ViCE: Self-Supervised Visual Concept Embeddings as Contextual and Pixel Appearance Invariant Semantic Representations

Robin Karlsson¹, Tomoki Hayashi¹, Keisuke Fujii¹, Alexander Carballo¹, Kento Ohtani¹, Kazuya Takeda¹,²
¹Nagoya University, ²Tier IV

Abstract

This work presents a self-supervised method to learn dense semantically rich visual concept embeddings for images inspired by methods for learning word embeddings in NLP. Our method improves on prior work by generating more expressive embeddings and by being applicable for high-resolution images. Viewing the generation of natural images as a stochastic process where a set of latent visual concepts give rise to observable pixel appearances, our method is formulated to learn the inverse mapping from pixels to concepts. Our method greatly improves the effectiveness of self-supervised learning for dense embedding maps by introducing superpixelization as a natural hierarchical step up from pixels to a small set of visually coherent regions. Additional contributions are regional contextual masking with nonuniform shapes matching visually coherent patches and complexity-based view sampling inspired by masked language models. The enhanced expressiveness of our dense embeddings is demonstrated by significantly improving the state-of-the-art representation quality benchmarks on COCO (+12.94 mIoU, +87.6%) and Cityscapes (+16.52 mIoU, +134.2%). Results show favorable scaling and domain generalization properties not demonstrated by prior work.

1. Introduction

Deep learning is recognized as the most potent modelling tool available for representation learning on unstructured data [6]. The universal approximation theorem theoretically proves that deep neural networks (DNN) unbounded in either depth [70] or width [54] can approximate any function arbitrarily well.

Progress in state-of-the-art (SOTA) performance on general computer vision tasks in the last decade has been based on supervised learning using relatively large datasets annotated with semantic information by human labelers [61]. Despite a decade of progress, arguments are made that the original promise of generalizable and robust computer vision deep learning models has not yet been achieved and
that the necessity of increasing the order of magnitude of labelled data is unsustainable in practice [27, 72]. Additionally, arguments can be made that learning from top-down categorization (i.e. “what it is”) from semantically vague and inconsistent human annotation could be limiting our pursuit of robust computer vision [36], and that instead learning through bottom-up association (i.e. “what it is like in a context”) is more akin to how visual concepts emerge for humans as supported by cognitive science [74,80,81,92] and similar to how word embeddings are learned in natural language processing (NLP) [50,75,76] as well as a motivation for capsule networks [93].

In NLP, self-supervised learning on massive unlabeled datasets approximating the real-world distribution of natural language sentences [38] is recognized as the de facto approach for leveraging the universal function approximation properties of DNNs, leading to the recent breakthrough training massive language models [9,32,95]. The crucial component that makes self-supervised learning successful in NLP is the fact that probabilistic enumeration of possible configuration spaces of natural sentences is computationally tractable and proved to be a highly useful learning signal source [2,23,76].

On the other hand, in the case of computer vision, high-resolution visual images consist of millions of high-dimensional and semantically meaningless pixels, making probabilistic enumeration over all possible configuration spaces computationally intractable and therefore limit transferability of contextual predictive self-supervised approaches known to be highly successful in NLP [63]. We propose that obtaining a means to partition an image into a small set of distinct regions encoded by a set of distinct and expressive semantic visual concept embeddings, analogous to how words in sentences are represented, is a necessary first step for unifying computer vision with NLP.

This work presents a novel method inspired by transferring principles for learning word embeddings [75,76,86] to the image domain. We devise how to train a model to represent images as a semantically rich embedding map partitioned into distinct, coherent regions, represented by a latent visual concept embedding (ViCE), similarly to how semantically rich word embeddings are discovered for words in the context of natural sentences. Essential aspects of our method are illustrated in Fig. 1, along with an embedding map visualization. Our working hypothesis is that there exists a strong analogy between how image context defines the meaning of individually semantically meaningless pixel regions and how sentence context defines the meaning of individually semantically meaningless categorical word tokens [50]. Viewing the generation of natural images as a stochastic process where a set of latent visual concepts give rise to observable pixel appearances, we formulate our method to learn the inverse mapping from observed pixels to latent visual concepts through self-supervised learning.

The contextual supervisory signal for learning word embeddings in NLP have been mentioned before as a conceptual motivator for pretext tasks for self-supervised computer vision pretraining methods [34]. However, to the best of our knowledge, our method is the first to consider learning dense visual embedding maps with the explicit intent to be used as input representations for downstream task models. By demonstrating the feasibility of representing images in terms of a small set of regions encoded by a set of distinct semantic visual concept embeddings, similarly to how semantic words embeddings partition sentences, we contribute towards realizing tractable probabilistic enumeration of configuration spaces for images and as a practical solution to the symbolic grounding problem [49] in vision. We hope our contribution will inspire further effort towards increasing the transferability of successful probabilistic methods from NLP to the visual domain and ultimately result in a similar breakthrough in self-supervised computer vision as the one experienced in NLP.

