ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Local Concept Embeddings in the Context of Self-Supervised Learning

Kim Paulke[1] Hans-Oliver Hansen[2] Thomas Martinetz[2]
Gesina Schwalbe[2]

1- Helmut Schmidt University, Hamburg - Germany
2- University of Lübeck, Germany

Abstract. This work investigates how self-supervised learning (SSL) frameworks encode semantic structure within their latent representations using the introspection technique *Local Concept Embeddings* (LoCE). We analyse three complementary SSL paradigms—contrastive (Barlow Twins), generative (Denoising Autoencoder), and predictive (Relative Patch Location)—all pretrained on the Cityscapes dataset and evaluated in a semantic segmentation setting. LoCE reveals that the Denoising Autoencoder produces the most distinct and coherent concept clusters (highest separability and clustering metrics), while Barlow Twins and RPL exhibit moderate structure and higher intra-class variability. Furthermore we find that greater latent disentanglement before fine-tuning correlates with improved segmentation performance, uncovering an interesting link between latent organization and downstream generalization.

1 Introduction

Self-supervised learning (SSL) has emerged as a powerful approach for representation learning in computer vision, enabling models to learn semantic and structural features from unlabelled data via pretext tasks [1, 2], boosting performance in downstream tasks such as classification and segmentation [1, 3]. However, conventional performance metrics such as mean Intersection over Union (mIoU) offer limited insight into how SSL models internally organise semantic concepts within their latent representations. Recent work in explainable AI and representation interpretability has characterized latent spaces using clustering, dimensionality reduction, and concept-based methods [4, 5]. Although, most approaches focus on global or feature-level statistics and rarely offer a detailed, local perspective on how semantic concepts are organised within the feature space across different SSL paradigms. To address this gap, this study employs *Local Concept Embeddings* (LoCE) [6], a framework designed to analyse the local semantic organisation of latent feature spaces.

Contributions. We systematically **compare three complementary SSL paradigms**: *contrastive* (Barlow Twins [2]), *generative* (Denoising Autoencoder), and *predictive* (Relative Patch Location). Each paradigm is pretrained on the Cityscapes dataset and fine-tuned for semantic segmentation. Our analysis demonstrates that (1) self-supervised pretraining yields **more coherent and semantically disentangled latent representations** than models trained from random initialization; and (2) models with **stronger concept separation consistently achieve higher segmentation performance**, underscoring the

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

link between representational structure and downstream task quality. In particular, contrastive and generative paradigms show more stable and interpretable concept boundaries compared to predictive ones.

2 Background and Related Work

2.1 Self-supervised Learning

SSL pretrains deep neural networks by leveraging intrinsic data structure to learn generalizable representations without human annotations, effectively transferring to downstream tasks like semantic segmentation. SSL encompasses three main categories: (1) *Contrastive* (e.g. SimCLR [1], **Barlow Twins** [2]): maximizes agreement between augmented views of the same image while separating others. (2) *Generative* (e.g., **denoising autoencoders** [7]): reconstructs corrupted inputs to capture semantic structure. (3) *Predictive* (e.g., **RPL** [8]): learns spatial or temporal relationships within data. While SSL improves generalization, particularly for dense prediction tasks, standard performance metrics (e.g., mIoU) lack insight into the internal semantic organization of learned representations. This motivates advanced representational analysis, specifically Local Concept Embeddings (LoCE) [6] as discussed in the following.

2.2 Explainable AI for Learned Representation Analysis

Explainable artificial intelligence (XAI) techniques aim to provide crucial transparency [9] about individual predictions [10] or the globally encoded knowledge inside black-box models like deep neural networks [11]. For the latter, a recent focus is on *concept-based explanations* that provide insights into how human-interpretable concepts are represented in DNN latent space [12]. This paradigm views the latent space as a semantic manifold where human-understandable concepts are represented as directions or subspaces rather than low-level features [11, 12]. A first technique to post-hoc analyse latent representations is TCAV [5], which uses labeled concept samples to learn representative vectors in intermediate feature layers. Extensions of these methods have explored nonlinear embeddings, contrastive pretraining, and manifold clustering to improve concept coherence and separability [12]. While these global concept approaches have proven effective for finding and comparing concept representations, they typically average concept representations across samples. Thus, they fail to capture concept distribution in latent space that are important for detailed interpretability, which was recently attended by the **Local Concept Embeddings (LoCE)** [6] technique.

LoCE represents a recent advancement in concept-based explainability, focusing on *local*, context-sensitive structures in latent representations. Unlike global methods like TCAV which assign a single concept vector, LoCE computes a dedicated embedding vector for each image-concept pair. This allows for nuanced analysis of how semantic concepts (e.g., *car*) are represented under varying contexts. LoCE operates post-hoc on internal layer activations, learning

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

projection vectors from latent activations to binary concept masks via a differentiable loss. These concept vectors can then be visualized (e.g., UMAP) or clustered (e.g., GMM) to quantify *concept separation*, *overlap*, and *sub-concept formation*. LoCE thus bridges the gap between local interpretability and global representation analysis, offering critical insights into SSL paradigms' inductive biases and generalization behavior.

