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

Robin Karlsson¹
karlsson.robin@g.sp.m.is.nagoya-u.ac.jp
Tomoki Hayashi¹
hayashi.tomoki@g.sp.m.is.nagoya-u.ac.jp
Keisuke Fujii¹
fujii@i.nagoya-u.ac.jp
Alexander Carballo¹
alexander@g.sp.m.is.nagoya-u.ac.jp
Kento Ohtani¹
ohtani.kento@g.sp.m.is.nagoya-u.ac.jp
Kazuya Takeda¹,²
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 \( O(1000) \) 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 [63, 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 Visual Concept Embedding (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 [1, 3, 11, 50, 52, 75]. 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 [20, 24]. 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 contrastive 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]. SwA V [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 [22] 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 [6, 15]. Other works [6, 15] 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 superpixelation** Prior work which visually groups pixels includes semi- and weakly supervised models [10, 13, 114], and methods bootstrapping from pretrained saliency [23] 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 $f_\theta$ 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 $s_i^{*k}$.

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 $f_\theta$ 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 $f_\theta$ 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.

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

We relate our approach to discovering semantic meanings for pixels to discovering semantic meanings for words in NLP similar to recent MIM works [5, 21, 96, 106]. 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 [53, 84]. We choose to demonstrate our method with the recent SOTA self-supervised learning method SwAV [43] to learn both $f_\theta$ 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 $O(1000)$, 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 $m$. $\beta$ 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 $L_{cl}$ is designed to simultaneously learn the mapping function $f_\theta$ in Eq. 1, and optimize the distribution of latent visual concepts $C$. The algorithm can be viewed as an extension of SwAV [44] to the problem of learning dense embedding maps. We refer to prior
work for an explanation of SwAV [4, 14, 29]. 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 \( \tilde{X} = \{ \tilde{X}^{(1,n)}, \ldots, \tilde{X}^{(M,n)} \} \) and corresponding superpixel map crops \( \tilde{A}^{(n)} = \{ \tilde{A}^{(1,n)}, \ldots, \tilde{A}^{(M,n)} \} \) 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 \( f_\theta \) transforms \( \tilde{X}^{(n)} \) into a normalized visual embedding tensor \( \tilde{Z}^{(n)} \in \mathbb{R}^{D \times h \times w} \). Next \( \tilde{Z}^{(n)} \) 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 \( \tilde{z}^{(i,m,n)} \) is computed to represent each region \( i \) and stored in \( T_Z \). Each vector \( \tilde{z}^{(i,m,n)} \) is scored in terms of compatibility or closeness to each visual concept vector \( C = (c^{(1)}, \ldots, c^{(K)}) \) by computing the following matrix product

\[
s^* = (\tilde{z}^*)^T C
\]

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 \( c^{(k)} \) is maximized for regions in the primary view \( m = 1 \) [4, 14]. We compute \( q^{(i)} \) efficiently by the Sinkhorn-Knopp algorithm [4, 14]. 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

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

where \( \sigma() \) is the softmax function and \( \tau \) 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 [14, 14].

4 Experiments


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 [14]. Each cluster is greedily assigned the majority label class, or optimally assigned by the Hungarian matching algorithm [14] to cover all classes. For linear model evaluation, we train a \( 1 \times 1 \) convolution
Table 1: Representation quality experiment results on low- and high-resolution images.