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 step up from pixels to a small set of visually coherent regions for improving the effectiveness and scalability of self-supervised learning for dense embedding maps. Further additions are contextual masking regions with nonuniform shapes matching visually coherent patches and complexity-based view sampling.

- Improve on the current SOTA method PiCIE [22] for unsupervised semantic segmentation by displaying superior representation quality benchmark results on COCO (+12.94 mIoU, +87.6%) and Cityscapes (+16.52 mIoU, +134.2%) and extending applicability to high-resolution images.

2. Related work

2.1. Self-Supervised representation learning

Many early works in applying self-supervised methods for computer vision focus on learning to solve pretext tasks as a substitute for learning from human annotations. Exemplary methods include patch prediction [34,84], jigsaw puzzles [10,79], image rotation prediction [41], and image colorization [120]. However, recent work [20,52] demonstrates that image-level embedding classification with cross-entropy minimization on large datasets [31,66] proves to
be an effective approach to self-supervised representation learning that approaches or even exceeds results obtained from supervised pretraining [16, 44].

Contrastive methods [20, 52, 101] are based on learning to encode images by discriminative latent embedding vectors. The embedding model is optimized to “pull together” embeddings of two augmented views of the same image representing the same content and “push away” embeddings of different images assumed to represent different content. While the true optimization objective requires optimization over all negative sample pairs, negative sampling approximates the objective by randomly sampling a large set of negative pairs to contrast with the positive pair. Recent non-contrastive methods [16, 47, 117] demonstrate approaches to avoid negative sampling to improve computational efficiency.

Clustering methods [3, 13–15, 113, 119] discovers a set of clusters or prototypes and learns discriminative embeddings for mapping views to a representative cluster. Contrary to contrastive methods, the objective does not have to be approximated as optimizing over the relatively small set of negative representative clusters is tractable. DeepCluster [13] iteratively performs K-means clustering over the entire dataset and subsequently learns an embedding model and an auxiliary classification head to predict the K-means cluster assignment. SeLA [3] presents a principled formulation of simultaneous clustering and representation learning with a single optimization objective by casting cluster assignment as an optimal transport problem [28, 64]. This approach naturally avoids degenerate solutions and removes gradient noise caused by needing to re-learn the cluster classification head every time cluster assignments are updated. SwAV [15] and ODC [119] demonstrate that clustering can be done online per batch to speed up learning.

Our method increases the effectiveness of dense representation learning for images by introducing superpixelization as a natural visual hierarchy for reducing superfluous pixel information. Our problem formulation can be viewed as a self-supervised extension for semantic segmentation.

### 2.2. Self-supervised computer vision

Recently self-supervised representation learning is being applied beyond classification to vision tasks involving explicit spatial information representation. In object detection [5, 29, 105, 106, 110, 112] self-supervision is generally used to learn expressive embeddings for plausible object proposal regions sampled randomly or heuristically [100]. Comparatively few works exist for semantic segmentation, as naively applying general self-supervised methods formulated as classification problems to learn to generate pixel-level embedding maps with millions of vectors is challenging. Existing works leverages self-supervised clustering approaches to learn coherent semantic groupings from pixels [22, 59, 82], autoregressive modeling [82], and GAN-based approaches [7, 19]. Other works [56, 104] leverages self-supervised depth map estimation [42, 71] for enhancing semantic segmentation performance. Recently, DINO [16] demonstrated that attention maps for “things” similar to those of unsupervised object segmentation methods naturally emerge for self-supervised Vision Transformer (ViT) [35, 103] models.

Our work improves the efficiency of self-supervised methods for pixel-level representation learning applicable to high-resolution images through a hierarchical superpixel partitioning of images, resulting in a small set of representative spatial-visual region embeddings for both “things” and “stuff”.

### 2.3. Word embeddings and visual tokens

In natural language processing (NLP) the basic representation of words is categorical tokens or one-hot vectors. Learning semantic embeddings for words using unsupervised methods as a pretraining task [75, 76, 86] offers significant improvements for downstream tasks compared with learning word embeddings as part of the task [99], and is the de facto elementary representation used by all recent language models [32, 62, 67, 114]. The metric of semantic similarity between words is co-occurrence in sentences [50]. Embedding models are optimized so that the embeddings for two words that often co-occur is close in vector space. A separate set of sampled word embeddings assumed to be unrelated are pushed away similarly to noise contrastive estimation [48, 101] to avoid degenerate solutions [43].

Our work extends the idea of word co-occurrence to the vision domain by discovering co-occurring latent visual patterns within a common region of multiple augmented views with different context information and pixel appearance.

