ViCE: Improving Dense Representation Learning by Superpixelization and Contrasting Cluster Assignment

Robin Karlsson[1]
karlsson.robin@g.sp.m.is.nagoya-u.ac.jp
Tomoki Hayashi[1]
hayashi.tomoki@g.sp.m.is.nagoya-u.ac.jp
Keisuke Fujii[1]
fujii@i.nagoya-u.ac.jp
Alexander Carballo[1]
alexander@g.sp.m.is.nagoya-u.ac.jp
Kento Ohtani[1]
ohtani.kento@g.sp.m.is.nagoya-u.ac.jp
Kazuya Takeda[1,2]
kazuya.takeda@nagoya-u.jp

[1] Graduate School of Informatics
Nagoya University
Aichi, Japan

[2] Tier IV Inc.
Tokyo, Japan

Abstract

Recent self-supervised models have demonstrated equal or better performance than supervised methods, opening for AI systems to learn visual representations from practically unlimited data. However, these methods are typically classification-based and thus ineffective for learning high-resolution feature maps that preserve precise spatial information. This work introduces superpixels to improve self-supervised learning of dense semantically rich visual concept embeddings. Decomposing images into a small set of visually coherent regions reduces the computational complexity by $\mathcal{O}(1000)$ while preserving detail. We experimentally show that contrasting over regions improves the effectiveness of contrastive learning methods, extends their applicability to high-resolution images, improves overclustering performance, superpixels are better than grids, and regional masking improves performance. The expressiveness of our dense embeddings is demonstrated by improving the SOTA unsupervised semantic segmentation benchmark on Cityscapes, and for convolutional models on COCO. Code is available at https://github.com/robin-karlsson0/vice.

1 Introduction

Progress in general computer vision tasks in the past decade has been based on supervised learning with large datasets annotated by human labelers [63]. Arguments are made that generalizable and robust computer vision models have not yet been achieved, and further increasing the amount of labeled data is unsustainable [28, 74]. One hypothesis is that learning

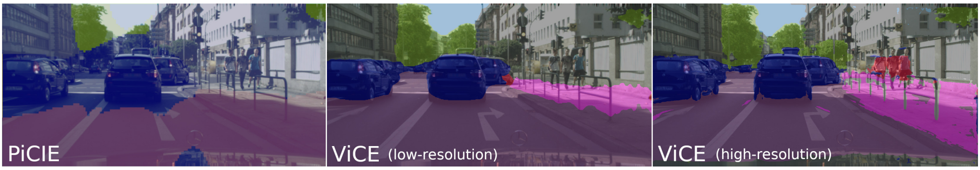

Figure 1: ViCE learns dense semantic embeddings from raw image data. Unsupervised semantic segmentation experiments show that our embedding maps are semantically richer and fit the content better compared to the SOTA baseline PiCIE [23]. Superpixelization further improves our results by enabling dense contrastive learning over high-resolution images.

from top-down categorization ("what it *is*") from semantically vague and inconsistent human annotation could be a limiting factor [35]. Instead, cognitive science tells us that learning from bottom-up association ("what it is *like*") may be more similar to how visual concepts emerge for humans [75, 81, 82, 93]. The success of bottom-up learning for word embeddings in natural language processing (NLP) [49, 76, 77] further strengthens the hypothesis. Recent self-supervised computer vision methods show promise in this direction with results approaching or even surpassing those of supervised methods [44]. However, these methods are classification-based and thus ineffective for learning high-resolution dense feature maps. Such maps are needed to associate semantic embeddings to spatial regions in vision inputs.

We introduce a method for improving the effectiveness of self-supervised classification methods for dense representation learning by decomposing images into a small set of visually coherent regions using superpixelization. We demonstrate how applying the method enables the contrasting cluster assignments method SwAV [14] to learn dense representations. The contributions of our paper are as follows:

- A new conceptual approach to represent high-resolution images as semantically rich embedding maps partitioned into distinct, coherent regions, represented by a latent **V**isual **C**oncept **E**mbedding (ViCE), analogous to word embeddings in NLP.
- Introduce superpixelization as a natural hierarchical region decomposition for dense contrastive learning in unsupervised semantic segmentation of high-resolution images. We demonstrate how to effectively implement self-supervised classification methods with region decomposition.
- Present SOTA unsupervised semantic segmentation results on Cityscapes, and for convolutional models on COCO.
- Experimentally demonstrate; Online contrasting cluster assignment [14] improves dense representation learning performance compared with offline clustering [12, 23]. Image decomposition by superpixelization improves performance, reduces computational time, and is more effective than grids. The ability to use high-resolution images improves performance. Contextual region masking improves performance.