3 Experimental Setup

Dataset and Preprocessing. We employed the Cityscapes Dataset, a standard benchmark for urban semantic segmentation. Comprising high-resolution images (1024×2048 pixels) from 50 German cities under diverse conditions, the dataset provides rich visual variability. For computational efficiency, all images were downscaled to 128×256 pixels. Standard augmentations, including random cropping, horizontal flipping, grayscale conversion, and color jitter, were applied.

The dataset includes 5000 finely annotated images (2975 training, 500 validation, 1525 test). We used the validation set for evaluation due to unavailable test labels. Semantic labels cover 30 categories, with 19 used for benchmarking. Cityscapes exhibits strong class imbalance, with four classes (*road, building, vegetation, car*) accounting for over 79% of pixels, posing a challenge for rare class generalization.

Self-supervised Pretraining. In this work three SSL pretext tasks were employed: contrastive (Barlow Twins), generative (Denoising Autoencoder), and predictive (Relative Patch Location). For each task, the encoder comprised the downsampling path of a U-Net [13]. Task-specific data augmentations were applied to facilitate robust and generalizable feature learning: **(1) Barlow Twins:** Augmentations were adapted from [2], including random resized crops, horizontal flipping, color jitter, and grayscale conversion. **(2) Denoising Autoencoder:** Primarily additive Gaussian noise ($\sigma = 0.1$) was applied to input images, compelling the network to reconstruct clean images and thus capture underlying structural information. **(3) Relative Patch Location:** Images were divided into a 3×3 grid. A randomly selected patch was then resized to the network's input dimensions. Grayscale conversion and horizontal flipping were applied to enhance input variability while preserving absolute positional information. The model was trained to predict the absolute position of each patch within the grid, thereby encouraging the learning of semantically meaningful spatial features from local image contexts.

Downstream Task and LoCE Analysis. The pretrained encoders were subsequently frozen and utilized as feature extractors for the downstream semantic segmentation task. For Local Concept Embeddings (LoCE) analyses, activations from the deepest encoder layer were extracted. Ground-truth labels, interpolated to match the spatial resolution of the encoder outputs, served as pixel-wise binary masks for each concept. This configuration ensures that the

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

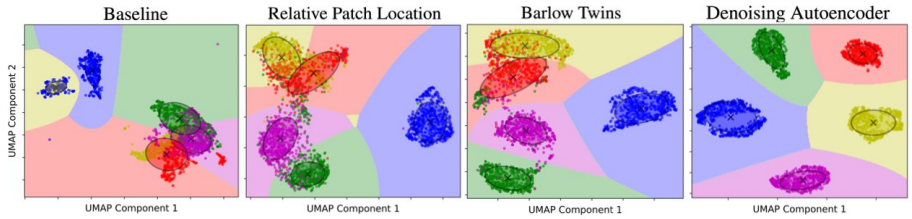

Fig. 1: Visualisation of LoCE with GMM decision boundaries illustrating the structure and separability of concepts across SSL paradigms. Colors denote classes: blue–road, yellow–car, green–vegetation, purple–building, red–person.

LoCE evaluation accurately reflects local semantic representations without interference from batch normalization or other post-processing effects. Overall, this experimental design enables a systematic investigation into how different SSL pretraining strategies influence the quality of learned representations and their utility for both downstream segmentation and local concept discovery.

4 Results

Our framework-independent analysis using Local Concept Embeddings (LoCE), as illustrated in Fig. 1, demonstrates that all three SSL approaches (contrastive, generative, predictive) structure semantically meaningful concepts in their latent spaces. For comparison, the Baseline (random initialization, no training) shows only weakly organized and strongly overlapping representations. Thus, models develop latent representations increasingly aligned with semantic categories even without explicit supervision. However, the degree of concept separation and cluster compactness varies notably between paradigms.

The Denoising Autoencoder (DAE) consistently produced the most distinct and compact clusters, clearly separating all five evaluated concepts (*road, building, vegetation, car, person*). Its reconstruction objective fosters sensitivity to both local image details and global scene context, yielding a semantically well-differentiated representation. In contrast, Barlow Twins and Relative Patch Location (RPL) exhibited stronger concept overlap, particularly between *car* and *person*, likely due to contextual co-occurrence and limited semantic disentanglement. While Barlow Twins effectively captures stable visual structures (e.g., *road, vegetation*), RPL emphasizes global layout regularities over fine-grained texture information, leading to reduced separability among visually similar or contextually adjacent concepts. These findings suggest paradigm-specific strengths: generative approaches achieve superior local concept isolation and semantic precision, while contrastive and predictive methods excel in capturing global structural regularities.