In computer vision, the bag of visual words model [37, 97, 98] decompose images into discriminative local image features typically extracted at keypoint locations by a SIFT detector [69]. The algorithm cluster features into a set of clusters or visual words which represent the image content. Later works propose to discover mid-level visual elements or words with richer visual semantics in the form of discriminative patches [96] and mode seeking [33] based on learning through iterative clustering and classification similar to recent self-supervised clustering methods [13] but for representative HOG features [30] in pixel space. More recently, extraction of latent embeddings or tokens for image patches is demonstrated by prior GNN methods [21, 65, 121]. The methods are based on learning transformation matrices or filters for converting local features generated by CNN backbones into latent embeddings. The Visual Transformer (VT) [108] applies this approach with added recurrence to generate visual tokens from current and previous spatial attention maps. These methods require a
where a set of latent visual concepts process. We model this process by a generative model set of latent semantic visual concepts through a generative model of natural images per-

idea, one can view the generation of natural images per-

ance as being generated from a particular pixel appear-

ance, and learning a mapping that consistently pre-

sents the same visual concept embedding map Z within the mutual subset of each view

P. The method is based on learning the inverse mapping

\[ P^{-1}(Z|X) \]

from observed pixels X to the latent visual embedding map Z through simultaneously discovering a set of latent visual concept C and a consistent mapping \( f_\theta \) from pixels to concepts using self-supervision.

Models of biological vision [73] support the existence of latent visual embeddings. According to current theories, the neural code generated by the retina and V1 visual cortex is highly structured, sparse, and robust [26, 89, 94], which higher-level visual centers process further for particular vi-

sion tasks [118]. We propose that our method thus can also be viewed as a means to learn an artificial retina and V1 visual cortex model, which similarly transforms pixels to rich, sparsely encoded features or concepts presumably equally suited for higher-level perception tasks as is evident in bio-

logical vision systems.

The working hypothesis is that contextual co-occurrence of abstract pixel patterns correspond to the same latent semantic visual concept in vector space analogous to how the discovery of latent semantic word embeddings in vector space [75, 76, 86] leverages word co-occurrence [50] and context [32, 87] in NLP. We propose to learn discriminative reoccurring abstract pixel patterns from a large, visual dataset by generating augmented views \( \tilde{X}^{(m)} \) containing a mutual image subset with different context and pixel appearance, and learning a mapping \( f_\theta \) that consistently predicts the same visual concept embedding map Z within the mutual subset of each view m.

\[ f_\theta(\tilde{X}^{(m)}) \simeq Z \quad \forall m \in (1, \ldots, M) \]
The visual concept vectors $c \in \mathbb{R}^D$ are constrained to lie on the surface of a $D$ dimensional unit hypersphere. Each dimension presumably corresponds to a distinct visual concept primitive or basis vector, and visual concepts are linear combinations of these primitives in vector space. The set of visual concept vectors $C$ are known and finite, theoretically allowing tractable probabilistic enumeration over possible configuration similar to recent successful probabilistic language modelling approaches in NLP [32, 90] presuming the vectors are accurately inferrable.

We apply the successful self-supervised learning method SwAV [15] to learn both the mapping function $f_\theta$ and the set of concepts $C$, though in principle any cluster-based self-supervised method [3, 13, 14, 119] can be applied. Fig. 2 shows an overview of our method to learn $f_\theta$, while Fig. 1 illustrates the self-supervised learning method for two concrete regions.

### 3.1. Hierarchy of visual regions

A critical enabler of our method is the introduction of a natural hierarchical step up from semantically meaningless pixels to a small set of visually coherent regions by superpixelization [91], as early experiments showed that naively applying vector comparison based self-supervised representation learning methods on embedding maps generated for high-resolution images proves to be highly inefficient.

Beyond improving efficiency, we propose that the ability to learn from high-resolution images is in itself a performance-boosting factor for dense representation learning, in comparison to prior methods [22, 59] limited to low-resolution images. Higher-resolution images allow for more aggressive geometric view augmentations, likely resulting in better resolution invariant concept mapping. Additionally, it is known in the semantic segmentation literature that training on high-resolution images is beneficial for learning to segment small objects such as poles and pedestrians [17].

Superpixel methods like Simple Linear Iterative Clustering (SLIC) [1] are adept at reducing a high-resolution image from millions of pixels into less than a thousand regions while still representing thin and small pixel patches like poles and pedestrian faces in road scene images as distinct regions. Superpixelation also proves to be a convenient approach to maintain a one-to-one spatial mapping between an arbitrary set of mutual regions in different views with strong geometric augmentation and nonuniform masking regions corresponding to visually coherent patches.

### 3.2. View generation

The following sections explain central aspects of how we generate augmented views suitable for discerning the latent semantic visual concepts through contextual and appearance invariance [20] and geometric equivariance [22]. Two augmented views are shown in Fig. 4 for reference.