2 Related work

Self-supervised visual representation learning Early works experimented with pretext tasks as a substitute for human annotations [9, 33, 41, 80, 86, 113]. Recent work demonstrates that image-level embedding classification with cross-entropy minimization on large datasets is a more effective approach capable of surpassing supervised pretraining [15, 44]. Contrastive methods [20, 24, 51, 95] learn discriminative latent embedding vectors for images by “pulling together" views of the same image, and “pushing away" embeddings of different images. Recent non-contrastive methods [15, 46, 111] demonstrate approaches to avoid negative sampling to improve computational efficiency. Clustering methods [4, 12, 13, 14, 68, 108, 112] simultaneously discovers a set of clusters or prototypes, and learns discriminative image embeddings. Contrary to contastive methods, the objective does not have to be approximated as optimizing over the entire set of negative representative clusters is tractable. DeepCluster [12] iteratively performs K-means clustering over the entire dataset and learns an embedding model and classification head to predict the cluster assignment. SeLA [4] presents a principled formulation for clustering and representation learning as a single optimization objective, by casting cluster assignment as an optimal transport problem [29, 65]. SwAV [14] and ODC [112] demonstrate that clustering can be done online per batch to increase learning efficiency.

Dense representation learning Recent clustering-based methods approach dense representation learning as an instance segmentation problem [16, 53, 67, 114] and regional feature correspondence [66, 99, 105]. These methods are purposed for pretraining backbones and generally output small feature maps (e.g. 7x7), in contrast to our method. Similarly to our method, VADeR [90] learns dense representations by contrasting pixel-level embeddings in augmented views. Our method improves on VADeR by allowing training on larger feature maps (512x512 vs. 56x56 px), more views, optimization without a negative sample memory bank, and contextual region masking. Self-supervised object detection [6, 30, 98, 100, 103, 107] learns expressive embeddings for plausible object proposal regions sampled randomly or heuristically [94]. Masked image modeling (MIM) [5, 21, 52, 106] demonstrates strong representation learning capability surpassing contrasting views. However, all these models output low-resolution feature maps. In contrast, our method ViCE generates precise object-fitting semantic partitioning even for high-resolution images.

Unsupervised semantic segmentation Existing works leverage self-supervised clustering approaches to learn coherent semantic groupings from mutual information [56, 83], geometric equivariance [23], and GAN-based approaches [7, 19]. Other works [54, 97] leverages self-supervised depth map estimation [42, 73] for enhancing semantic segmentation performance. Recently, DINO [15] demonstrated that attention maps for semantic objects naturally emerge for self-supervised Vision Transformer (ViT) models [34, 96]. STEGO [47] presents a method to distill features from DINO and achieve SOTA results. Our work improves learning efficiency also on high-resolution images by contrasting cluster assignment over superpixels.

Image decomposition by superpixeliation Prior work which visually groups pixels includes semi- and weakly supervised models [37, 64, 109], and methods bootstrapping from pretrained saliency [39] and contour detector [55, 60, 115] models. We utilize visual grouping without depending on pretraining and not only as an inductive bias, but to perform contrastive learning over a set of visually coherent regions instead of individually meaningless pixels. Ouyang *et al*. [84] uses self-supervised learning to map superpixel regions between augmented views for transferring semantic labels in annotated samples to corresponding regions in unannotated samples. [59, 78] uses superpixels to refine the unsupervised segmentation output. In contrast, our method uses superpixels to learn semantics from high-resolution images without annotated data.

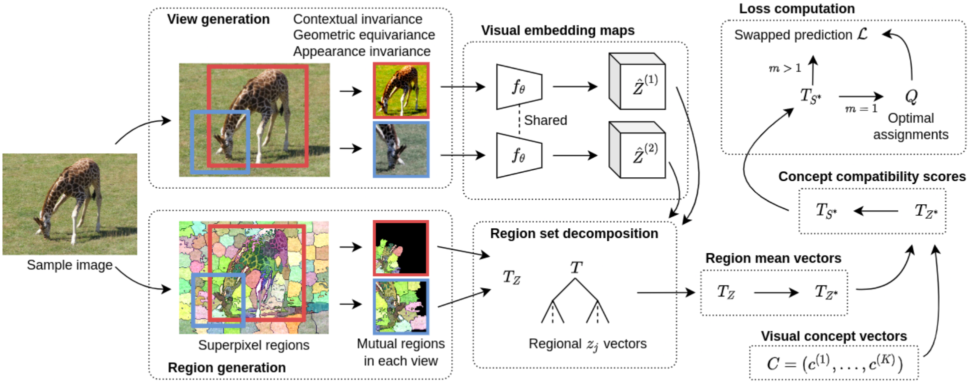

Figure 2: Overview of ViCE. A training iteration starts by generating M augmented views. First, we partition the image into I mutually common superpixel regions. The model f_θ transforms view images into visual concept embedding maps $\hat{Z}^{(m)}$. All vectors z_j are arranged in a tree structure T_Z used to conveniently organize indices of corresponding regions. A mean vector z_i^* is computed for each region. Next, we score each z_i^* in terms of closeness to each concept vector $c^{(k)}$, resulting in region-specific score vectors s_i^*.