Quantitative evaluation, summarized in Table 1, corroborates these observations. The DAE consistently achieved superior scores across all clustering metrics. Specifically, it recorded the highest Inter-Class Distance (24.51), Separability Ratio (5.35), and Silhouette Score (0.73), indicative of highly compact

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Table 1: Quantitative metrics for local concept separation and clustering quality across SSL paradigms.

		Baseline	**Barlow Twins**	**Denoising**	**RPL**
Distance	Mean Intra-Class Distance ↓				
	(car / person / road)	**1.67/1.83/2.17**	5.57/5.54/4.83	4.825/3.35/5.768	5.42/5.58/3.78
	(vegetation/building)	**2.16/2.05**	8.89/5.35	4.470/ 4.47	6.66/4.56
	Mean Inter-Class Distance ↑	4.99	17.8	**24.51**	14.99
	Separability Ratio ↑	2.52	2.95	*5.35*	2.88
	Silhouette Score ↑	0.43	0.54	**0.73**	0.53
Clustering	Homogeneity ↑	0.45	0.73	**1**	0.64
	Completeness ↑	0.49	0.74	**1**	0.64
	V-Measure ↑	0.47	0.74	**1**	0.64
	Adjusted Rand Index ↑	0.27	0.71	**1**	0.6
	NMI ↑	0.47	0.73	**1**	0.64
	ER ↓	0.51	0.15	**0**	0.24

and well-separated clusters. All clustering-based metrics-Homogeneity, Completeness, V-Measure, Adjusted Rand Index (ARI), and Normalised Mutual Information (NMI)-reached their theoretical maxima (1.0) for the DAE, with an Error Rate (ER) of 0.0. Barlow Twins showed moderate separability (0.7-0.74 across various metrics), whereas RPL and the random baseline demonstrated weaker latent structure characterized by overlapping concepts and higher intra-class variance. Notably, no method exhibited distinct sub-concept formation, suggesting that LoCE captures a semantically coherent but non-hierarchical organization of concepts.

In downstream **semantic segmentation** on Cityscapes, SSL pretraining consistently improved performance over random initialization. The DAE achieved the best results (mIoU 0.4573, acc. 0.8712), effectively capturing fine structural details. Barlow Twins performed competitively (mIoU 0.4214, acc. 0.8639), though with slightly less sensitivity to small-scale structures. RPL showed smaller gains (mIoU 0.3485, acc. 0.8203), indicating limitations in representing local visual variations. Across all methods, pretraining particularly benefited rare and structurally complex classes.

These findings strongly align with our LoCE analyses: models exhibiting higher concept separation and lower concept overlap consistently achieved better generalization and segmentation accuracy. Specifically, the DAE's superior local concept isolation translated directly to higher mIoU. Overall, generative SSL objectives foster fine-grained semantic differentiation, contrastive learning promotes robust global structure encoding, and predictive models like RPL capture spatial regularities. This strong interplay highlights LoCE's potential as a diagnostic tool for understanding how different SSL paradigms encode semantic information and for connecting latent structure to downstream task performance.

5 Conclusion

This work investigated the impact of self-supervised pretraining on semantic segmentation, focusing on the structure and quality of latent representations

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

and their effect on downstream performance. Using LoCE, we analyzed how Barlow Twins, Denoising Autoencoder, and Relative Patch Location capture semantic concepts in local image regions. All methods learned meaningful representations, but the Denoising Autoencoder achieved the clearest separation of all considered concepts, including closely related categories such as *car* and *person*, which resulted in superior segmentation performance. Barlow Twins captured robust, structured background features, while RPL showed moderate improvements over the baseline. Our results demonstrate a strong correlation between the semantic coherence of latent representations and downstream task performance. These findings suggest that SSL methods optimising both local detail and global context produce more informative representations for complex segmentation tasks.

Acknowledgments GS acknowledges support through the project "chAI" funded by the German Federal Ministry of Research, Technology and Space (BMFTR), grant no. 16IS24058. The authors are solely responsible for the content of this publication.

References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*. PmLR, 2020.

[2] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*. PMLR, 2021.

[3] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[4] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proc. Nat. Academy of Sciences*, 117(48), 2020.

[5] Been Kim et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*. PMLR, 2018.

[6] G. Mikriukov, G. Schwalbe, and K. Bade. Local concept embeddings for analysis of concept distributions in vision dnn feature spaces. *IJCV*, 2025.

[7] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.

[8] C. Doersch, A. Gupta, and A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[9] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3), 2017.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[11] G. Schwalbe and B. Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38(5), 2024.

[12] J. H. Lee, G. Mikriukov, G. Schwalbe, S. Wermter, and D. Wolter. Concept-based explanations in computer vision: Where are we and where could we go? In *ECCV*. Springer, 2024.

[13] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 2015.