![Figure 3. View generation centers sampled from a probability mask representing image complexity measured by the Canny edge detection algorithm [12]. This sampling scheme improves embedding quality by generating a higher degree of content-rich views.](image)

![Figure 4. Examples of two generated view pairs. The top row displays the actual view feed to the model. The middle row illustrates the mutual pixel representation upsampled for visualization purposes. The bottom row showcases mutual superpixel regions colored by region index.](image)

#### 3.2.1 Contextual invariance

View generation centers $(x, y)^*$ are sampled in content-rich regions with high complexity to increase the variety of concepts and thus better satisfy the equipartitioning of concepts assumption [3, 15] as shown in Fig. 3. This approach is analogous to how subsampling frequent words is known to result in speedup and regularity for learning word embeddings [76]. We found that probabilistic sampling of points from a Gaussian filtered Canny edge detection map [12] with a small uniform random probability is a good measure of image complexity. Contextually different views $X^{(m)}$ are generated by sampling $M$ view centers $(x, y)^{(m)}$ around a mutual generation center point $(x, y)^*$ while ensuring a mutual common subset region exists for all views.

Masking words in sentences and learning to predict them from context is a fundamental self-supervised pretraining approach for language modeling [32]. Our method similarly masks superpixels regions encompassed by a randomly
sampled masking region with random noise as a means to learn robust features and alleviate the shortcut learning problem [39]. Prior self-supervision work [4,35,51] demonstrates the effectiveness of applying rectangular masking to generate a learning signal. We propose that masking by superpixel regions improves effectiveness by allowing a finer decoupling of contextual features by aligning regions with visually coherent patches. We also propose region masking as a robust alternative to image inpainting for learning global image statistics without involving a GAN [58].

3.2.2 Geometric equivariance

We generate geometrically equivariant views in two ways. In the first approach, a resize coefficient \( \beta^{(m)} \) from the resize range \([\beta_{\text{min}}, \beta_{\text{max}}]\) is sampled for each view \( m \). \( \beta \) determines the size of the cropped out view region (see the red and blue crop regions in Fig. 1 and Fig. 2). All differently sized view crops are resized to the common view size, thus enforcing the model to learn resolution invariant representations. The second approach resizes the input image to obtain views with a wider context in different resolutions. All views are randomly flipped horizontally.

3.2.3 Appearance invariance

All views are appearance augmented by random color distortion and Gaussian blurring before normalization following the self-supervised vision literature [20, 107, 111] as a means to learn appearance invariant concepts.

3.3. Learning algorithm

The learning objective is designed to simultaneously learn a mapping \( f_0 \) from augmented views \( \tilde{X}^{(m)} \) to a latent visual embedding map \( Z \) as defined in Eq. (2), as well as optimizing the distribution of latent visual concepts \( C \). We decompose the learning problem into two sub-objectives; the coherence objective \( L_{co} \) enforces uniform embeddings for all pixels within each visually coherent region, while the clustering objective \( L_{cl} \) optimizes \( f_\theta \) and \( C \).

The algorithm can be viewed as an extension of SwAV [15] to the problem of learning dense embedding maps for high-resolution images. This section focuses on modifications introduced by our method and refers to prior work for an explanation of SwAV [3, 15, 28, 60]. The rest of this section explains the flow of a training iteration based on Fig. 2. Learning for two concrete regions is illustrated in Fig. 1. We provide pseudocodes in Appendix A.

A training iteration starts by randomly sampling \( N \) images \( X^{(n)} \in \mathbb{R}^{3 \times H \times W} \) with height \( H \) and width \( W \). Each image is partitioned into superpixel regions represented as matrices \( A^{(n)} \in \mathbb{R}^{H \times W} \) with each element specifying the region a pixel belongs to. Next, a set of \( M \) augmented views \( \tilde{X}^{(n)} = \{ \tilde{X}^{(1,n)}, \ldots, \tilde{X}^{(M,n)} \} \) are generated for each image \( X^{(n)} \) by applying cropping and transformations as explained in Sec. 3.2. Here \( h \) and \( w \) denotes the common view size. The superpixel matrices \( A^{(n)} \) are cropped and geometrically transformed identically to the corresponding views, resulting in a set of superpixel view crops \( \tilde{A}^{(n)} = \{ \tilde{A}^{(1,n)}, \ldots, \tilde{A}^{(M,n)} \} \) of only mutual regions existing in all views. Examples of generated views and superpixel crops are shown in Fig. 4. Finally the set of all views \( \tilde{X}^{(m,n)} \) are concatenated into single tensor \( \tilde{X} \in \mathbb{R}^{B \times 3 \times h \times w} \) where \( B = NM \) denotes the total number of views.

A model \( f_0 \) transforms \( \tilde{X} \) into a normalized visual embedding tensor \( \tilde{Z} \in \mathbb{R}^{B \times D \times h \times w} \). After decomposing \( \tilde{Z} \) into view-specific embedding maps \( \tilde{Z}^{(b)} \) and unrolling them into row vector matrices \( \tilde{Z}^{(b)} \in \mathbb{R}^{hw \times D} \), each embedding vector \( z \in \mathbb{R}^D \) 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

\[
S^{(b)} = \tilde{Z}^{(b)} C, \quad \forall b \in \{1, \ldots, B\}, \quad S^{(b)} \in \mathbb{R}^{hw \times K}
\]

with \( C \in \mathbb{R}^{D \times K} \) represented as a matrix. Note that the cosine distance between vectors \( z \) and \( c \) equals the dot product \( z \cdot c \) as both vectors are normalized.

