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

Robin Karlsson (1), Tomoki Hayashi (1), Keisuke Fujii (1), Alexander Carballo (2), Kento Ohtani (1), Kazuya Takeda (1,3)
(1) Nagoya University, (2) Gifu University, (3) TIER IV

Email: karlsson.robin@g.sp.m.is.nagoya-u.ac.jp Full version: https://arxiv.org/abs/2111.12460v3 Github: https://github.com/robin-karlsson0/vice

Self-supervised dense representation learning

- Recent **self-supervised classification models** are better or equal to fully supervised models
- However, these methods are **ineffective for learning dense representations**
  - Typically reduce images into a **tiny feature 7x7 map** that discards precise spatial information about content
  - **Prohibitive computational demand** for contrastive learning methods [1] on high-resolution images

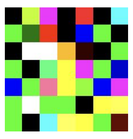
Regional decomposition by superpixels

- Decompose images into a **small set of visually coherent regions**
  - Adjacent similar pixels represent same semantics (no information loss)
  - Superpixel regions conform to image content (unlike grids)
- **Reduce computational complexity** by **O(1000)**
  - Improve efficiency of contrastive methods on high-resolution images
  - Experiments show that learning from high-resolution images is beneficial

1280x720 pixels ⇒ 921,600 pixels/vectors

Resnet50

Spatial information about content is lost

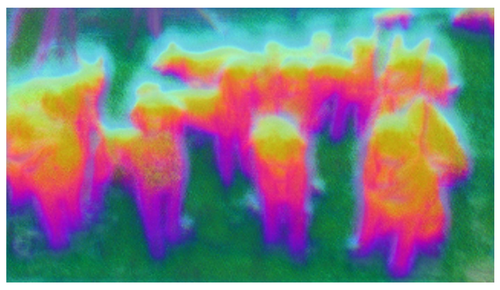


Embedding RGB visualizations (7x7)

Superpixelization

2304 vectors (99.75% reduction)

ViCE



Embedding RGB visualizations (1280x720)

ViCE: Visual Concept Embeddings

- View natural **images X** as a result of a **dense representation Z** of **latent visual concepts C = (c_1, … , c_K)** transformed by a **stochastic generative process f(X|Z)**
- Stochasticity: Assume a single latent Z representation maps to many potential images X
  - Same visual content but different pixel appearance
- Method: Learn an **approximative inverse function f'(Z|X)** and a **set C**