<table>
<thead>
<tr>
<th>Model</th>
<th>COCO</th>
<th>Cityscapes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mIoU</td>
<td>Acc.</td>
</tr>
<tr>
<td>ResNet50 [50]</td>
<td>C 27</td>
<td>8.9</td>
</tr>
<tr>
<td>MoCoV2 [22]</td>
<td>C 27</td>
<td>10.40</td>
</tr>
<tr>
<td>DINO* [15]</td>
<td>C 27</td>
<td>9.60</td>
</tr>
<tr>
<td>IIC [56]</td>
<td>C 27</td>
<td>6.71</td>
</tr>
<tr>
<td>PiCIE [23]</td>
<td>C 27</td>
<td>13.84</td>
</tr>
<tr>
<td></td>
<td>C 27</td>
<td>14.60</td>
</tr>
<tr>
<td></td>
<td>C 27</td>
<td>9.27</td>
</tr>
<tr>
<td></td>
<td>C 27</td>
<td>10.75</td>
</tr>
<tr>
<td></td>
<td>C 27</td>
<td>12.42</td>
</tr>
<tr>
<td></td>
<td>C 27</td>
<td>14.77</td>
</tr>
<tr>
<td></td>
<td>C 128</td>
<td>16.66</td>
</tr>
<tr>
<td></td>
<td>C 256</td>
<td>17.98</td>
</tr>
<tr>
<td>Linear</td>
<td>C 256</td>
<td>25.49</td>
</tr>
<tr>
<td></td>
<td>C 256</td>
<td>29.38</td>
</tr>
<tr>
<td>ViCE (low-res)</td>
<td>C 27</td>
<td>11.40</td>
</tr>
<tr>
<td></td>
<td>C 27</td>
<td>11.55</td>
</tr>
<tr>
<td></td>
<td>C 128</td>
<td>16.66</td>
</tr>
<tr>
<td></td>
<td>C 256</td>
<td>17.98</td>
</tr>
<tr>
<td>Linear</td>
<td>C 256</td>
<td>25.49</td>
</tr>
<tr>
<td>ViCE (high-res)</td>
<td>C 27</td>
<td>12.81</td>
</tr>
<tr>
<td></td>
<td>C 27</td>
<td>19.52</td>
</tr>
<tr>
<td></td>
<td>C 128</td>
<td>21.48</td>
</tr>
<tr>
<td></td>
<td>C 256</td>
<td>22.48</td>
</tr>
<tr>
<td>Linear</td>
<td>C 256</td>
<td>31.55</td>
</tr>
<tr>
<td>STEGO* [47]</td>
<td>C 27</td>
<td>28.20</td>
</tr>
<tr>
<td></td>
<td>C 27</td>
<td>21.00</td>
</tr>
<tr>
<td>Linear</td>
<td>C 27</td>
<td>30.40</td>
</tr>
</tbody>
</table>

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 [1]. The FIFO queue consists of 5K score vectors $s^*$ per GPU. The model is optimized using the LARS optimizer [100] 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 [100] 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Resolution</th>
<th>Configuration</th>
<th>Cluster mIoU</th>
<th>Linear mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCO</td>
<td>Low</td>
<td>RN50, FPN</td>
<td>19.37</td>
<td>27.63</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>RN50, DLV3+</td>
<td><strong>21.77</strong></td>
<td><strong>29.38</strong></td>
</tr>
<tr>
<td>Cityscapes</td>
<td>Low</td>
<td>RN18, FPN</td>
<td>21.48</td>
<td>31.55</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>RN18, DLV3+</td>
<td><strong>25.23</strong></td>
<td>30.40</td>
</tr>
</tbody>
</table>

4.1 Representation quality experiments

Table 1 presents results on low-resolution image experiments. C K denotes evaluation with K clusters, ◊ denotes reproduced results with optimal cluster assignment, * 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 [10, 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.

<table>
<thead>
<tr>
<th></th>
<th>Low-resolution Cityscapes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPN 1px Masking Grid 10px Super 10px DLV3+</td>
</tr>
<tr>
<td>mIoU</td>
<td>29.66 30.42 31.30 31.55 11.56</td>
</tr>
<tr>
<td>Time</td>
<td>34h 4min 31h 6min 5h 31min 5h 37min</td>
</tr>
<tr>
<td></td>
<td>High-resolution Cityscapes</td>
</tr>
<tr>
<td></td>
<td>FPN 1px FPN super 20px DLV3+ grid 20 px DLV3+ super 20px</td>
</tr>
<tr>
<td>mIoU</td>
<td>- 8.98 25.53 29.38 29.38</td>
</tr>
<tr>
<td>Time</td>
<td>92h 20min (est.) 4h 55min 10h 1min 6h 16min</td>
</tr>
</tbody>
</table>

Table 4: Domain generalization performance

<table>
<thead>
<tr>
<th>Training data domain</th>
<th>Evaluation data domain</th>
<th>mIoU</th>
<th>aAcc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cityscapes</td>
<td>Cityscapes</td>
<td>30.40</td>
<td>87.00</td>
</tr>
<tr>
<td>COCO</td>
<td>Cityscapes</td>
<td>34.14</td>
<td>86.10</td>
</tr>
</tbody>
</table>

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, JP-MJOP1612).

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

References


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