Next all matrices \( \tilde{Z}^{(b)} \) and \( S^{(b)} \) are decomposed into tree structures \( T_Z \) and \( T_S \) where leaf nodes conveniently store row vector embedding maps \( \tilde{Z}^{(i,m,n)} \in \mathbb{R}^{J} \times D \) and score matrices \( S^{(i,m,n)} \in \mathbb{R}^{J \times K} \) indexed by region \( i \) in view \( m \) of image \( n \). Here \( J \) denotes the number of vectors or members \( z^{(j)} \) in region \( i \) of each view and image. At this point non-mutual regions are discarded.

Central to our method is the computation of a single representative embedding vector \( z^{(i,m,n)} \) and score \( s^{(i,m,n)} \) for each region \( i \) in all views and images, corresponding to the embedding vector \( z^{(j)} \) in each region \( i \) closest to some concept vector \( c \)

\[
z^{(i,m,n)*} = \arg \max_{z^{(j)}} z^{(j)} \cdot c \quad \forall k \in \{1, \ldots, K\}
\]

together with the corresponding score vector for \( z^{(j)} \). The representative vectors for all regions are stored in a separate trees \( T_{Z^*} \) and \( T_{S^*} \).

Regional semantic consistency is enforced by the coherence objective \( L_{co} \) formulated as an \( l_1 \) loss which pulls all embeddings \( z^{(j)} \) for a region \( i \) towards the corresponding representative embedding \( z^{(i,m,n)*} \) for all views \( m \) in images \( n \)

\[
L_{co} = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{j=1}^{I} \sum_{i=1}^{J} \sum_{j=1}^{J} ||z^{(i,m,n)*} - z^{(j)}||_1
\]

where \( I \) denotes the number of regions.

The clustering objective \( L_{cl} \) is formulated as a swapped prediction problem [15] whereby the representative score
vectors $s^{(i,m)*}$ for each region $i$ in all secondary views $m > 1$ are used to predict the optimal assignment of concepts $q^{(i)}$ efficiently computed by the Sinkhorn-Knopp algorithm [3, 28] for the same region in the primary view $m = 1$ for each image $n$

$$L_{cl} = -\frac{1}{NM} \sum_{n=1}^{N} \sum_{m=2}^{M} \sum_{i=1}^{I} q^{(i)} \log \sigma \left( \frac{1}{7} s^{(i,m)*} \right)$$  \hspace{1cm} (6)

where $\sigma()$ is the softmax function. We compute $q^{(i)}$ as the optimal soft assignment [15] of visual concepts $C$ given score vectors $s^{(i,1)*}$ for each region $i$ in the primary view with the constraint that all concepts are equally distributed among the embedding vectors $z^{(i,1)*}$. A FIFO queue of accumulated $s^{(i,1)*}$ vectors improves the empirical approximation of a uniform distribution of concepts [3, 15].

The final optimization objective $L$ is simply the average of the two objectives $L = 0.5(L_{co} + L_{cl})$. Experiments show that $L_{co}$ tend to approach zero and thus acts as a constraint while $L_{cl}$ discovers latent visual concepts and the mapping from pixels to visual embeddings.

4. Experiments

We implement ViCE in the self-supervised learning library framework VISSL [46] written in Pytorch [83] and evaluate embedding representation quality on the semantic segmentation downstream tasks using a linear classification model. For evaluation, we leverage the semantic segmentation framework MMSegmentation [24] with minimal modifications for the benefit of reproducibility of results and fair performance comparison. Our primary comparative baseline for dense representation learning is the recent SOTA unsupervised semantic segmentation model PiCIE [22] based on DeepCluster [15].

4.1. Implementation details

We use the recent SOTA semantic segmentation model DeepLabV3+ [18] with a ResNet-50 backbone [53] for generating visual embeddings maps $Z \in \mathbb{R}^{D \times H \times W}$ with the same spatial resolution as the input image. Two embeddings $z^{(a)}$ and $z^{(b)}$ are compared for semantic similarity using the dot product $z^{(a)}, z^{(b)}$ on normalized vectors. This operation is equivalent to comparing two word embeddings by cosine distance [75, 76].

