H. Lau, and Stephen Lin. Distilling localization for self-supervised representation learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 10990–10998. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17312.

[58] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6612–6619. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.700. URL https://doi.org/10.1109/CVPR.2017.700.