3 ViCE: Visual Concept Embeddings

The concept of “the thing in itself" in Kantian philosophy denotes the existence of objects as they are independent of observation. Similarly, one can view natural images perceived by a photometric sensor to be generated from a set of latent semantic visual concepts. We model this process by a model $f(X|Z)$ that generates the observable pixel appearance X of semantic entities represented by a set of latent visual concepts $C = (c^{(1)}, \ldots, c^{(K)})$, encoded into a dense embedding map Z. Our method is based on learning a function f_θ to approximate the inverse mapping $f^{-1}(Z|X)$ while simultaneously discovering the set of latent visual concepts C. The problem of finding the inverse mapping is called vision as inverse graphics [25, 61, 62]. We propose to learn a mapping f_θ that predicts the same visual concept embedding map $Z \in \mathbb{R}^{D \times H \times W}$ with the same spatial resolution as the input image $X \in \mathbb{R}^{3 \times H \times W}$ for all mutually co-occurring abstract pixel patterns generated from augmented views $\tilde{X}^{(m)}$. All views contain one subregion representing the same content, but with different pixel appearances and surrounding context.

$$f_\theta(\tilde{X}^{(m)}) \simeq Z \quad \forall m \in (1, \ldots, M) \tag{1}$$

We relate our approach to discovering semantic meanings for pixels to discovering semantic meanings for words in NLP similar to recent MIM works [6, 21, 106, 116]. Methods to learn semantically rich word embeddings [76, 77, 88] are based on co-occurrence [49] and context [32, 89] of individually meaningless tokens. Each visual concept vector c corresponds to a distinct visual concept primitive or basis vector, and visual concepts are linear combinations of these primitives. The set of concepts C is known and finite, ensuring tractable probabilistic enumeration over possible configuration akin to successful probabilistic language modeling approaches in NLP [32, 91]. We choose to demonstrate our method with the recent SOTA self-supervised learning method SwAV [14] to learn both f_θ and C, though in principle any cluster-based self-supervised method can be used. Fig. 2 shows an overview of our method.

3.1 Decomposing images into visually coherent regions

A high-resolution image contains millions of individually meaningless and mostly redundant pixels. However, it is known that training on high-resolution images is beneficial for learning to segment small objects such as poles and pedestrians [17]. Nevertheless, naively applying self-supervised representation learning methods based on vector comparison on high-resolution embedding maps is inefficient. To solve this problem, we propose to decompose the image into a small set of visually coherent regions using superpixelization [92] and apply representation learning methods to this greatly reduced set of elements. Superpixel methods like Simple Linear Iterative Clustering (SLIC) [1] reduce elements by $\mathcal{O}(1000)$, transforming an image from millions of pixels into less than a thousand regions. We choose SLIC because of advantages [2] such as more uniform region distribution compared to graph-based methods [36]. In contrast to grid decomposition, which is the standard for ViT models [15, 34], superpixels can preserve detail by representing thin and small patches like poles as distinct regions while requiring 75% fewer elements on average with the same base element size. While in this paper our objective is to show that even the simplest form of region decomposition is useful, it is likely that leveraging learning-based superpixelization methods [3, 71, 101] can further improve performance.

3.2 View generation and contextual region masking

We generate augmented views for discerning the latent semantic visual concepts through photometric invariance [20], and geometric equivariance [23]. We introduce region masking as an additional augmentation for contextual invariance shown to improve performance. To generate views with different contexts, we first sample a center point $(x,y)^*$ in the image. Sampling is done in content-rich regions to better satisfy the equipartitioning of concepts assumption [4, 14] for each training batch. We found that probabilistic sampling from a Gaussian filtered Canny edge detection map[11] is a useful measure of image content. Views $\tilde{X}^{(m)}$ are generated by sampling M view centers $(x,y)^{(m)}$ around $(x,y)^*$ while ensuring a mutual image subregion exists. We generate geometrically equivariant views by first sampling a resize coefficient $\beta^{(m)}$ for each view m. β determines the size of the cropped view region as exemplified by the red and blue crop regions in Fig. 2. All view crops are resized to the common view size, thus enforcing the model to learn resolution invariant representations. All views are randomly flipped horizontally. All views are augmented by random color distortion and Gaussian blurring before normalization to learn appearance invariant visual concepts [20, 102, 104]. A ratio of superpixel regions is masked with noise as a means to learn robust features and alleviate the shortcut learning problem [40]. We provide the view generation algorithm as pseudocode in the Supplementary.

3.3 Learning algorithm

The objective $\mathcal{L}_{cl}$ is designed to simultaneously learn the mapping function f_θ in Eq. 1, and optimize the distribution of latent visual concepts C. The algorithm can be viewed as an extension of SwAV [14] to the problem of learning dense embedding maps. We refer to prior work for an explanation of SwAV [4, 14, 29, 58]. The rest of this section explains the flow of a training iteration as visualized in Fig. 2. We provide pseudocodes in the Supplementary.