4.2. Training Details

We conduct experiments on 32 V100 32 GB GPUs distributed over eight nodes. Each GPU loads four high-resolution images and generates five augmented views with a size of $400 \times 400$ pixels, resulting in a total batch size of 128 images or 640 views. We upsample small images to a minimum size of 500 pixels before generating views. For generating superpixels we use SLIC [1] implemented in OpenCV [8] with average region size 20 px, smoothness factor (i.e. ruler) 10, and optimize for 5 iterations. Maximal mask coverage is 25% with a minimum common superpixel region preservation ratio of 25%. View resize coefficients $\beta$ are sampled between 0.5 to 2. The embedding dimension $D$ is 128, and the number of visual concepts $C$ is 128. We use the same set of hyperparameters in all benchmarking experiments. The models are initialized by the default Pytorch pre-trained ResNet 50 weights. See Appendix B for an additional hyperparameter study.

Parameters for the clustering objective $L_{co}$ are the same as SwAV [15] with temperature $\tau$ set to 0.1, regularization parameter $\epsilon$ equaling 0.05, and 3 iterations for the Sinkhorn-Knopp optimization algorithm [28]. The FIFO queue consists of 5K representative score vectors $s^*$ per GPU. The model is optimized using the LARS optimizer [115] with weight decay $10^{-6}$. The learning rate (LR) schedule consists of a linear warmup phase followed by cosine decay [68, 77]. We set the peak LR using the linear LR scaling rule [45] with the optimal base LR 0.04 found through a set of single node 4 GPU experiments.

4.3. Representation quality experiments

We evaluate the representation quality of the visual concept embedding map $Z$ generated by ViCE through performance on a downstream semantic segmentation task by a linear $1 \times 1$ segmentation model. Note that our method works on high-resolution images and thus does not need to downsample images like prior methods [22, 59].

The first experiment involves training and evaluating ViCE on generic natural images represented by the COCO-Stuff164K dataset [11, 66]. The model is trained for 12 epochs. Evaluation is performed on a reduced set of 27 classes obtained by merging the original 172 semantic classes following [59]. The results in Tab. 1 demonstrate the leap in performance achieved by ViCE, improving the prior SOTA by +12.94 mIoU (+87.6%). Two additional 4 and 8 epoch experiments are performed to assess further learning potential. The performance curve in Fig. 5 shows that ViCE is expected to continue improve linearly with more training. Viewed together with the hyperparameter experiments results in Appendix B, this result indicates favorable compute and model complexity scaling properties for our method not shown in prior work [22, 59].

The second experiment involves high-resolution road scene images. ViCE is trained on a set of 1M varying road scene images collected from eight public datasets [40, 55, 57, 78, 88, 102, 109, 116]. We evaluate performance on the Cityscapes dataset [25]. A linear classification model is used to evaluate ViCE, while the other methods employ a clustering model. The results in Tab. 2 show the same leap in performance despite not training on Cityscapes im-
Embedding extractor & mIoU & Acc.
IIC [59] & 13.26 & 51.49 \\
Mod. DeepCluster [22] & 13.76 & 50.79 \\
PiCIE [22] & 14.77 & 54.75 \\
ViCE (ours) & **27.71** & **65.47** \\

Table 1. Representation quality experiment results on COCO. Performance is evaluated using a linear classifier.

Embedding extractor & mIoU & Acc.
IIC [59] & 6.35 & 47.88 \\
Mod. DeepCluster [22] & 7.06 & 40.67 \\
PiCIE [22] & 12.31 & 65.50 \\
ViCE* (ours) & **28.83** & **85.51** \\

Table 2. Representation quality experiment on Cityscapes. *Note that ViCE is not trained on Cityscapes unlike the other methods.

Figure 5. The linear performance curve of the COCO representation quality experiment indicates performance is expected to improve proportionally with more training epochs.

Figure 6. Visualizing the embedding maps shows how ViCE discovers distinct semantic visual entities and learns the mapping from pixels, such as similar semantic representations for trees, ground, heads, and torsos.

torsos. We visualize embedding maps by linearly reducing the dimensionality of each vector $z$ using PCA [85] and rescale values to the RGB range.

5. Conclusion

In this work, we presented a new self-supervised representation learning method ViCE for dense embedding maps based on learning a mapping from pixel appearances to latent visual concepts. ViCE significantly improved on the prior SOTA through a hierarchical decomposition of an image into a small set of visually coherent regions. Experiments demonstrated favorable scaling and domain generalization properties for our method.

As for limitations, we observed that ViCE does not seem efficient to learn a backbone from scratch and therefore needs to rely on another self-supervised method [15,20] for initialization. We believe this limitation indicates a possibility for further improvement to our method. We identified the potential for racial discrimination caused by representing people of different ethnicity by separate visual embeddings. However, we found no indications of this trait in our embedding visualizations. Nevertheless, we urge the research community to ensure ethical fairness when applying self-supervised models. Training self-supervised models negatively impact the environment through large energy consumption. We believe our research will contribute towards realizing general vision models with efficient downstream task models and conserve resources in the long run.
References

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

[2] Rie Kubota Ando, Tong Zhang, and Peter Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research (JMLR), 6(11):1817–1853, 2005.

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

[4] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. ArXiv, abs/2106.08254, 2021.