* Superpixels only used during training

Table 8: Average inference time for a high-resolution image

	Segmentation model	Cluster model	Linear model
[msec]	57	2395	15

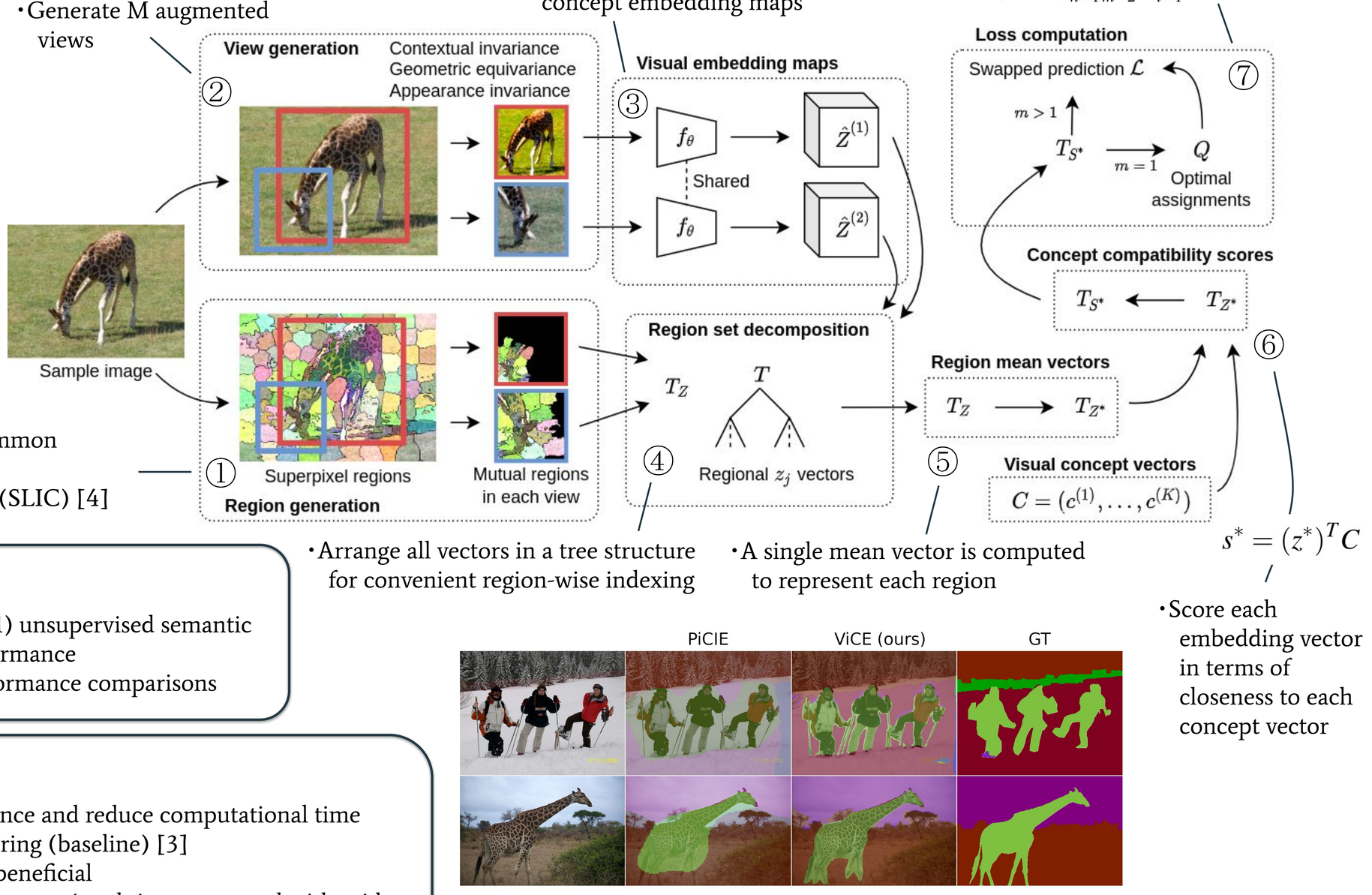
Learning method

Fundamental learning signals

- **Same content** with **different pixel appearance** by data augmentation [2,3]
- Semantic meaning from **spatial co-occurrence and context** (similar to learning word embeddings)

- Generate M augmented views
- Transform views into visual concept embedding maps
- Objective optimize the similarity of embeddings within all views, as well as the distribution of visual concepts

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



- Partition image into I mutually common superpixel regions
- Simple Linear Iterative Clustering (SLIC) [4]

- Arrange all vectors in a tree structure for convenient region-wise indexing
- A single mean vector is computed to represent each region

$$s^* = (z^*)^T C$$

- Score each embedding vector in terms of closeness to each concept vector

Experiments

- Evaluate representation quality through 1) unsupervised semantic segmentation and 2) linear model performance
- Low-resolution and high-resolution performance comparisons

Results

- **Image decomposition** improves performance and reduce computational time
- **Online clustering** (ViCE) > Offline clustering (baseline) [3]
- Learning from **high-resolution images** is beneficial
- **Superpixels** improve performance and computational time compared with grids
- **General vision models** learn more useful embeddings also when comparing on a narrow domain

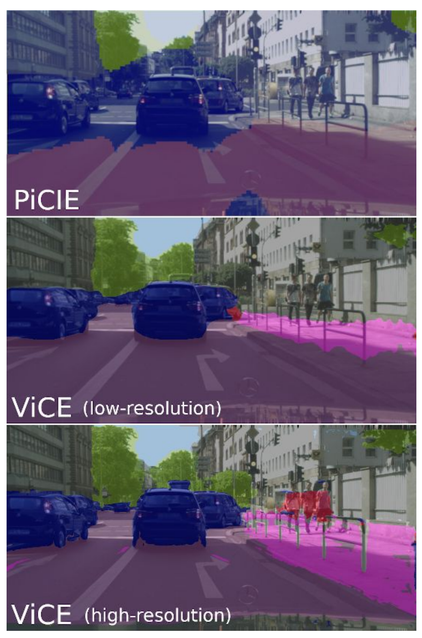


Table 1: Representation quality experiment results on low- and high-resolution images.

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◇	14.60	48.37		C 27◇	11.85	64.29
	C 27⋆	9.27	38.31		C 27⋆	8.80	82.48
	C 128⋆	10.75	49.81		C 128⋆	7.97	56.52
	C 256⋆	12.42	**66.02**		C 256⋆	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⋆	11.55	50.49		C 27⋆	19.52	80.34
	C 128⋆	16.66	52.33		C 128⋆	21.48	81.55
	C 256⋆	17.98	54.92		C 256⋆	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⋆	**21.77**	64.75	ViCE (high-res)	C 256⋆	**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	-	-

C K: evaluation w. K clusters, ◇ reproduced results, ⋆ greedy assignment, ∗ ViT-based models

Conclusions

- Present a **new SOTA unsupervised semantic segmentation** method ViCE for learning to generate **dense embedding maps for high-resolution natural images**
- **Decomposing images** by superpixelization improves the effectiveness of classification-based self-supervised methods
- **Superpixels perform better** than conventional grid decomposition
- Hope our work will **raise interest in incorporating non-uniform image decomposition techniques** to improve other self-supervised computer vision methods including ViT-based models [5]

References: [1] Caron M. et al. NeurIPS, 2020 [2] Chen T. et al. ICML 2020 [3] Cho J. et al. CVPR 2021
[4] Achanta R. et al. PAMI, 2274-2285, 2012 [5] Caron M. et al. ICCV, 2021

Acknowledgements: This work was supported by supported by JST SPRING (Grant Number JPMJSP2125), JSPS KAKENHI (21H04892, 21429770), OPERA (JP-MJOP1612)
The computation was carried out through the "General Projects" program on the supercomputer "Flow" at the Information Technology Center, Nagoya University

NAGOYA UNIVERSITY