A training iteration starts by partitioning an image $X^{(n)} \in \mathbb{R}^{3\times H\times W}$ with height H and width W into a superpixel region map $A^{(n)} \in \mathbb{R}^{H\times W}$, with integer values specifying every pixel's region index. Next, a set of M augmented views $\tilde{X}^{(n)} = \{\tilde{X}^{(1,n)}, \dots, \tilde{X}^{(M,n)}\}$ and corresponding superpixel map crops $\tilde{A}^{(n)} = \{\tilde{A}^{(1,n)}, \dots, \tilde{A}^{(M,n)}\}$ of size h and w are generated for each image as explained in Sec. 3.2. $\tilde{A}^{(n)}$ is processed to contain only mutual regions existing in all views. The learned function f_θ transforms $\tilde{X}^{(n)}$ into a normalized visual embedding tensor $\hat{Z}^{(n)} \in \mathbb{R}^{D\times h\times w}$. Next $\hat{Z}^{(n)}$ is decomposed region-wise into row vectors $z_j \in \mathbb{R}^D$ and stored in a tree structure T_Z used to conveniently organize indices of corresponding regions i in view m of image n. Vectors of non-mutual regions are discarded. A single mean vector $z^{(i,m,n)*}$ is computed to represent each region i and stored in T_{Z^*}. Each vector $z^{(i,m,n)*}$ is scored in terms of compatibility or closeness to each visual concept vector $C = (c^{(1)}, \dots, c^{(K)})$ by computing the following matrix product

$$s^* = (z^*)^T C \tag{2}$$

with $C \in \mathbb{R}^{D\times K}$ represented as an optimizable weight matrix. Note that the dot product $z \cdot c$ equals the cosine distance as both vectors are normalized. All regional score vectors $s^{(i,m,n)*}$ are stored in a tree structure T_{S^*}. The concept assignments $q^{(i)}$ are determined by optimally distributing $s^{(i,m,n)*}$ uniformly over all concepts $c^{(k)}$ so that the overall compatibility between all $s^{(i)}$ and $c^{(k)}$ are maximized for regions in the primary view $m = 1$ [14]. We compute $q^{(i)}$ efficiently by the Sinkhorn-Knopp algorithm [4, 29]. A FIFO queue of accumulated $s^{(i,1,n)*}$ vectors is used to improve the empirical approximation of a uniform distribution of concepts [4, 14]. The swapped prediction learning objective [14] is

$$\mathcal{L}_{cl} = -\frac{1}{N(M-1)} \sum_{n=1}^{N} \sum_{m=2}^{M} \frac{1}{I} \sum_{i=1}^{I} q^{(i)} log\, \sigma\left(\tfrac{1}{\tau} s^{(i,m)*}\right) \tag{3}$$

where $\sigma()$ is the softmax function and τ is temperature. Two normalized embeddings $z^{(a)}$ and $z^{(b)}$ are compared for semantic similarity using the dot product. This operation is equivalent to comparing two word embeddings by cosine distance [76, 77].

4 Experiments

We implement ViCE in the self-supervised learning framework VISSL [45] based on PyTorch [85]. The quality of learned embeddings are evaluated on the COCO-Stuff164k [10, 69] reduced to 27 classes [56] and the Cityscapes [27] benchmark datasets. We use the framework MMSegmentation [26] for evaluation and visualization. Our comparative baseline for dense representation learning is the SOTA unsupervised semantic segmentation CNN model PiCIE [23] based on DeepCluster [12]. We experiment with ResNet 18 and 50 backbones [50] and two decoder architectures; the SOTA model DeepLabV3+ (DLV3+) [18] for high-resolution images, and the Feature Pyramid Network (FPN) [70] used in our baseline.

We evaluate the semantic richness and spatial accuracy of the resulting embedding maps using clustering and linear models. For unsupervised semantic segmentation we compute a set of K clusters based on output embeddings using FAISS [57]. Each cluster is greedily assigned the majority label class, or optimally assigned by the Hungarian matching algorithm [48] to cover all classes. For linear model evaluation, we train a 1×1 convolution

Table 1: Representation quality experiment results on low- and high-resolution images.

Model		mIoU	Acc.	Model		mIoU	Acc.
	COCO				*Cityscapes*		
ResNet50 [50]	C 27	8.9	24.60	ResNet50 [50]	C 27	-	-
MoCoV2 [22]	C 27	10.40	9.60	MoCoV2 [22]	C 27	-	-
DINO* [15]	C 27	9.60	30.50	DINO* [15]	C 27	-	-
IIC [56]	C 27	6.71	21.79	IIC [56]	C 27.	6.35	47.88
PiCIE [23]	C 27	13.84	48.09	PiCIE [23]	C 27	12.31	65.50
	C $27^{\diamond}$	14.60	48.37		C $27^{\diamond}$	11.85	64.29
	C $27^{\star}$	9.27	38.31		C $27^{\star}$	8.80	82.48
	C $128^{\star}$	10.75	49.81		C $128^{\star}$	7.97	56.52
	C $256^{\star}$	12.42	**66.02**		C $256^{\star}$	12.71	**89.86**
	Linear	14.77	54.75		Linear	-	-
PiCIE+H [23]	C 27+100	14.40	50.0	PiCIE+H [23]	C 27+100	-	-
ViCE (low-res)	C 27	11.40	28.91	ViCE (low-res)	C 27	12.81	31.87
	C $27^{\star}$	11.55	50.49		C $27^{\star}$	19.52	80.34
	C $128^{\star}$	16.66	52.33		C $128^{\star}$	21.48	81.55
	C $256^{\star}$	17.98	54.92		C $256^{\star}$	21.24	81.72
	Linear	25.49	62.78		Linear	**31.55**	86.33
				No pretrain	Linear	24.84	82.99
ViCE (high-res)	C $256^{\star}$	**21.77**	64.75	ViCE (high-res)	C $256^{\star}$	**25.23**	84.28
	Linear	**29.38**	**68.16**		Linear	30.40	**87.0**
STEGO* [47]	C 27	28.20	56.90	STEGO* [47]	C 27	21.00	73.20
	Linear	41.00	76.10		Linear	-	-

layer without a nonlinear activation function. All models are trained and evaluated on separate train and validation sets. Note that the visual concepts learned by ViCE during training are not used for evaluation, and it is therefore fair to compare ViCE and baseline performance as long as the number of clusters is the same in both evaluation models.