[5] Amir Bar, Xin Wang, Vadim Korotov, Colorado Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. DETReg: Unsupervised pre-training with region priors for object detection. ArXiv, abs/2106.04550, 2021.

[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Presentation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 35(8):1798–1828, 2013.

[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] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In NeurIPS, volume 33, pages 1877–1901, 2020.

[10] 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.

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

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

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

[14] 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.

[15] 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.

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

[17] Liang-Chieh Chen, George Papandreou, Jiasen Lu, Vladlen Koltun, and Lucidic Denoyer. Unsupervised object segmentation by redrawing. In NeurIPS, volume 32, pages 12705–12716, 2019.

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

[19] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Shuicheng Yan, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In CVPR, pages 433–442, 2019.

[20] 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.

[21] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In IJCAI, pages 160–167, 2008.

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

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

[24] Gabriel Cristobal, Laurent Perrinet, and Matthias Kell. Bio- logically Inspired Computer Vision: Fundamentals and Applications. Aug. 2015.

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

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

[27] 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.
A. Pseudocodes

Algorithm 1 explains the generation of $M$ views for a batch of $N$ images. The algorithm samples an image $X^{(n)}$ and computes a superpixel index map $A^{(n)}$. $M$ views are generated from the sampled image and superpixel index map. Each of these views are randomly masked before being resized to the same pixel dimension. Only mutual regions existing in all views are kept. All views are geometrically augmented by random horizontal flipping, and appearance augmented by color distortion and randomly blurred. All generated views are gathered and converted into a 4D tensor.

### Algorithm 1 View generation

\[
\begin{align*}
\hat{X} &:= \{ \} \quad \triangleright \text{Empty sets} \\
\hat{A} &:= \{ \} \\
&\text{for } n \in \{1, \ldots, N\} \text{ do} \\
&\quad X^{(n)} \sim \text{dataloader} \quad \triangleright \text{Sample an image} \\
&\quad A^{(n)} := \text{superpixels}(X^{(n)}) \\
&\quad \hat{X}^{(n)}, \hat{A}^{(n)} := \text{gen_views}(X^{(n)}, A^{(n)}) \\
&\quad \# \hat{X}^{(n)} = \{ \hat{X}^{(1, n)}, \ldots, \hat{X}^{(M, n)} \} \\
&\quad \# \hat{A}^{(n)} = \{ \hat{A}^{(1, n)}, \ldots, \hat{A}^{(M, n)} \} \\
&\quad \hat{X}^{(n)}, \hat{A}^{(n)} := \text{mask_views}(\hat{X}^{(n)}, \hat{A}^{(n)}) \\
&\quad \hat{X}^{(n)}, \hat{A}^{(n)} := \text{resize_views}(\hat{X}^{(n)}, \hat{A}^{(n)}) \\
&\quad \hat{X}^{(n)}, \hat{A}^{(n)} := \text{mutual_regions}(\hat{X}^{(n)}, \hat{A}^{(n)}) \\
&\quad \hat{X}^{(n)}, \hat{A}^{(n)} := \text{appearance_aug}(\hat{X}^{(n)}) \\
&\quad \hat{X} := \hat{X} + \hat{X}^{(n)} \quad \triangleright \text{Add new views to set} \\
&\quad \hat{A} := \hat{A} + \hat{A}^{(n)} \\
&\text{end for} \\
\end{align*}
\]

Algorithm 2 explains the learning algorithm. The model $f_\theta$ generates an embedding map $\tilde{Z}$ from the image view tensor $\hat{X}$. The single tensor $\tilde{Z}$ is decomposed into $B$ tensors $\tilde{Z}^{(b)}$ each corresponding to a single view. Next, four trees are created to contain the latent visual embeddings $z$, concept compatibility scores $s$ for all elements in each mutual region $i$, as well as representative vectors $z^*$ and $s^*$. The two objectives $L_{co}$ and $L_{cl}$ are computed using the trees. The model parameters $\theta$ and set of visual concept vectors $C$ are optimized based on the mean loss $L$.

The coherence objective $L_{co}$ explained in Algorithm 3 is based on minimizing the distance between all embeddings $z^{(j)}$ and the representative embedding $z^{(i)*}$ for each region $i$ in all views $m$ of all images $n$.