We conduct experiments on 32 V100 32 GB GPUs. Each GPU loads four images, and generates five augmented views. High- and low-resolution views correspond to 512×512 pixels and 256×256 pixels, respectively. The resulting total batch size is 128 images with 640 views. To generating superpixels, we use SLIC [1] implemented in OpenCV [8] with average region size 20 px. Maximal mask coverage is 25 %. The view resize coefficients β are sampled between 0.5 to 2. The embedding dimension D and the number of visual concepts C are 128. We use the same set of hyperparameters in all experiments. A hyperparameter study is given in the Supplementary. Parameters for the objective $\mathcal{L}_{cl}$ are the same as SwAV [14]. The FIFO queue consists of 5K score vectors s^* per GPU. The model is optimized using the LARS optimizer [110] with weight decay 10^{-6}. The learning rate (LR) schedule is linear warmup followed by cosine decay [72, 79]. We set the peak LR using the linear LR scaling rule [43] with a base LR 0.04 for a single 4 GPU node. We initialize models with the default PyTorch pretrained weights obtained by training on ImageNet [31] for 600 epochs. However, our method can learn from random initialization as shown in Table 1. Timing information is given in the Supplementary.

Table 2: Performance of best models trained on high- and low-resolution images

Dataset	Resolution	Configuration	Cluster mIoU	Linear mIoU
COCO	Low	RN50, FPN	19.37	27.63
	High	RN50, DLV3+	**21.77**	**29.38**
Cityscapes	Low	RN18, FPN	21.48	**31.55**
	High	RN18, DLV3+	**25.23**	30.40

4.1 Representation quality experiments

Table 1 presents results on low-resolution image experiments. C K denotes evaluation with K clusters, $\diamond$ denotes reproduced results with optimal cluster assignment, $\star$ denotes greedy assignment, and $*$ denotes ViT-based models. The best CNN-based cluster and linear model results are written in bold. Both ViCE (low-res) and PiCIE [23] use the same ResNet 18 backbone, FPN decoder, and 320×320 px image downsampling procedure for fair comparison. All ViCE models are trained for 4 epochs for COCO, and 24 epochs for Cityscapes, respectively. We trained and evaluated our PiCIE models using the official code [23]. Our high-resolution and overclustered model achieves SOTA results on Cityscapes, and on COCO for convolutional models. The generic image COCO results show that ViCE is adept at discovering concepts using overclustering [38]. We believe this property stems from online clustering being more stable than offline clustering methods [14, 112]. The Cityscapes results show ViCE improving on PiCIE in all experiments. ViCE performs better than the SOTA ViT-based model STEGO [47] on Cityscapes with high-resolution and overclustering. We trained our best high-resolution C 256* COCO model in 64 h and the equivalent PiCIE model in 52 h. Fig. 1, 3 shows clustering output visualizations. Table 2 shows that the best high-resolution models improves on the best low-resolution models evaluated on high-resolution images. Note that effectively training on high-resolution images is made possible by superpixelization. Results for varying superpixel sizes and performance are given in the Supplementary.

4.2 Ablation studies

The upper section of Table 3 provides an ablation study for low-resolution images evaluated by a linear model. The first column represents the baseline ViCE model using an RN18 backbone and FPN decoder [70] without region decomposition. The second columns indicate gains from random masking. The third and fourth column shows gains from applying grid and superpixel region decomposition. The final column indicates that utilizing the more complex DLV3+ decoder [18] is detrimental in the case of low-resolution images. We speculate this is because atrous convolutions in high-resolution decoders skip relevant neighboring information in tiny feature maps. The first column in the bottom section of Table 3 is empty, as learning dense embeddings for high-resolution images without superpixelization is computationally intractable. The second column showcase the radical difference in using superpixelization. The third column demonstrates the importance of utilizing a high-resolution decoder. The final column shows how superpixels are better than grids with equivalent base element sizes.

Table 3: Representation quality ablation study on low- and high-resolution images.