The clustering objective $L_{cl}$ shown in Algorithm 4 first computes an optimal assignment of visual concepts $Q$ based on the representative scores in the first view $m = 1$. The loss is minimized when predicted visual embeddings in secondary views $m \geq 1$ are closer to the optimally assigned visual concept vectors for each region $i$ in all views $m$ of
Algorithm 4 Clustering objective

\[
\mathcal{L}_{cl} := 0 \\
Q := \text{optimal assignment}(T_{S,*}) \\
\text{for } n \in \{1, \ldots, N\} \text{ do} \\
\quad \text{for } m \in \{2, \ldots, M\} \text{ do} \\
\quad \quad \text{for } i \in \{1, \ldots, I\} \text{ do} \\
\quad \quad \quad q^{(i)} := Q(n, i) \\
\quad \quad \quad s^{(i)*} := T_{S,*}(n, m, i) \\
\quad \quad \quad p^{(i)} := \sigma(\frac{1}{2}s^{(i)*}) \\
\quad \quad \quad \mathcal{L}_{cl} := -q^{(i)} \log p^{(i)} \\
\quad \text{end for} \\
\quad \mathcal{L}_{cl} := \mathcal{L}_{cl}/I \\
\text{end for} \\
\mathcal{L}_{cl} := \mathcal{L}_{cl}/(N(M-1))
\]

B. Hyperparameter study

We quantify the effect of hyperparameter choices by running a set of COCO representation quality experiments for four epochs where we change a single parameter in an otherwise static baseline configuration. The experiments are listed in Tab. 3. We compare results to the baseline experiment with view size 500 px, maximal mask coverage 50 %, 128 concepts, queue size of 5K vectors, four views, embedding dimension $D_{64}$, and modest view resize range (0.5, 1.5).

The results indicate that smaller view sizes are favorable, as also noted in DINO [16]. Modest masking proves to be better than no masking. The ideal number of concepts needs to be found by experiments. Increasing the number of views improves representation learning, as also noted in SwAV [15]. Larger embedding size $D$ results in more expressive embeddings. The benefit of increasing $D$ is confirmed by an additional experiment using smaller 400 px view sizes to fit training jobs in GPU memory. All benchmark experiments presented in the main paper use the optimal hyperparameters found in this study.

C. Visualizations

Representation quality is evaluated qualitatively through visualizations in Fig. 7 displaying output for linear semantic segmentation models on Cityscapes images. We visually observe several improvements in our embeddings over the SOTA unsupervised segmentation model PiCIE [22]; semantics better fit actual object boundaries, clear discrimination between sidewalk and road, and semantics for smaller objects such as pedestrians and bicycles are inferrable. The comparison illustrates the difference between 28.83 and 12.31 mIoU for ViCE and PiCIE, respectively. Note that we evaluate ViCE using a linear $1 \times 1$ convolution model while PiCIE [22] uses a linear overclustering model. The original paper [22] shows that the performance of both evaluation models is comparable. We choose to visualize content-rich and thus more challenging images from the “Frankfurt” validation set for our model.

Fig. 8 shows visualizations of the secondary total information experiment. Judging by the detail for large and small semantics, it is apparent that the generated visual embeddings contain more information than directly interpretable using a linear model as compared with Fig. 7. The model trained on latent visual embedding maps generated by ViCE can generate semantic segmentation output of the same visual quality as the model trained on raw pixels. Quantitative evaluation results are 70.39 and 79.31 mIoU for the embedding and pixel semantic segmentation model, respectively. From these results, we conclude that the learned visual embedding map simultaneously possesses highly discriminative and spatially precise information extractable using a nonlinear model, as well as being directly interpretable by a linear model.

Fig. 9 shows additional visualizations of images with people from the COCO-Stuff164K validation set [111]. The mapping for faces is found to be relatively consistent between people of different ages and ethnicity, as well as varying context, further demonstrating our method’s ability to discover contextual and pixel appearance invariant semantic representations of visual entities or concepts without human supervision or heuristic proposals [5, 100].

| Hyperparameter change | $\Delta$ mIoU |
|-----------------------|---------------|
| View size 500 → 400 px | +2.34 (+12.7%) |
| Masking ratio 50% → 25% | +1.52 (+8.3%) |
| Masking ratio 50% → 0% | +1.35 (+7.4%) |
| #Concepts 128 → 64 | -0.45 (-2.5%) |
| #Concepts 128 → 256 | -0.59 (-3.2%) |
| Queue size 5K → 10K | -0.73 (-4.0%) |
| #Views 4 → 3 | -1.22 (-6.6%) |
| Emb. size $D_{64}$ → 32 | -1.48 (-8.1%) |
| Resize range (0.5, 1.5) → (0.15, 2.0) | -1.86 (-10.1%) |

Table 3. Hyperparameter experiments. The left column indicates the changed parameter, and the right column shows the resulting change in performance. All parameters besides the specified one remain unchanged.
Figure 7. Visualizations for representation quality experiments using a linear semantic segmentation model. The left side shows results generated based on the visual embedding map generated by our ViCE model on full-resolution Cityscapes images. The right side shows results from PiCIE [22] on a different set of downsampled images.
Figure 8. Visualizations for representation quality experiments using a nonlinear semantic segmentation model trained on latent visual embedding maps generated by ViCE and a regular baseline model trained on pixel images. Both models generate output of comparable quality, implying that the visual embedding map retains a significant degree of information.
Figure 9. Visualization of embedding maps for images with people of different ages and ethnicity. The results showcase the consistency for which our method learns to map faces with varying appearances, observed in diverse contexts, to a similar latent visual semantic concept.