Low-resolution Cityscapes

	FPN 1px	Masking	Grid 10px	Super 10px	DLV3+
mIoU	29.66	30.42	31.30	**31.55**	11.56
Time	34h 4min	31h 6min	5h 31min	**5h 31min**	5h 37min

High-resolution Cityscapes

	FPN 1px	FPN super 20px	DLV3+ grid 20 px	DLV3+ super 20px
mIoU	-	8.98	25.53	**29.38**
Time	92h 20min (est.)	4h 55min	10h 1min	**6h 16min**

Table 4: Domain generalization performance

Training data domain	Evaluation data domain	mIoU	aAcc
Cityscapes	Cityscapes	30.40	**87.00**
COCO	Cityscapes	**34.14**	86.10

4.3 Domain Generalization experiment

In Table 4 we show how ViCE benefits when learning from a large general visual domain. Training on COCO and evaluating on Cityscapes with a linear model increases performance from 30.40 to 34.14 (+3.74) mIoU by improving the distinctiveness of complex classes like "Traffic sign". Our findings show that general vision models can learn more useful features compared to narrow vision models even when applied in the narrow domain. The recent SOTA model STEGO [47] similarly uses a backbone trained on ImageNet only.

4.4 Qualitative evaluation

Fig. 4 visualizes dense embedding maps to demonstrate how ViCE discovers distinct semantic visual entities or concepts from natural images without human supervision or proposals heuristics [6, 94]. For example, persons are represented differently from the ground surface, and human faces and bodies are semantically similar. We visualize embedding maps by PCA dimensionality reduction [87] and scale each z to the RGB range.

5 Conclusion

We present a new SOTA self-supervised unsupervised semantic segmentation method ViCE for learning to generate dense embedding maps. Our experiments quantitatively demonstrate that decomposing images by superpixelization improves the effectiveness of classification-based self-supervised methods, particularly for high-resolution images, and also achieves better performance than conventional grid decomposition. We hope our work will raise interest in further incorporating non-uniform image decomposition techniques to improve self-supervised computer vision methods including ViT-based models like DINO [15] and other dense representation learning methods [66, 90, 99, 105].

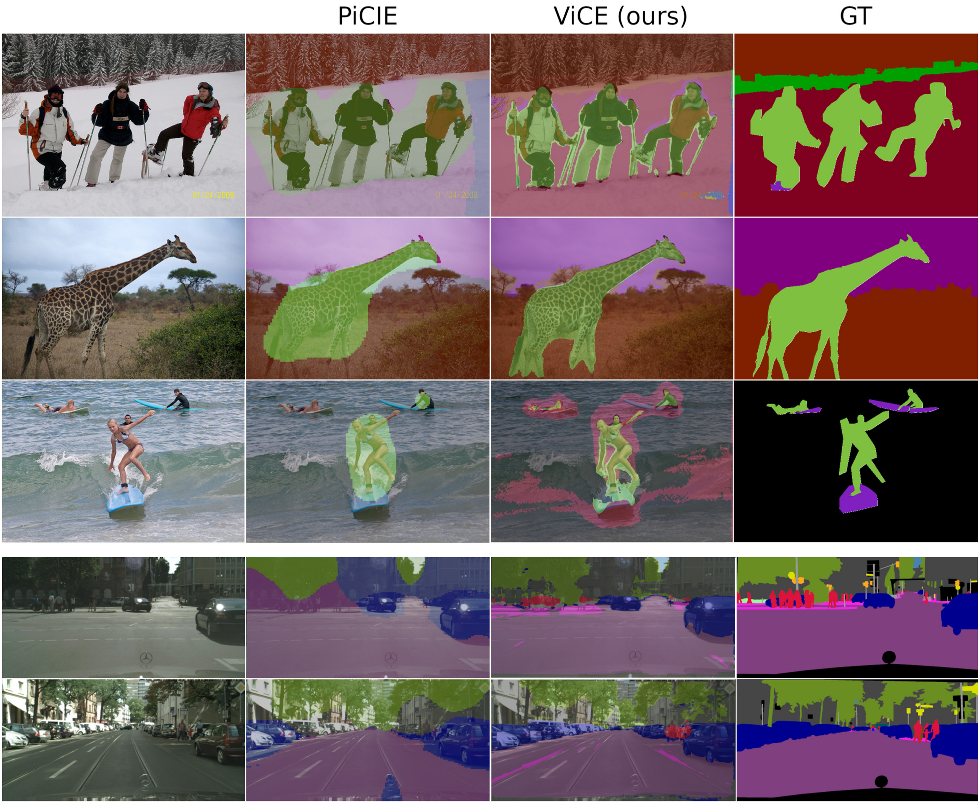

Figure 3: Output cluster visualizations on COCO (top) and Cityscapes (bottom).

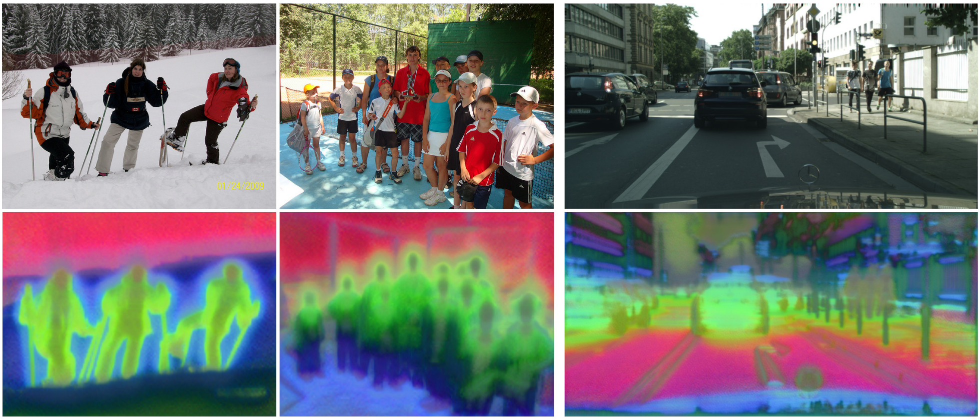

Figure 4: Dense embedding maps visualized as RGB images.

Acknowledgements

This work was financially supported by JST SPRING, Grant Number JPMJSP2125. The authors would like to take this opportunity to thank the "Interdisciplinary Frontier Next-Generation Researcher Program of the Tokai Higher Education and Research System".

The work was financially supported by JSPS KAKENHI, Grant Number 21H04892.

This research was supported by Program on Open Innovation Platform with Enterprises, Research Institute and Academia, Japan Science and Technology Agency (JST, OPERA, JPMJOP1612).

The computation was carried out through the "General Projects" program on the supercomputer "Flow" at the Information Technology Center, Nagoya University.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels. In *EPFL Technical Report*, volume 149300, 2010.

[2] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[3] Pablo Arbelaez, Jordi Pont-Tuset, Jon Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.

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

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2022.

[6] Amir Bar, Xin Wang, Vadim Kantorov, Colorado Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. DETReg: Unsupervised pre-training with region priors for object detection. 2022.

[7] Adam Bielski and Paolo Favaro. Emergence of object segmentation in perturbed generative models. In *NeurIPS*, volume 32, 2019.

[8] Gary Bradski. The OpenCV Library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000.

[9] Silvia Bucci, Antonio D'Innocente, Yujun Liao, Fabio Maria Carlucci, Barbara Caputo, and Tatiana Tommasi. Self-supervised learning across domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[10] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018.

[11] John F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, PAMI-8(6):679–698, 1986.

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

[13] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *ICCV*, pages 2959–2968, 2019. doi: 10.1109/ICCV.2019.00305.

[14] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, volume 33, 2020.

[15] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021.

[16] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. *ICCV*, pages 7526–7534, 2021.

[17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40:834–848, 2018.

[18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 833–851, 2018.

[19] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In *NeurIPS*, volume 32, pages 12705–12716, 2019.

[20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.

[21] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv*, 2022.

[22] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. In *ArXiv*, 2020.

[23] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *CVPR*, pages 16794–16804, 2021.

[24] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[25] Moreno Comellas. *Vision as inverse graphics for detailed scene understanding*. PhD thesis, University of Edinburgh, July 1956.

[26] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

[27] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[28] Yann Le Cunn. Self-supervised learning (keynote talk). AAAI, 2020.

[29] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, volume 26, pages 2292–2300, 2013.

[30] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. UP-DETR: Unsupervised pre-training for object detection with transformers. In *CVPR*, pages 1601–1610, 2021.

[31] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.

[33] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015.

[34] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[35] Alexei A. Efros. Self-supervision for learning from the bottom up (invited talk). ICLR, 2021.

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

[37] Gianni Franchi, Nacim Belkhir, Mai Lan Ha, Yufei Hu, Andrei Bursuc, Volker Blanz, and Angela Yao. Robust semantic segmentation with superpixel-mix. In *BMVC*, 2021.

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

[39] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *ICCV*, 2021.

[40] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

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

[42] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019.

[43] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *ArXiv*, abs/1706.02677, 2017.

[44] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, and Piotr Bojanowski. Self-supervised pretraining of visual features in the wild. *ArXiv*, abs/2103.01988, 2021.

[45] Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. VISSL. https://github.com/facebookresearch/vissl, 2021.

[46] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, volume 33, pages 21271–21284, 2020.

[47] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondances. In *ICLR*, Apr. 2022.

[48] Kuhn W. Harold. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[49] Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954. doi: 10.1080/00437956.1954.11659520.

[50] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[51] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

[52] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CVPR*, 2022.

[53] Olivier J. H'enaff, Skanda Koppula, Jean-Baptiste Alayrac, Aäron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. *ICCV*, pages 10066–10076, 2021.

[54] 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 *CVPR*, pages 11130–11140, 2021.

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

[56] Xu Ji, João F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, pages 9865–9874, 2019.

[57] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[58] Tim Kaiser and Nikolas Adaloglou. Understanding SwAV: Self-supervised learning with contrasting cluster assignments. https://theaisummer.com/swav/, 2021.

[59] Asako Kanezaki. Unsupervised image segmentation by backpropagation. In *ICASSP*, 2018.

[60] Tsung-Wei Ke, Jyh-Jing Hwang, Yunhui Guo, Xudong Wang, and Stella Yu. Unsupervised hierarchical semantic segmentation with multiview cosegmentation and clustering transformers. *CVPR*, 2022.

[61] Daniel Kersten and Alan Yuille. Vision as bayesian inference: analysis by synthesis? *Trends Cogn Sci.*, 10(7), 2006.

[62] Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as bayesian inference. *Annual Review of Psychology*, 55(1):271–304, 2004.

[63] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, volume 25, pages 1097–1105, 2012.

[64] Suha Kwak, Seunghoon Hong, and Bohyung Han. Weakly supervised semantic segmentation using superpixel pooling network. In *AAAI*, 2017.

[65] Bruno Lévy and Erica L. Schwindt. Notions of optimal transport theory and how to implement them on a computer. *Computers & Graphics*, 72:135–148, 2018.

[66] Xiaoni Li, Y. Zhou, Yifei Zhang, Aoting Zhang, Wei Wang, Ning Jiang, Haiying Wu, and Weiping Wang. Dense semantic contrast for self-supervised visual representation learning. 2021.

[67] Xiaoni Li, Y. Zhou, Yifei Zhang, Aoting Zhang, Wei Wang, Ning Jiang, Haiying Wu, and Weiping Wang. Dense semantic contrast for self-supervised visual representation learning. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[68] Yunfan Li, Hu Peng, Liu Zitao Peng Dezhong, Tianyi Zhou, and Peng Xi. Unsupervised semantic segmentation by contrasting object mask proposals. In *AAAI*, 2021.

[69] 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 *ECCV*, pages 740–755. Springer, 2014.

[70] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, 2017.

[71] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, 2020.

[72] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

[73] Juan Luis, Gonzalez Bello, and Munchurl Kim. Forget about the LiDAR: Self-supervised depth estimators with MED probability volumes. In *NeurIPS*, volume 33, pages 12626–12637, 2020.

[74] Gary Marcus and Ernest Davis. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon Books, USA, 2019. ISBN 1524748250.

[75] Douglas L. Medin and Marguerite M. Schaffer. Context theory of classification learning. *Psychological Review*, 85:207–238, 1978.

[76] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[77] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[78] Seyyed Ehsan Mirsadeghi, Ali Royat, and Hamid Rezatofighi. Unsupervised image segmentation by mutual information maximization and adversarial regularization. *IEEE Robotics and Automation Letters (RA-L)*, 6:6931–6938, 2021.

[79] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, pages 6706–6716, 2020.

[80] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016.

[81] Robert M. Nosofsky. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, 115:39–57, 1986.

[82] Robert M. Nosofsky, John K. Kruschke, and Stephen C. McKinley. Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology*, 18:211–233, 1992.

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

[84] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *ECCV*, 2020.

[85] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, pages 8026–8037, 2019.

[86] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. doi: 10.1109/CVPR.2016.278.

[87] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, 1901.

[88] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[89] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, pages 2227–2237, Jun. 2018.

[90] Pedro H. O. Pinheiro, Amjad Almahairi, Ryan Y. Benmalek, Florian Golemo, and Aaron C. Courville. Unsupervised learning of dense visual representations. In *NeurIPS*, 2020.

[91] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. In *Open AI*, 2018.

[92] Xiaofeng Ren and Jitendra Malik. Learning a classification model for segmentation. In *ICCV*, volume 1, pages 10–17, 2003.

[93] Eleanor H. Rosch. Natural categories. *Cognitive Psychology*, 4(3):328–350, 1973. doi: https://doi.org/10.1016/0010-0285(73)90017-0.

[94] Jasper R.R. Uijlings, Koen E.A. Van de Sande, Theo Gevers, and Arnold W.M. Smeulders. Selective search for object recognition. *IJCV*, 104:154–171, 2013. doi: 10.1007/s11263-013-0620-5.

[95] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

[96] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, volume 30, page 6000–6010, 2017.

[97] Tuan-Hung Vu, Himalaya Jain, Max Bucher, Matthieu Cord, and Patrick Pérez. DADA: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, pages 7364–7373, 2019.

[98] Xinlong Wang, Zhang Rufeng, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, pages 3024–3033, 2021.

[99] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021.

[100] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. In *NeurIPS*, volume 34, 2021.

[101] Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Learning super-features for image retrieval. In *ICLR*, 2022.

[102] Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *ICML*, volume 139, pages 11112–11122, 2021.

[103] Tete Xiao, Colorado J. Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *ICCV*, pages 10539–10548, 2021.

[104] Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *ICLR*, 2021.

[105] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *CVPR*, 2021.

[106] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *CVPR*, 2022.

[107] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Ching-Feng Lin. Instance localization for self-supervised detection pretraining. In *CVPR*, pages 3987–3996, 2021.

[108] Linxiao Yang, Ngai-Man Cheung, Jiaying Li, and Jun Fang. Deep clustering by Gaussian mixture variational autoencoders with graph embedding. In *ICCV*, pages 6439–6448, 2019. doi: 10.1109/ICCV.2019.00654.

[109] Sheng Yi, Huimin Ma, Xiang Wang, Tianyu Hu, Xi Li, and Yu Wang. Weakly-supervised semantic segmentation with superpixel guided local and global consistency. *Pattern Recognition*, 124:108504, 2022.

[110] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *ArXiv*, abs/1708.03888, 2017.

[111] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, pages 12310–12320, 2021.

[112] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Loy. Online deep clustering for unsupervised representation learning. In *CVPR*, pages 6688–6697, 2020.

[113] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV*, pages 649–666. Springer, 2016.

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

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

[116] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Loddon Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *ICLR*, 2022.